Detecting and Characterizing Web Bot Traffic in a Large E-commerce Marketplace

Haitao Xu[1], Zhao Li[2], Chen Chu[2], Yuanmi Chen[2], Yifan Yang[2], Haifeng Lu[2],
Haining Wang[3], and Angelos Stavrou[4]

[1] Arizona State University, Glendale, AZ 85306, USA
hxu@asu.edu
[2] Alibaba Group, Hangzhou, China
{lizhao.lz,chuchen.cc,yuanmi.cym,yifan.yy,haifeng.lhf}@alibaba-inc.com
[3] University of Delaware, Newark, DE 19716, USA
hnw@udel.edu
[4] George Mason University, Fairfax, VA 22030, USA
astavrou@gmu.edu

Abstract. A certain amount of web traffic is attributed to web bots on the Internet. Web bot traffic has raised serious concerns among website operators, because they usually consume considerable resources at web servers, resulting in high workloads and longer response time, while not bringing in any profit. Even worse, the content of the pages it crawled might later be used for other fraudulent activities. Thus, it is important to detect web bot traffic and characterize it. In this paper, we first propose an efficient approach to detect web bot traffic in a large e-commerce marketplace and then perform an in-depth analysis on the characteristics of web bot traffic. Specifically, our proposed bot detection approach consists of the following modules: (1) an Expectation Maximization (EM)-based feature selection method to extract the most distinguishable features, (2) a gradient based decision tree to calculate the likelihood of being a bot IP, and (3) a threshold estimation mechanism aiming to recover a reasonable amount of non-bot traffic flow. The detection approach has been applied on Taobao/Tmall platforms, and its detection capability has been demonstrated by identifying a considerable amount of web bot traffic. Based on data samples of traffic originating from web bots and normal users, we conduct a comparative analysis to uncover the behavioral patterns of web bots different from normal users. The analysis results reveal their differences in terms of active time, search queries, item and store preferences, and many other aspects. These findings provide new insights for public websites to further improve web bot traffic detection for protecting valuable web contents.

1 Introduction

Web bots, the programs generating automated traffic, are being leveraged by various parties for a variety of purposes. Web bots are generating a significant volume of web traffic everyday. The 2018 annual report of Distil Networks [2] reveals that web bots account for 42.2% of all website traffic while human traffic makes up the rest 57.8%. The bot landscape is fairly polarized between benign bots and malicious bots [1]. A benign bot mainly refers to a search engine bot that abides by the robot.txt industry opt-in standard and could add value to publishers or advertisers. A malicious bot enables high-speed abuse and attacks on websites. Unsavory competitors and cyber-criminals leverage malicious bots to perform a wide array of malicious activities, such as brute force login, web scraping, adversarial information retrieval, personal and financial data harvesting, and transaction frauds [3, 34].

E-commerce portals is among the sites hit hardest by malicious bots: according to the report [3], about 20% of traffic to e-commerce portals is from malicious bots; malicious bots even generated up to 70% of Amazon.com traffic [4]. As one of the largest e-commerce companies in the world, Alibaba also observed a certain amount of malicious bot traffic to its two main subsidiary sites, i.e., Taobao.com and Tmall.com. In this paper, first, we proposed a novel and efficient approach for detecting web bot traffic. We then implemented and deployed the approach on Taobao/Tmall platforms, and it shows that the detection approach performed well on those large websites by identifying a large set of IP addresses (IPs) used by malicious web bots. Second, we conducted an in-deep behavioral analysis on a sample of web bot traffic to better understand the distinguishable characteristics of web bot traffic from normal web traffic initiated by human users.

In particular, we first presented a bot IP detection algorithm, which consists of two steps: 1) we proposed an Expectation Maximization (EM)-based feature selection method to select the features eligible for determining whether an incoming visitor is a bot; and 2) we proposed to employ a decision tree to combine all the selected features to produce an overall value. We computed a threshold to the decision tree result which optimally recovers the non-bot traffic curve over time. In addition, we dissected one-month long malicious web bot traffic sample and examined the unique behavioral patterns of web bots from normal users.

We analyzed interaction logs containing a one-month time window of randomly sampled visits to Taobao or Tmall from more than 99,000 bot IP addresses (BIPs). Note that bots are unlike normal logged-on users and do not have an associated unique user ID. In addition, a bot may change its IP frequently (e.g., within minutes) and it is impossible to establish a one-to-one relationship between a bot and a BIP. Hence, to be accurate, we consider a BIP rather than a bot as the investigated subject in this work. For a comparative analysis, we also obtained a sample set of more than 97,000 normal users and their interaction logs in the same month.

Our analysis results show that BIPs have unique behavioral patterns and different preferences on items and stores in comparison to the normal logged-on users. Specifically, within the same time period, a BIP could generate 10 times more search queries and clicks than a normal user. Also, a BIP tends to visit the same item multiple times within one day probably for the purpose of periodically monitoring the dynamics of the item. A BIP visits more stores daily than a normal user and prefers to visit the stores with middle or lower reputation grades. By characterizing the malicious bot traffic, we are able to provide e-commerce sites with insights of detecting malicious bot traffic and protect valuable web contents on the e-commerce marketplaces.

The remainder of this paper is organized as follows. Section 2 presents our bot IP detection approach and its application on Taobao/Tmall platforms. Section 3 describes our dataset and presents our investigation results about malicious bot traffic. Section 4 discusses limitations and future work. Section 5 surveys the related work, followed by our conclusions in Section 6.

2 Bot IP Detection Methodology

Our proposed bot IP detection approach consists of two steps. First, we develop an Expectation-Maximization (EM)-based feature extractor to obtain an abnormal score for each IP, and identify suspicious Bot IPs whose abnormal scores are larger than a threshold (2.1). Second, we build a decision tree based on the suspicious label and features of IPs and extract explainable rules from the decision tree (2.2). Furthermore, we demonstrate the effectiveness of our detection approach by applying the resulting rules on Taobao/Tmall platforms (2.3).

2.1 Using EM-based Abnormal Score To Generate Labels

In this section, we develop an EM-based approach and define an abnormal score for each IP.

Intuitively we assume that the distribution of any feature in the candidate pool is a mixture of two different distributions that describe normal traffic samples and suspicious traffic ones, respectively. It is reasonable since the normal traffic samples were generated by normal users from normal IPs while the others are not. With this assumption, the EM algorithm is introduced to estimate the parameters of the two distributions [13]. An IP may be suspicious if the distance between the two distributions is large enough. We present the details of our EM-based modeling and feature extraction procedure as follows.

EM-based Modeling Suppose we have N IPs, a feature of interest (e.g., click-through rate[5]), and a set of corresponding IP-wise values $X = \{x_1, \cdots, x_N\}$. We randomly sampled the same feature of 1,000 IPs in a normal period and in a abnormal period, respectively. We computed the distributions of 1,000 normal feature values and 1,000 abnormal feature values. As shown in Figures 1 and 2, the logarithm of feature values from normal IPs roughly follows a Normal distribution, while the feature values from suspicious IPs nearly follow a mixture of two normal distributions.

[5] Click-through rate is calculated as the total number of clicks on the product detail webpage divided by the number of impressions of the product information in the Taobao/Tmall search engine return results.

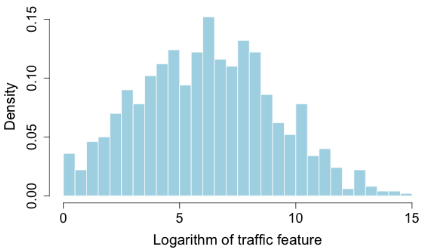

Fig. 1: Distribution of normal traffic of a candidate feature.

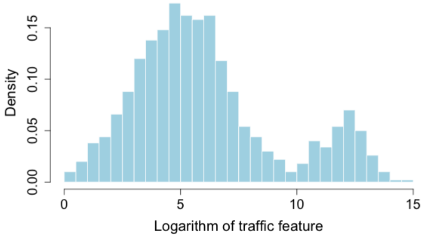

Fig. 2: Distribution of suspicious traffic of a candidate feature.

We define the mixture of two Normal distributions with the density function $p(x)$:

$$p(x|\Theta) = \alpha_1 p_1(x|\theta_1) + \alpha_2 p_2(x|\theta_2), \tag{1}$$

where $p_i(.|\theta_i)$ is a Gaussian density function with parameter $\theta_i = \{\mu_i, \sigma_i\}$, and α_i denotes the non-negative mixture weight and $\alpha_1+\alpha_2 = 1$. Under this assumption, x is from a population composed of two Normal-distributed sub-groups, which can not be observed directly. The sub-group indicator $z_i(x)$ is defined as $z_i(x) = 1$ when the sample x is from the i-th distribution, and therefore $z_1(x) + z_2(x) \equiv 1$. Unless explicitly stated, $z_i(x)$ is simplified to z_i in the latter context.

In this model, one p_i represents the distribution of normal customer behavior while the other describes the suspicious one. The nuisance parameter α_i quantifies the probability whether the sample is from suspicious group or not. The product of all probability density functions (PDFs), according to the expression (1), is the full likelihood under the i.i.d. assumption. Equivalently, the following log-likelihood is used:

$$\log L(X, \Theta) = \log \prod_k p(x_k|\Theta) = \sum_{k=1}^{N} \log p(x_k|\Theta) \quad = \sum_{k=1}^{N} \Big(\log \sum_{i=1}^{2} \alpha_i p_i(x_k|z_i, \theta_i) \Big) \tag{2}$$

This formula could be maximized by the EM algorithm, consisting of three main steps. The EM algorithm repeats the last two steps (i.e., E and M steps) until the convergence criterion is met.

Initialization-step: starting from an initial estimate of θ_i randomly.

E-step: Given the parameters of the distributions, calculate the probability that an IP k comes from distribution i. Denote the current parameter values as $\Theta = \{\mu_1, \mu_2, \sigma_1, \sigma_2\}$. Compute the probability $\omega_{k,i}$ for all IPs k, $1 \le k \le N$ and two mixture components $i = 1, 2$ as

$$\omega_{k,i} = p(z_{k,i} = 1|x_k, \Theta) = \frac{p_i(x_k|z_i, \theta_i) \cdot \alpha_i}{\sum_{m=1}^{2} p_m(x_k|z_m, \theta_m) \cdot \alpha_m} \tag{3}$$

Note that for each IP k, $\omega_{k,1} + \omega_{k,2} = 1$.

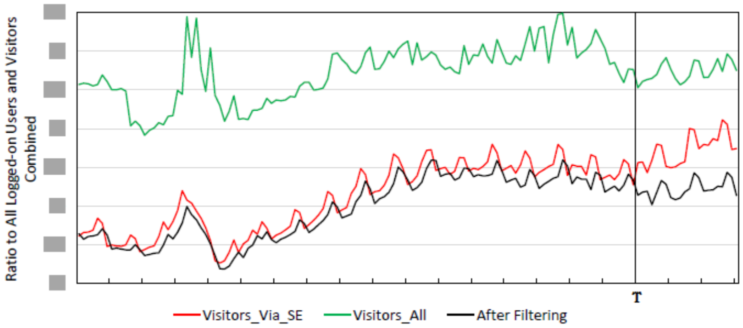

Fig. 3: An abnormally high proportion of unregistered visitors were observed in year 2015. They were then detected and removed by applying our bot detection algorithm on the Alibaba data.

M-step: Given the probabilities calculated in E-step, update the distribution parameters. Let $N_i = \sum_{k=1}^{N} \omega_{k,i}, i = 1, 2$, and we have

$$\alpha_i^{new} = \frac{N_i}{N} \tag{4}$$

$$\mu_i^{new} = (\frac{1}{N_i}) \sum_{k=1}^{N} \omega_{k,i} \cdot x_k \tag{5}$$

$$\sigma_i^{new} = (\frac{1}{N_i}) \sum_{k=1}^{N} \omega_{k,i} \cdot (x_k - \mu_k^{new})^2 \tag{6}$$

Convergence criterion: The convergence is generally detected by calculating the value of the log-likelihood after each iteration and halting when it appears not to be changing in a significant manner from one iteration to the next.

Abnormal Score We consider an IP to be suspicious if the distance between the estimated two distributions is large enough. We define an empirical *abnormal score* of an IP i on feature j as

$$S_{i,j} = \frac{|\mu^{\star}_{i,j,2} - \mu^{\star}_{i,j,1}|}{\max\{|\mu^{\star}_{i,j,1}|, |\mu^{\star}_{i,j,2}|\}} \tag{7}$$

An IP is suspicious if its score $S_{i,j}$ is greater than a certain threshold θ [13].

Threshold Selection To determine the suspicious threshold θ, we resort to the efforts of human experts. We consider two types of visitor traffic (i.e., not-logged-on) to the Taobao/Tmall platforms: visitor traffic coming through the search engine (SE) of the platforms (termed as SE visitor traffic) and visitor traffic taking other ways to enter the platforms. Normally, the ratio of SE visitor traffic to all visitor traffic received by platforms is quite stable. As shown in Figure 3[6], SE visitor traffic (represented by the middle red curve) and all visitor

[6] Both the y-axis and x-axis (denoting the time in 2015) values are deliberately hidden for confidentiality reasons.

traffic (i.e., the top green curve) kept the same pace in growth during consecutive days spanning over a few months in 2015; However, since a changepoint (i.e., a time point, marked as the vertical line T in Figure 3) in 2015, all visitor traffic decreased, but the SE visitor traffic had a significant increase. Web bot traffic could be the major contributor to the abnormal increase in the SE visitor traffic.

By applying EM-based approach on the data of the whole period of time, we can obtain abnormal score $S_{i,j}$ for each IP i and each feature j. To simplify the threshold selection, we define a score for each IP i: $\bar{S}_i = max_j(S_{i,j})$. And we consider an IP is suspicious if $\bar{S}_i > \theta$. Human experts then choose the best value of θ by manually adjusting the threshold to make sure that the trend of the after-filtering-curve (the bottom black one in Figure 3) is more similar to all-visitor-curve (the top green one in Figure 3, especially the end part of that period of time).

2.2 Decision Tree Modeling and Rules' Selection

With the effort of human experts, the EM-based abnormal scores can be used to detect part of the Bot IPs. However, the detection is based on some independent rules, that is, each rule is derived from just one feature. And thus those rules can hardly capture Bot IPs' behaviors. Moreover, the manually-adjusted threshold θ may also result in decrease on the evaluation metric *recall*. We address these problems in three steps:

1. To introduce decision tree, we leverage tens of features to model the suspicious label generated by the EM-based approach.
2. To do feature selection, we conduct cross validation.
3. To generate rules from the resulting decision tree, we adopt the methods introduced in [15].

Following the previous steps, we obtain a list of selected features for the abnormal visitor traffic in 2015, described as follows[7]:

F_1 : the percentage of visits made by non-login users

F_2 : the percentage of clicking on suggested queries

F_3 : the percentage of HTTP requests with empty referer field

F_4 : the total number of search result page view

F_5 : the total number of distinct query keyword divided by F_4

The consequent rules generated from the resulting decision tree can be described in the form of $R_1 \wedge R_2 \wedge R_3 \wedge (R_4 \vee R_5)$ where,

$$\begin{aligned} R_1 &: \ F_1 > 0.9 \\ R_2 &: \ F_2 < 0.1 \\ R_3 &: \ F_3 > 0.7 \\ R_4 &: \ F_4 > 50 \wedge F_5 > 0.9 \\ R_5 &: \ F_4 > 100 \wedge F_5 > 0.7 \end{aligned}$$

[7] When you search a keyword in e-commerce website, the resulting page is the so-called search result page. A search result page view is a page view of search result, for example searching a keyword or going to the next page in search result page.

We note that any derived information (e.g., the thresholds shown above) does not represent the true scenarios in Taobao/Tmall platforms.

2.3 Model Validation

To demonstrate the effectiveness of the model, we validate the generated rules using the data labels generated by both the EM-based approach and online test. With the model, we achieve the *precision* of 95.4% and the *recall* of 92%, which implies that the rules are quite effective in detection of bot IPs.

As for the online test, we deploy the rules on Taobao/Tmall platforms. We observe that in Figure 3 the abnormal increase (represented as the part of red curve since the changepoint T) falls back to the normal black curve after filtering bot IP data.

3 Characterization

3.1 Dataset for Analysis

Table 1: Summary of the dataset.

Visitor	Client Type (%)	Searches (%)	Clicks (%)
BIP (99,140)	PC (99.9%)	92.7%	98.7%
	Wireless (20.0%)	7.3%	1.3%
User (97,109)	PC (63.4%)	17.6%	10.4%
	Wireless (91.4%)	82.4%	89.6%

By broadly sampling the BIPs detected by our bot IP detection approach, we obtained 99,140 BIPs and the associated interaction logs. In addition, we retrieved a sample of 97,109 users and their interaction logs in the same month for comparative analysis. The interaction logs detail the activities conducted by each visitor regardless of whether the visitor is logged on. For a BIP, its activities on an e-commerce site are represented by its search and click behaviors, while a logged-on user may also present transaction-related behaviors such as adding items to cart and checking out.

Loaded on PC or mobile devices. An initial examination of the interaction logs reveals that 99.9% (99,089) BIPs were loaded on the PC devices, which contributed to 92.7% searches and 98.7% clicks while 20% BIPs were loaded on the wireless[8] devices, which generated 7.3% searches and 1.3% clicks. Note that some BIPs may be loaded on both PC and wireless devices. Considering the quite scarce activities presented by BIPs on wireless devices, we focus on the 99,089 BIPs presenting search and click behaviors on the PC clients. Additionally, among the 97,109 logged-on users[9], 63.4% (61,521) were logged on the PC devices and generated 17.6% searches and 10.4% clicks while 91.4% users were logged on the wireless devices and launched 82.4% searches and 89.6% clicks.

[8] We use the two terms “wireless” and “mobile” interchangeably.

[9] Note that not all not-logged-on visitors were deemed as bots by Alibaba IT teams. In addition, a user could be logged on both PC and wireless devices.

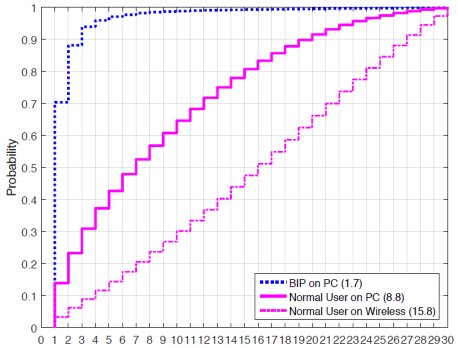

Fig. 4: CDF of active days per BIP and normal user in the same month of the changepoint T. Numbers in parentheses denote the mean values.

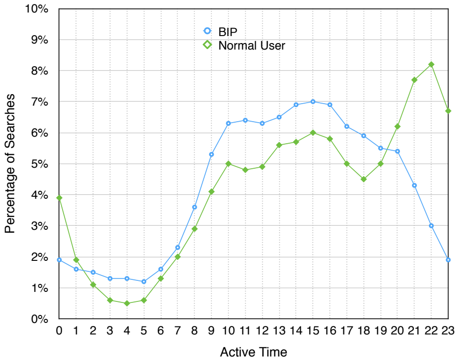

Fig. 5: Distribution of search queries launched by BIPs and normal users during the 24 hours on one day.

The statistical results shown in Table 1 indicate that most BIPs were loaded on PC devices while normal users preferred to browse the shopping sites on the wireless clients such as smartphones and pads. We focus on three kinds of visitors: BIPs on PC, users on PC, and users on wireless.

Next, we attempt to reveal unique browsing patterns and infer the hidden intents of web bots by characterizing each major step of their browsing activities and making comparisons with normal users.

3.2 Browsing Time and Origin

We first examine how many days a BIP was active in the month, their most active time on a day, the number of MIDs (Machine IDentification numbers) used by BIPs during one month, and their origin countries.

Active Days within One Month. Figure 4 depicts the CDF of the days during which a BIP or a user was observed to generate web traffic in the same month of the changepoint T. It shows that about 88% of BIPs were active for only one or two days and the mean value of active days per BIP is 1.7 days. Different than BIPs, logged-on users were active for more days. About 86% of users on PC were active for more than one day, about 48% were active for at least one week, and about 22% active for more than two weeks. The mean value of active days is 8.8. Users on wireless were more active. About 97% of users on wireless were active for more than one day, about 80% were active for at least one week, about 60% active for more than two weeks, and about 30% active for more than three weeks. The mean value is 15.8 days. The results are consistent with the fact that mobile revenue accounts for 65% of core retail business of Alibaba in the quarter ended December 2015 [6].

Takeaway: Most BIPs were observed active for at most two days probably due to the fact that web bots change IP addresses frequently to avoid the detection of their brutal crawling activities.

Active Time on One Day. It is interesting to know at which time BIPs and normal users are most active during one day. We measured the degree of being

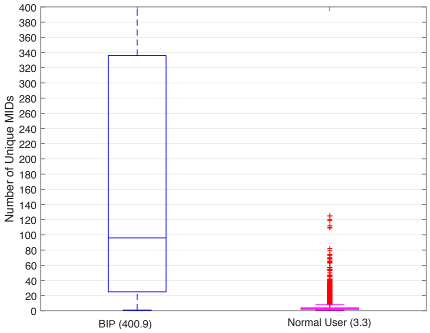

Fig. 6: Number of MIDs used per BIP and normal user within one month. Numbers in parentheses in the x-axis labels denote the mean values.

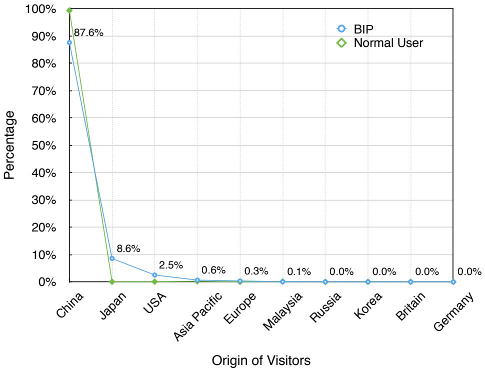

Fig. 7: Distribution of origin countries of BIPs and normal users.

active at a time based on the percentage of search queries made during that time. Figure 5 shows the distribution of search queries made by BIPs and users during 24 hours on one day. Evidently, BIPs and normal users presented different patterns. Normal users were most active during hours 20 to 23, consistent with previous reports [8, 9], and not so active during the working hours between 9 and 19, while BIPs were not so active during the hours from 20 to 23 but quite active during the working hours.

Takeaway: Web bots are not active in the time period (hours 20 to 23) during which normal users are most active, implying that bot developers only run the bots during their working hours.

Number of MIDs Used within One Month. A website usually creates a cookie string for a newly incoming visitor and stores it in the browser cookie file for session tracking. Alibaba's e-commerce sites generate an MID based on a visitor's cookie string for identifying her in the interaction logs. Each time the cookie is deleted by a visitor, the sites would generate a new MID when she returns back. The boxplot in Figure 6 depicts the number of MIDs used per BIP and normal user in the same month of the changepoint T. For each box in the figure, its bottom corresponds to the number of MIDs on the 25th percentile, its top corresponds to the value on the 75th percentile, and the line across the box corresponds to the median value. On average, a BIP was corresponding to up to 401 MIDs within just one month. Given the results shown in Figure 4 that a BIP was observed active in one month for only 1.7 days, we speculate that a BIP may clear its cookies up to hundreds of times each day for evading tracking by the e-commerce sites. By contrast, a normal user was associated with only 3.3 MIDs on average although she was observed active for 8 to 15 days within one month on average, as shown in Figure 4. The result makes sense since a persistent cookie only remains valid during the configured duration period and a new MID would be generated for the user when her cookie becomes invalid.

Takeaway: A BIP may clear its cookies up to hundreds of times a day to avoid tracking.

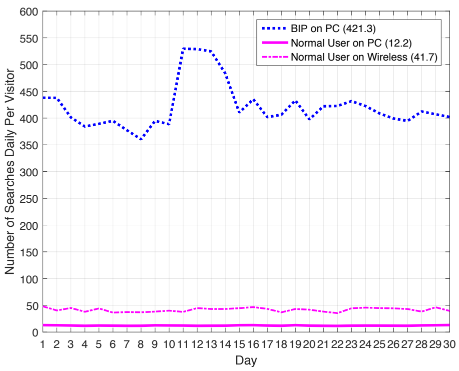

Fig. 8: Number of search queries made daily per BIP and normal user during their active days in the same month of the changepoint T. Numbers in parentheses denote the mean values.

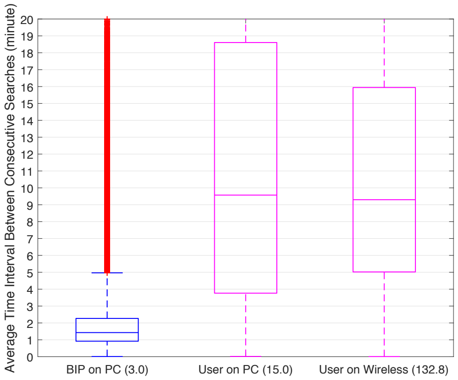

Fig. 9: Average time interval (minutes) between consecutive search queries made by BIPs and normal users. Numbers in parentheses in the x-axis labels denote the mean values.

Origin Country of Visitors. We also compared the origin countries of BIPs to those of normal users in an attempt to identify the origin countries of web bots. Figure 7 depicts the distribution of the origin countries for both BIPs and users. It shows that 99.4% of users were from China, 0.04% were from Japan, and 0.06% from USA. The result makes sense since currently Alibaba keeps its focus on China and Chinese shoppers constitute the majority of its consumers. For the BIPs, 87.6% were from China, 8.6% were from Japan, and 2.5% from USA. Comparatively, the percentages of Japan and USA have risen up.

Takeaway: China, Japan, and USA are the top three countries where bots were launched.

3.3 Statistics of Searches and Clicks

We present the statistics about the searches and clicks made by BIPs and normal users in the one month we investigated.

Daily Number of Searches in One Month. We examined how many search queries submitted daily by BIPs and users during their active days to pinpoint the difference in behavior patterns. Figure 8 depicts the number of search queries made daily per BIP and normal user in the same month of the changepoint T. It shows that BIPs generated an exceptionally large number of search queries on Alibaba e-commerce sites each day. On average, each BIP launched 421.3 search queries daily on the sites. It would be quite unusual if normal users have had made so many queries on Taobao or Tmall for their interested items, since it is quite boring to manually launch hundreds of searches given that each search involves typing keywords and clicking on the search button. In contrast, on average, a normal user on PC generated about 12 search queries daily and a user on wireless made about 42 searches daily, fewer than one tenth of search queries daily made by BIPs. Thus, unlike BIPs, normal users do not search for items excessively. In addition, the result that wireless users made

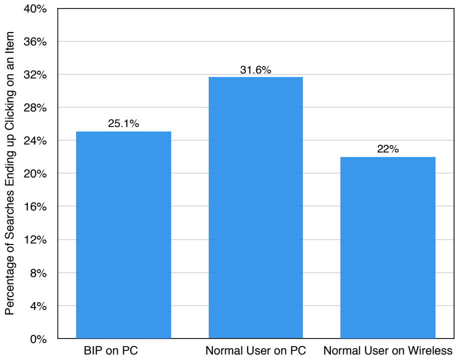

Fig. 10: Percentage of search queries made by BIPs and users that end up clicking on an item.

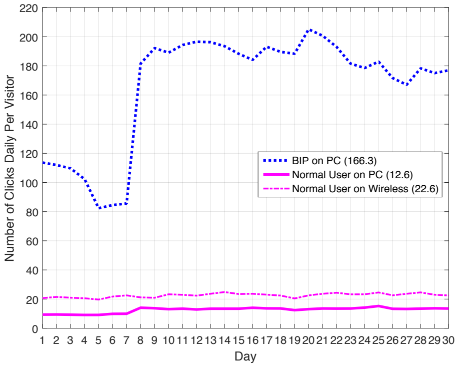

Fig. 11: Number of clicks generated daily per BIP and normal user during their active days in the same month of the change-point T. Numbers in parentheses denote the mean values.

more search queries again confirms that nowadays users prefer online shopping on mobile devices [11].

Takeaway: Bots made about 10 times more search queries daily than normal users.

Time Interval between Consecutive Search Queries. The time interval between consecutive search queries represents the degree of activity. Note that by consecutive search queries, we mean the search queries that happened in sequence on the same day while not necessarily in the same session. Thus the below results represent an upper bound of the time interval between consecutive search queries in one session. Figure 9 depicts the average time interval in minute between consecutive searches made by BIPs and users. Specifically, 25% of BIPs had the time interval of less than 1 minute; 50% of BIPs launched the next search query within 1.5 minutes; and 75% had the time interval of less than 2.2 minutes. In contrast, for users on PC, the median value of the time interval was 9.5 minutes, 75% had the time interval of more than 3.8 minutes, 25% had the interval of more than 18.7 minutes, and the mean value was 15 minutes. Users on wireless had much longer time intervals than BIPs. 75% had the time interval of more than 5 minutes, 50% had the time interval of more than 9.2 minutes and 25% had the time interval of more than 16 minutes. The mean value was 132.8 minutes, exceptionally high due to some outliers.

Takeaway: BIPs behaved much aggressively in launching searches. They had much shorter time intervals between consecutive search queries than normal users, about one fifth of that of the latter on average.

Search Queries Ending up Clicking on an Item. When a search query is submitted, the e-commerce search engine will return back all relevant items. Then the visitor could browse the results pages and choose one or more items to click through to the item detail pages. It is interesting to examine the percentage of search queries that finally lead to further clicks on the items. Figure 10 shows that about 25% of search queries launched by BIPs, 31.6% of search queries by

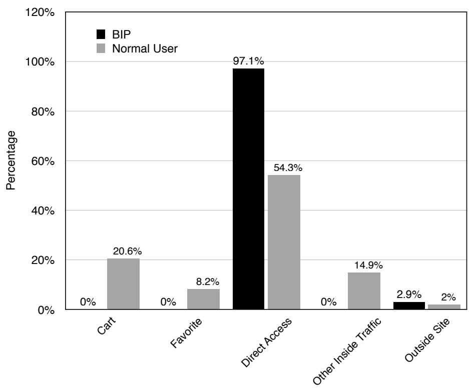

Fig. 12: Breakdown of the paths to the clicks performed by BIPs and users without precedent search queries.

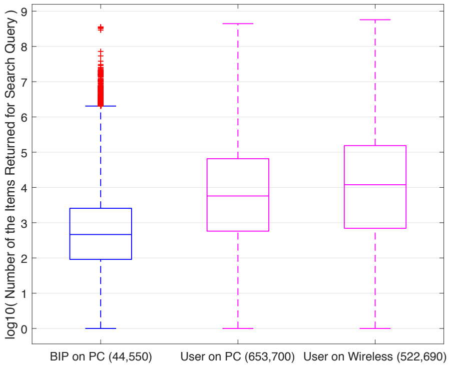

Fig. 13: Number of the items returned for a search query made by BIPs and normal users. Numbers in parentheses in the x-axis labels denote the mean values.

users on PC, and 22% of search queries by users on wireless led to further clicks to the item detail pages. Thus, BIPs and users do not present much difference in this metric.

Daily Number of Clicks in One Month. Figure 11 displays the number of clicks made daily per BIP and normal user. On average, a BIP launched 166.3 clicks daily while on average a normal user on PC performed 12.6 clicks daily and a user on wireless generated a bit more clicks with 22.6 clicks daily.

Takeaway: A BIP performed many more clicks daily than a normal user, about 10 times the clicks made by the latter.

Clicks Without Precedent Searches. Normally, a visitor searches in the e-commerce search engine for the desired items and then clicks on one or more items from the results pages to continue browsing their detail pages. Finally, the visitor chooses an item and adds it to the cart. After that, she may continue to pay online and place an order. Or, she may just leave the platform and return back for checking out several days later, which is also very common. In the latter case, the interaction logs about the visitor for her return would record that the visitor made direct access to the item through her cart or favorite[10] and did not make any precedent search queries.

A statistical analysis of the interaction logs shows that about a half of clicks made by all those three kinds of visitors were not preceded by any search queries. Next we attempt to explore from what path those clicks without precedent search queries were made. Examination of the interaction logs reveals that such clicks were made through one of the following ways: (1) clicking on an item in a cart; (2) clicking on an item in the favorite; (3) direct access to the item detail page through the page URL; (4) access to the item detail page through other traffic directing paths inside the e-commerce site; and (5) access to the item page via advertisements or redirecting from the outside sites. Figure 12 provides a

[10] Note that the favorite here refers to the favorite feature provided by the e-commerce sites, rather than the bookmark features of modern web browsers.

breakdown of the paths to the clicks performed by BIPs and users without precedent search queries. It shows that nearly all (97.1%) of such clicks performed by BIPs were made by direct access to items' detail pages via URLs[11], and the rest clicks (2.9%) originated from the outside sites. Comparatively, 54.3% of such clicks made by normal users were generated by direct access via the detail page URLs, 20.6% of clicks were generated on the carts, 8.2% were generated on the favorites, 14.9% of clicks were the traffic directed by the e-commerce site through other means, and 2% of clicks originated from outside sites probably by clicking through the advertisements displayed on the outside sites.

Takeaway: The results indicate that web bot designers may first obtain a collection of URLs of the item detail pages and then leverage the bots to automatically crawl each item detail page by following the URLs. The reason why normal users also made direct access to the item detail pages via their URLs for about a half of their clicks could be that many users have the habit of saving the detail page links of the interested items to their web browser bookmarks, rather than adding the items to the favorite of the e-commerce sites.

3.4 Returned Results for Search Queries and Subsequent Clicks

For a search query, Alibaba's built-in search engine usually returns tens of thousands of items. Close scrutiny of the number of returned items, the results pages visited, and click position on a result page may reveal distinct patterns of BIPs.

Number of Returned Items for a Search Query. The number of returned items may reflect whether a search query is elaborately made up. Specific search queries are typically responded with limited but desired results. Figure 13 gives a comparison between BIPs and normal users in terms of the returned items for each of their search queries. It shows that BIPs typically received a much smaller number of returned items for each search query. Specifically, for each search query, BIPs got 91 items returned on the 25th percentile, 462 items in the median, 2,565 items on the 75th percentile, and 44,550 items on average. In comparison, users received much more items returned for their search queries either on PC or on wireless. For each search query made by users on PC clients, the number of returned items is 576 on the 25th percentile, 5,731 in the median, 65,356 on the 75th percentile, and 653,700 in the mean. The search queries submitted by users on wireless clients were returned even more items, with the median value of 11,947 and the 75th percentile of 153,765.

Takeaway: Search queries made by BIPs were often responded with much fewer items, more exactly, about an order of magnitude fewer than the items returned for a search query made by normal users. This result could be attributed to two factors: long and complicated search queries, and searching for unpopular items. Combined with previous findings that BIPs usually launch long search queries and tend to search for not so popular items in the e-commerce search engine, one could conclude that BIPs were indeed using long and elaborately

[11] In essence, a direct click on an item is the same as a direct access to the item's detail page via its URL.

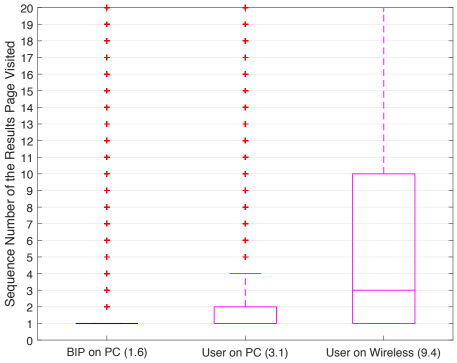

Fig. 14: Sequence number of the results pages visited by BIPs and normal users. Numbers in parentheses in the x-axis labels denote the mean values.

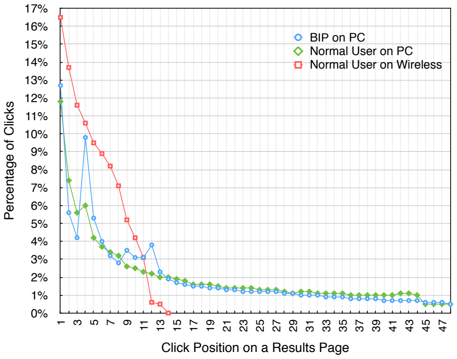

Fig. 15: Distribution of the click traffic on each position of a results page visited by BIPs and normal users.

crafted search queries to crawl data on the Alibaba marketplace. However, their intents are still ambiguous and cannot be quickly determined.

Sequence Number of the Results Pages Visited. Among the results pages returned for a search query, choosing which page to visit is an interesting feature to explore. The boxplot in Figure 14 describes the statistics about the sequence number of the results pages visited by BIPs and normal users. It shows that nearly all BIPs only visited the first results page. Comparatively, in addition to the first results page, normal users may often go further to visit the next pages. For normal users on PC, about one third[12] of their navigations were observed beyond the first results page, and for about 20% visits they navigated beyond the second page. Users on wireless demonstrated even much deeper visits. Specifically, the sequence number of results pages visited had the median value of 3, which means that users on wireless browsed the third results page and/or the deeper pages for half of their visits. The sequence number on the 75th percentile was 10, indicating that for about 25% visits, users on wireless browsed the 10th results page and/or the beyond. It makes sense that users on wireless usually navigate to deeper pages since the number of items listed in each results page is only about 10 for mobile devices due to their small screen sizes while each results page could contain about 50 items on PCs. Thus users on wireless had to navigate more pages for the interested items.

Takeaway: Most web bots only browsed the first results page, indicating that web bots were only interested in the items in the top listings.

Click Positions on a Results Page. A results page usually displays tens of items highly relevant to a search query. A visitor typically browses those displayed items and chooses several of them to click on for further review and comparison before making a purchasing decision. We analyzed all clicks made by BIPs and users on the results pages to examine the distribution of click traffic at each position on a results page. Figure 15 depicts such a distribution for

[12] The number was calculated with the dataset and cannot be inferred from Figure 14.

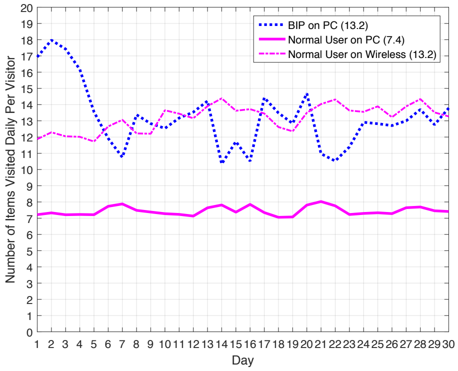

Fig. 16: Number of items visited daily per BIP and normal user. Numbers in parentheses denote the mean values.

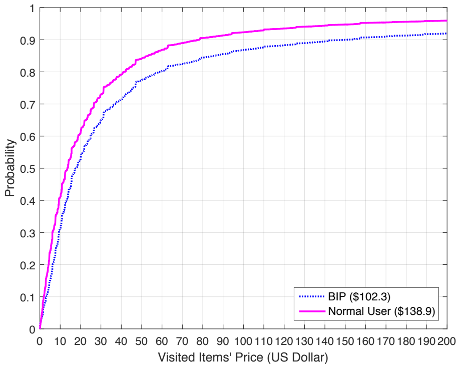

Fig. 17: CDF of the price of the items visited by BIPs and normal users. Numbers in parentheses denote the mean values.

BIPs and users. The figure shows that the items at the first position of results pages received the most clicks: 12.7% clicks of BIPs, 11.8% clicks of users on PC, and 16.5% clicks of users on wireless. Overall, the amount of the received click traffic decreased sharply with the larger ranking positions on the results pages, especially compared with that of the top positions. Compared to the click traffic to an items at the first position, the clicks received by an items at the second position decreased by 7.1% for BIPs, 4.4% for users on PC, and 2.8% for users on wireless. However, some unusual results were also observed. Nearly 10% click traffic from BIPs were directed to the items at the fourth position, significantly larger than the click traffic received by the items at the third or the second positions. We do not exactly know why but it seems that users on PC also preferred to click on the items at the 4th position than the items at the 3rd position. In addition, since a results page on wireless devices usually contains 10 to 20 items, thus it makes sense that the curve representing users on wireless reaches the x-axis at the position 14.

Takeaway: Both web bots and normal users generated the most clicks on the items at the first position on a results page while web bots were also observed to generate a significant proportion of click traffic to the items at the fourth position.

3.5 Visited Items and Sellers

In this part, we characterize the items whose detail pages were viewed by BIPs and normal users, and the stores which accommodate those items.

Number of Items Visited Daily Per Visitor. We first examined how many items were visited[13] by BIPs and normal users. Figure 16 depicts the number of items visited daily per BIP and normal user in the same month of the changepoint T. We found that overall the number of the items visited daily per BIP and normal user is steady. On average, a BIP visited 13.2 items each

[13] An item is deemed visited if its detail page is viewed or retrieved.

day, a user on wireless visited the same number of items per day, and a user on PC visited fewer items each day, with 7.4 items visited. Thus, BIPs did not behave abnormally in terms of the number of items visited each day. However, previous results (Figure 11) show that clicks made daily per BIP was about 10 times the clicks performed daily per normal user. This leads to the conclusion that a BIP may visit one item multiple times per day, more exactly, about 10 times the frequency of a normal user typically visiting an item. And again, users on wireless seem more active than users on PC.

Takeaway: A web bot may visit one item multiple times within one day, probably for monitoring the dynamic of its interested items periodically.

Price of the Items Visited. We also examined the distribution of the price of the items visited by BIPs and normal users, which is depicted in Figure 17. The two curves follow a similar distribution. Both BIPs and normal users showed great interest in the cheap items, and cheaper items received more visits. Specifically, the items with prices less than 10 US dollars were most visited by both BIPs and users, with the occupation ratio of 30% and 40%, respectively. The items with the prices between 10 and 20 US dollars received 20% and 23% visits from BIPs and users, respectively. About 12% and 13% visits from BIPs and users were for the items with the prices between 20 and 30 US dollars. About 78% items visited by BIPs and 84% by users had the prices less than 50 US dollars.

Takeaway: Overall, the items visited by web bots were a bit more expensive than those visited by normal users. Most items listed on the Alibaba marketplace are very cheap and cheap items are much more popular than the expensive ones.

Number of Sellers Visited Daily Per Visitor. Each item belongs to a store. We also attempted to explore the characteristics of the stores accommodating items. We first examined the number of stores visited daily per BIP and normal user. Figure 18 shows that a BIP visited 12 stores daily on average, about twice the number of stores visited daily per normal user. Specifically, a user on wireless visited about 6 stores per day on average. Considering that a user on wireless visited about 13 items per day on average shown in Figure 11, we estimate that a user on wireless may view about two items' detail pages in one store per day on average, more than the number of items visited daily per user on the PC client in one store.

Takeaway: A BIP visited twice the number of stores by a normal user daily.

Reputation Grade of the Stores Visited. Based on the trading volume and positive reviews, Alibaba's e-commerce sites Taobao and Tmall divide stores into twenty grades [10], going from one to five stars, then one to five diamonds, then one to five crowns, and lastly one to five red crowns. A high grade often implies the items on the store sell quite well and receive positive customer reviews. Figure 19 provides a breakdown of the reputation grades of the stores ever visited by BIPs and normal users. It shows that the stores visited most by BIPs had diamond or star grades, representing the middle grades or lower. Specifically, about 55% stores had diamond grades and 25.4% had star grades. In contrast, normal users seemed to have preferences for the stores with middle grades or

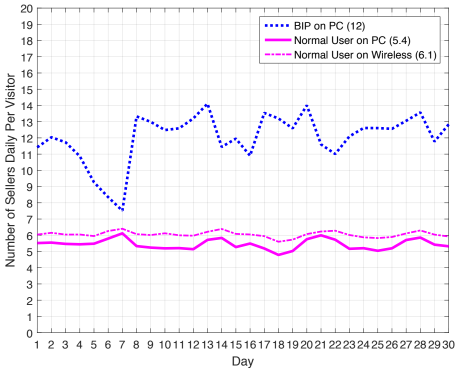

Fig. 18: Number of stores visited daily per BIP and normal user. Numbers in parentheses denote the mean values.

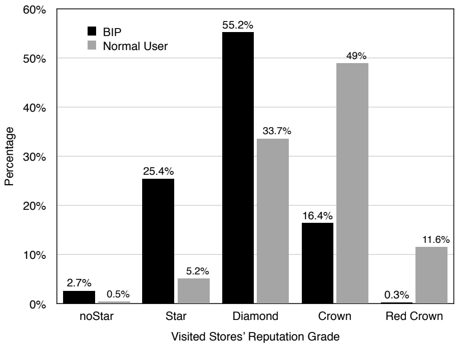

Fig. 19: Breakdown of the reputation grades of the stores ever visited by BIPs and normal users.

higher. Among the stores visited by them, nearly a half had the crown grades and a third had diamond grades. In addition, BIPs and normal users differed markedly on the stores with the lowest and highest grades. A newly open store or a store without a good sales record usually has a grade of less than one star. The figure shows that such stores occupied 2.7% of all stores visited by BIPs and only 0.5% for normal users. The red crown grades represent the highest grades for the stores on the e-commerce sites. We found that 11.6% stores visited by normal users were with red crown grades while only 0.3% stores visited by BIPs had the highest grades.

Takeaway: Web bots preferred to visit the stores with middle reputation grades or lower.

4 Limitations and Future Work

Our current two-step bot detection approach assumes the logarithm of each candidate feature follows a mixture of two Gaussian distributions. Although it has been successfully applied to realistic log data for bot detection, the assumption may not always hold. In addition, the proposed approach involves human experts to ensure accuracy. In the future work, more general methods such as linear mixture model (LMM) and semi/non-parametric (NP) model could be introduced. The LMM assumes the data follows a mixture of several Gaussians and take into account more covariate features. Meanwhile, deep neural network (DNN) and other deep learning methods are proved much powerful for classification tasks. However, neither LMM nor DNN can be directly applied to our problem since they cannot estimate corresponding parameters and make further inferences without positive and negative IP samples detected by our approach. In addition, we cannot disclose what distinguishable features were trained and used for bot detection in this work because of the data confidentiality policy of our partner e-commerce marketplace.

5 Related Work

Our work is closely related to previous work in the areas of Web traffic characterization and automated traffic detection. Ihm et al. [12] analyzed five years of real Web traffic and made interesting findings about modern Web traffic. Meiss et al. [17] aimed to figure out the statistical properties of the global network flow traffic and found that client-server connections and traffic flows exhibit heavy-tailed probability distribution and lack typical behavior. Lan et al. [18] performed a quantitative analysis of the effect of DDoS and worm traffic on the background traffic and concluded that malicious traffic caused a significant increase in the average DNS and web latency. Buehrer et al. [19] studied automated web search traffic and click traffic, and proposed discriminating features to model the physical indicator of a human user as well as the behavior of automated traffic. Adar et al. [20] explored Web revisitation patterns and the reasons behind the behavior and finally revealed four primary revisitation patterns. Goseva-Popstojanova et al. [21] characterized malicious cyber activities aimed at web systems based on data collected by honeypot systems. They also developed supervised learning methods to distinguish attack sessions from vulnerability scan sessions. Kang et al. [31] proposed a semi-supervised learning approach for classifying automated web search traffic from genuine human user traffic. Weng et al. [13] developed a system for e-commerce platforms to detect human-generated traffic leveraging two detectors, namely EM-based time series detector and graph-based detector. Su et al. [14] developed a factor graph based model to detect malicious human-generated "Add-To-Favorite" behaviors based on a small set of ground truth of spamming activities.

Some other previous work focuses on detecting automated traffic, including web bot traffic. Suchacka et al. [22] proposed a Bayesian approach to detect web bots based on the features related to user sessions, evaluated it with real e-commerce traffic, and computed a detection accuracy of more than 90%. McKenna [23] used honeypots for harvesting web bots and detecting them, and concluded that web bots using deep-crawling algorithms could evade their honeypots-based detection approach. To address the issue of web bots degrading the performance and scalability of web systems, Rude et al. [24,26] considered it necessary to accurately predict the next resource requested by a web bot. They explored a suite of classifiers for the resource request type prediction problem and found that Elman neural networks performed best. Finally, they introduced a cache system architecture in which web bot traffic and human traffic were served with separate policies. Koehl and Wang [28] studied the impact and cost of the search engine web bots on web servers, presented a practical caching approach for web server owners to mitigate the overload incurred by search engines, and finally validated the proposed caching framework. Gummadi et al. [29] aimed to mitigate the effects of botnet attacks by identifying human-generated traffic and servicing them with higher priority. They identified human-generated traffic by checking whether the incoming request was made within a small amount of time of legitimate keyboard or mouse activity on the client machine.

Jamshed et al. [30] presented another effort on suppressing web bot traffic by proposing deterministic human attestation based on trustworthy input devices, e.g., keyboards. Specifically, they proposed to augment the input devices with a trusted platform module chip. Goseva-Popstojanova et al. [21] characterized malicious cyber activities aimed at web systems based on data collected by honeypot systems. They also developed supervised learning methods to distinguish attack sessions from vulnerability scan sessions. Kang et al. [31] proposed a semi-supervised learning approach for classifying automated web search traffic from genuine human user traffic. Comparatively, we present an EM-based feature selection and rule-based web bot detection approach, which is straightforward but has been evaluated to be effective.

One main goal of web bot traffic to the e-commerce sites could be to infer the reputation system and item ranking rules in use, which could be then manipulated by insincere sellers to attract buyers and gain profits. One work [32] reported the underground platforms which cracked the reputation systems and provided seller reputation escalation as a service through by hiring freelancers to conduct fraud transactions. In addition, Kohavi et al. [33] recommended ten supplementary analyses for e-commerce websites to conduct after reviewing the standard web analytics reports. Identifying and eliminating bot traffic was suggested to be done first before performing any website analysis. This also justifies the value of our work.

6 Conclusion

Web bots contribute to a significant proportion of all traffic to e-commerce sites and has raised serious concerns of e-commerce operators. In this paper, we propose an efficient detection approach of web bot traffic to a large e-commerce marketplace and then perform an in-depth behavioral analysis of a sample web bot traffic. The bot detection approach has been applied to Taobao/Tmall platforms and performed well by identifying a huge amount of web bot traffic. With a sample of web bot traffic and normal user traffic, we performed characteristic analysis. The analytical results have revealed unique behavioral pattens of web bots. For instance, a bot IP address has been found to stay active for only one or two days in one month but generate 10 times more search queries and clicks than a normal user. Our work enables e-commerce marketplace operators to better detect and understand web bot traffic.

Acknowledgement. We would like to thank the anonymous reviewers for their valuable feedback. This work was partially supported by the U.S. NSF grant CNS-1618117 and DARPA XD3 Project HR0011-16-C-0055.

References

1. Good bots are going too far. `https://goo.gl/uVjvs3`
2. Distil Networks: The 2018 Bad Bot Report. `https://goo.gl/Ysmz34`
3. Distil Networks: The 2016 Bad Bot Report. `https://goo.gl/76T3YQ`
4. Distil Networks: The 2015 Bad Bot Report. `https://goo.gl/duH5Dy`
5. Hubert W Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. In *American Statistical Association Journal*, 1967.

6. Alibaba Group Quarterly Report. `https://goo.gl/LLehdY`
7. Apparel: Most Popular Category on Alibaba. `https://goo.gl/2KfSVq`
8. Taobao Users' Shopping Habits in 24 Hours. `https://goo.gl/za6EB2`.
9. Taobao Online Shoppers Behavior. `https://goo.gl/YgiJL1`.
10. Sellers' Reputation Grade on Alibaba. `https://goo.gl/HBQ9MT`
11. Alibaba Group's September Quarter 2015 Results. `goo.gl/25X7JN`.
12. S. Ihm, and V.S. Pai. Towards Understanding Modern Web Traffic. In *IMC* 2011.
13. H. Weng, Z. Li and et al. Online E-Commerce Fraud: A Large-scale Detection and Analysis. In *ICDE* 2018.
14. N. Su, Y. Liu and et al. Detecting Crowdturfing "Add to Favorites" Activities in Online Shopping. In *WWW* 2018.
15. J. R. Quinlan Generating production rules from decision trees. In *IJCAI* 1987
16. L. Breiman. Classification and regression trees. Routledge 2017.
17. M. Meiss, F. Menczer, and A. Vespignani. On the Lack of Typical Behavior in the Global Web Traffic Network. In *WWW* 2005.
18. K. Lan, A. Hussain, and D. Dutta. Effect of Malicious Traffic on the Network. In *PAM* 2003.
19. G. Buehrer, J.W. Stokes, and K. Chellapilla. A Large-scale Study of Automated Web Search Traffic. In *Airweb* 2008.
20. E. Adar, J. Teevan, and S.T. Dumais. Large Scale Analysis of Web Revisitation Patterns. In *CHI* 2008.
21. K. Goseva-Popstojanova, G. Anastasovski, A. Dimitrijevikj, R. Pantev, and B. Miller. Characterization and Classification of Malicious Web Traffic. In *Computers and Security* 2014.
22. G. Suchacka, and M. Sobków. Detection of Internet robots using a Bayesian approach. In *IEEE 2nd International Conference on Cybernetics (CYBCONF)* 2015.
23. S. F. McKenna. Detection and classification of Web robots with honeypots. In *Naval Postgraduate School* 2016.
24. H. N. Rude. Intelligent Caching to Mitigate the Impact of Web Robots on Web Servers. In *Wright State University* 2016.
25. D. Doran, and S.S. Gokhale. Discovering new trends in web robot traffic through functional classification. In *Seventh IEEE International Symposium on Network Computing and Applications* 2008.
26. H.N. Rude, and D. Doran. Request type prediction for web robot and internet of things traffic. In *ICMLA* 2015.
27. E. Pujol, P. Richter, B. Chandrasekaran, G. Smaragdakis, A. Feldmann, B.M. Maggs, and K. Ng. Back-office web traffic on the internet. In *IMC* 2014.
28. A. Koehl, and H. Wang. Surviving a search engine overload. In *WWW* 2012.
29. R. Gummadi, H. Balakrishnan, P. Maniatis, and S. Ratnasamy. Not-a-Bot: Improving Service Availability in the Face of Botnet Attacks.. In *NSDI* 2009.
30. M.A. Jamshed, W. Kim, and K. Park. Suppressing bot traffic with accurate human attestation. In *Proceedings of the first ACM asia-pacific workshop on Workshop on systems* 2010.
31. H. Kang, K. Wang, D. Soukal, F. Behr, and Z. Zheng. Large-scale Bot Detection for Search Engines. In *WWW* 2010.
32. H. Xu, D. Liu, H. Wang, and A. Stavrou. E-commerce Reputation Manipulation: The Emergence of Reputation-Escalation-as-a-Service. In *WWW* 2015.
33. R. Kohavi, and R. Parekh. Ten Supplementary Analyses to Improve E-commerce Web Sites. In *SIGKDD Workshop* 2003.
34. C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas. DDoS in the IoT: Mirai and other botnets. In *Computer 50, no. 7 (2017): 80-84*.