

BotGraph: Web Bot Detection Based on Sitemap

Yang Luo¹, Guozhen She^{1,2}, Jinwan Huang^{1,3}, Peng Cheng¹ and Yongqiang Xiong¹

¹Microsoft Research

²Fudan University

³Beihang University

{yangluo, v-gushe, t-jihua, pengc, yqx}@microsoft.com

Abstract

The web bots have been blamed for consuming large amount of Internet traffic and undermining the interest of the scraped sites for years. Traditional bot detection studies focus mainly on signature-based solution, but advanced bots usually forge their identities to bypass such detection. With increasing cloud migration, cloud providers provide new opportunities for an effective bot detection based on big data to solve this issue. In this paper, we present a behavior-based bot detection scheme called BotGraph that combines sitemap and convolutional neural network (CNN) to detect inner behavior of bots. Experimental results show that BotGraph achieves $\sim 95\%$ recall and precision on 35-day production data traces from different customers including the Bing search engine and several sites.

1 Introduction

According to Incapsula’s report [Incapsula, 2016], about 51.8% of Internet traffic in 2016 are performed by automatic bots instead of human. These bots include search engines, price scrappers, Email harvesters and even Trojan which could launch DDoS attacks. These bots not only cause the leakage of business data, but also consume significant bandwidth and server overload. For the past few years, as more businesses migrate their websites to clouds, it becomes the responsibility of the cloud provider to offer an effective bot mitigation solution for its customers.

Based on whether authentication is required, most commonly seen bots can be partitioned into two broad categories: social bots, which target for social networks and web bots, which target for general websites. Compared to social bots, detecting web bots is more challenging because of two reasons. First, it is difficult to identify a website user (or bot) as there is no concept like the user account (or hard to retrieve it as a cloud provider) in the web traffic. Leveraging the client IP address seems to be a feasible method for user identification. Nevertheless, it can be faked easily via proxy. Second, social bot detection can be customized and tuned for a specific social media. However, for web bots, especially as a cloud provider, there would be millions of websites hosted

in the cloud. Each site provides distinct services to its customers. Recognizing the bots among all the web traffic for all sites requires a universal scheme that works for heterogeneous scenarios. In this paper, we focus on the detection of web bots. We use the term “bot” to refer to web bots in following sections.

There are several traditional ways to perform bot detection, e.g., UserAgent blacklist, IP rate limiting, device fingerprint recognition, etc. However, maintaining IP or UserAgent blacklist requires huge effort to maintain the blacklist database. Moreover, a bot can easily use a proxy IP address or modify its UserAgent to a normal browser. Detecting device fingerprint such as mouse movement and JavaScript engine validation is often a better way, but it usually relies on client-side JavaScript code, which is an invasive technique. Unfortunately, advanced bots can still bypass such detection by utilizing real browser environments like headless Chrome. In summary, all these methods rely on bot’s identities or feature codes, which can be easily bypassed by advanced bots via faking their identities. Detecting bots via their behaviors instead of their identities would be a better way.

In this paper, we categorize the features of web traffic into two types: identity features and behavior features. Then we introduce a behavior-based bot detection approach called BotGraph. BotGraph is performed in three steps. First, we define the concept of sitemap and propose three ways to build the sitemap for a site. Second, each user session is mapped to a subgraph of the sitemap. The subgraph contains information about which URL patterns the client has visited and the corresponding access frequencies. Third, a 2-dimensional image is generated from the above subgraph. Thus the task of bot detection has become an image classification problem. We use the state-of-the-art techniques like CNN to classify the images into two categories: bot or non-bot. We evaluate BotGraph on various datasets including Bing search engine and several sites from different industries. The result shows that our method can achieve $\sim 95\%$ precision and recall on most of the datasets.

The remainder of this paper is organized as follows. Section 2 elaborates on the related work. Section 3 presents our behavior-based bot detection scheme called BotGraph. Section 4 brings the experimental results. Section 5 discusses about our drawbacks. Section 6 concludes the paper.

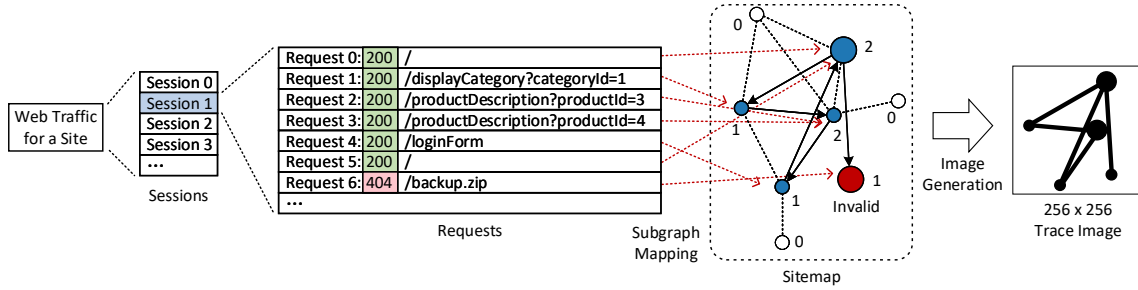


Figure 1: Architecture of BotGraph.

2 Related Work

Bot detection is highly demanded for both website owners and service providers. Google Analytics [Google, 2019] provides built-in filters to help their customers filter bot requests based on geolocation information of the client IP address. Distil, Akamai and ShieldSquare are selected as three leaders in the bot detection area, based on the whitepaper published by Forrester [Forrester, 2018]. Akamai [Akamai, 2018] uses pre-defined bot signature database as well as a legitimate service whitelist. It also allows its users to customize bot detection rule. ShieldSquare [ShieldSquare, 2019] provides both identity-based and behavior-based bot detection. The identity-based method utilizes client-side JavaScript to collect parameters like browser fingerprints. The behavior-based analysis is based on characteristic in terms of number of pages visited per session, time spent on each page, frequency of repeat visits, and so on. This is similar to our solution, but we describe user behaviors in a graph based on sitemap instead, which contains more unique features for a client. Distil [Networks, 2018] provides several bot detection methods such as known violator blacklist, biometric data validation like finger swiping and mouse movement. HTTP’s UserAgent is used to recognize the category of the bots. They also provide a machine-learning based method, which needs to be trained for one week before being ready to use. However, UserAgent-based bot detection is not feasible for advanced bots, as they can easily hide themselves by modifying its value.

For feature extraction, some network intrusion datasets [KDD, 1999; CSIC, 2009; UNB, 2014] are provided as the ground truth, some methods were brought out based on which, A review [Jaafar *et al.*, 2019] introduced a method which encode the numerical features into holistic metrics like total requests, standard deviative time and the percentage of POST requests, a topic-based model latent dirichlet allocation (LDA) was introduced [Lagopoulos *et al.*, 2017] by Athanasios to encode the semantic information like words in the URLs and postfix representing the type of target resource(eg. html, pdf, asp) into digit feature vectors by introducing the statistic-based concepts like topic variance and topic similarity. Besides that, auto-encoder model [Zong *et al.*, 2018] is proposed to facilitate the feature extraction procedure. Deep-Defense [Yuan *et al.*, 2017] introduced an algorithm which splitted the traffic logs into several segments with the same

shape, then encoded the request information in each line of segment to numerical matrix, which would then be fed to recurrent neural network (RNN) based model. Similar to BotGraph, this model can capture the context information among requests. However, the inference efficiency of RNN model highly relies on the length of segments, while BotGraph can perform much more stable.

For detection methods, an improved support vector machine (SVM) algorithm is [Schölkopf *et al.*, 1999] proposed as the general approach, More recently, then a boost method called XGBoost [Chen and Guestrin, 2016] is applied to improve both the efficiency and accuracy of detection. Inspired by the trend towards deep learning, the energy-based deep learning model [Zhai *et al.*, 2016] showed up, and GAN-based model [Zenati *et al.*, 2018] also dabbled in the scope of anomaly detection.

Besides the CNN model used in this work, we also investigated using graph convolutional network (GCN) [Kipf and Welling, 2016] to train the client’s behavior in the web traffic. GCN is different with CNN by accepting structured graphs instead of images as input. However, it is not suitable for our scenario as it can only perform node classification inside one graph instead of classifying multiple graphs.

Overall, most previous researches rely on identity features such as client IP address or UserAgent for bot detection, which can be easily defeated by advanced bots. A behavior-based bot detection method should be proposed to differentiate advanced bots from normal users.

3 Our Approach

3.1 Overview

The architecture of BotGraph is shown in Figure 1. There are three steps: first, we need to build the sitemap for the site. We provide three ways to do this. Second, we map the requests in a session to a subgraph of the sitemap. Third, we generate the 2-dimensional trace image from the subgraph, which transforms the bot detection task into an image classification problem. Finally, we use CNN-based model to classify those generated images into two categories: bot and non-bot. We will provide the details of BotGraph as follows.

3.2 Basic Concepts

Request. In this paper, we use the term “request” to represent the HTTP request from the client to the server (aka

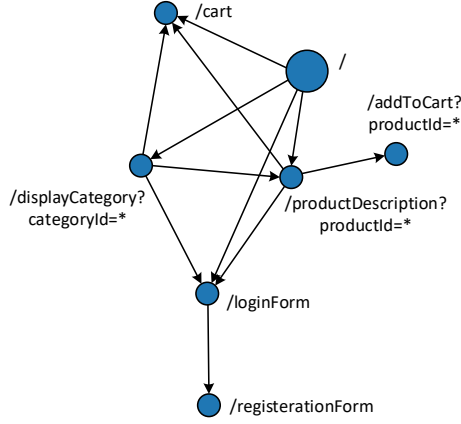


Figure 2: Sitemap of a simple online shopping site.

Field	Description
Timestamp	Request time, e.g., 2019-1-12 04:00:07
HttpMethod	HTTP request method, e.g., GET, POST.
RequestUri	The path in URL, e.g., /books/desc?id=1
Status	HTTP status code, e.g., 200, 404.
Host	“Host” field in request header.
UserAgent	“User-Agent” field in request header.
ClientIp	Client’s IP address.

Table 1: Request fields.

the site) together with the corresponding response. A request usually contains many fields, a portion of which is shown in Table 1.

Session. Bots usually scrape the pages with a large number of requests. To describe such a behavior, a necessary pre-processing step is partitioning the requests into sessions. A session identifies a unique client (normal browser or bot) that performs the accesses.

3.3 Features

For the bot detection task, there are two types of fields in a request: identity fields and behavior fields.

Identity fields. Identity fields are used to identify the client or server. Specifically, *UserAgent*, *ClientIp* are identity fields for the client. *Host* is the identity field for the server.

Behavior fields. Behavior fields are used to describe the access behavior of the client, including fields like *RequestUri* and *Status*.

Currently, identity fields are widely used in traditional bot mitigation schemes such as IP rate limiting and UserAgent blacklisting. However, if the bots fake their identity through using IP proxy pool, or tamper its *UserAgent*, those methods would fail. Therefore, a feasible way would be detecting the bots via their behavior instead of their identity. The behavior fields used in BotGraph are as follows:

RequestUri, Status. These two fields play the central role in describing a bot’s behavior. We map *RequestUri* of each request into a sitemap node. *Status* is used to determine whether

it is a valid mapping. The details would be discussed in Section 3.5.

As BotGraph detects bots on a per-session basis, as input, we assume the requests for a site are grouped into sessions and sorted by *Timestamp*. Next we will introduce how we perform bot detection based on sitemap and CNN.

3.4 Sitemap Retrieval

Google first introduced the Sitemaps protocol to describe a site’s content [Google, 2005]. A sitemap is a XML file that lists all the URLs for a site. It allows search engines to crawl the site more efficiently. In this paper, we extend the list-formatted sitemap into a graph. The sitemap for a site is defined as: $G = (V, E)$, in which:

- G : a directed graph.
- V : set of nodes. Each node represents a URL pattern incorporating multiple URLs. For example, Both */page?id=1* and */page?id=2* belong to the same pattern: */page?id=**.
- E : set of directed edges from one node to another. Take two nodes: v_1 and v_2 as example, if the HTML content of the web page with pattern v_1 has one or more hyperlink (typically via HTML $\langle a \rangle$ tag) pointing to a URL of pattern v_2 , then we say v_1 has an edge to v_2 .

We show an example of sitemap for a simple online shopping site [Shah, 2016] in Figure 2. This sitemap shows the basic functionality of the site, including registration, login, product view, cart, etc.

There are several ways to build the sitemap for a site, including active crawling, passive sniffing and self-providing.

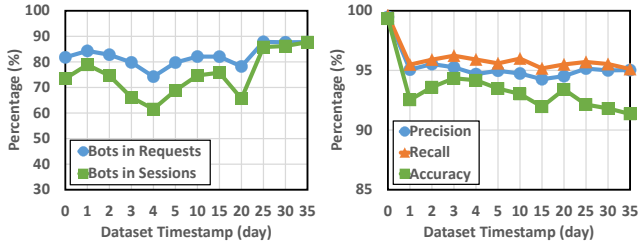
Active crawling. Active crawling requires to run a crawler to build the sitemap of the site. The crawling typically starts from the homepage and enters each hyperlink from the current page recursively. Each URL pattern is retrieved only once to reduce the number of pages need crawling. This is based on the assumption that web pages with same URL pattern have the same page structure and similar hyperlinks of the same URL patterns.

Passive sniffing. In passive sniffing, the URLs of site’s traffic are monitored, learned and then used to build the sitemap. This scheme is less intrusive than active crawling. However, the sitemap may be incomplete limited by the amount of sniffed traffic.

Self providing. The site provides its own sitemap for bot detection. By this way we could gain the most precise sitemap, but requires non-trivial work from the site owner.

3.5 Subgraph Mapping

As shown in Figure 1, we can map *RequestUri* of each request in a session into a node in the sitemap (in blue color). For each node in sitemap, we use the term “access frequency” to indicate the number of requests mapped to it. Moreover, two adjacent requests in the session can determine an edge in the sitemap (in solid line). Those mapped nodes and determined edges form a subgraph of the original sitemap. The



(a) Number of bots. (b) Precision, recall and accuracy.

Figure 3: Performance on search engine dataset spanning 35 days.

Site	#Nodes	#Edges	How Is It Retrieved
Search engine	542	73441	Self providing
News site	51	1473	Active crawling
University site	134	2498	

Table 2: Sitemaps of our datasets.

generated subgraph contains information about the URL patterns the client has visited as well as the corresponding access frequencies.

There are occasions in which non-existent URLs are accessed and status code 404 is responded (see Request 6 in Figure 1). This could be blamed for the bot’s attempts to access a previously crawled and cached URL which has already been removed, or just brute-force attacks against vulnerable URLs like `/backup.zip` and `/apmserv5.2.6.rar`. These URLs cannot be mapped to any node in the sitemap. To resolve this, a node called “INVALID” is manually added to the sitemap as a container of all non-existent URLs.

3.6 Image Generation

In this section, we discuss about how to generate the trace image from a given subgraph in the section above. There are two kinds of elements in the trace image: spot and line. The procedure of image generation is described as follows: First we draw the black-filled spot for each node in the subgraph, the central coordinate and radius of which will be dwelt on later. Then we draw the straight line to bridge each pair of nodes containing an edge in original subgraph, the width of which is a session-free constant.

Page Affinity

To determine the position of each node in the sitemap, we use the Verlet algorithm [Verlet, 1967] to generate their coordinates. The Verlet algorithm performs molecular dynamics simulation based on Newton’s equation of motion. The intuition is that a link between two nodes in the sitemap generates an attractive force, which tends to make them nodes closer. The algorithm use the iteration formula to reach the final balanced states for all nodes. In this way, the affinity of original web pages is visually exhibited on the generated sitemap.

It is notable that the coordinates are generated only once for a sitemap (site) and shared by all the subgraphs derived from this sitemap, which would not cause significant overhead when processing lots of sessions.

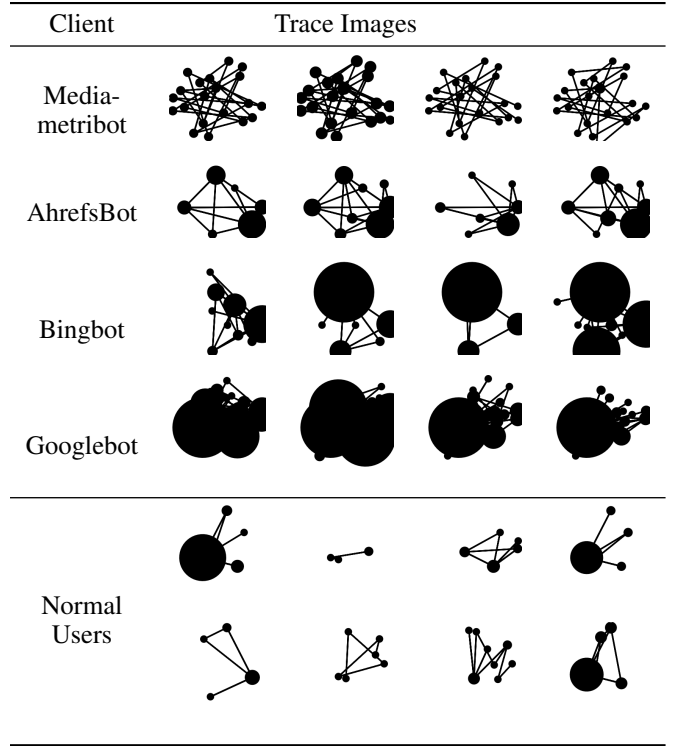


Table 3: Trace images of different sessions.

Access Frequency

The access frequency for each URL pattern is an important feature for bot detection as a bot usually needs to access certain type of pages repeatedly, e.g., scraping product description pages like `/product?id=*`. We use the spot radius to represent such access frequency for a sitemap node. Define $r = f(x)$, r is the radius of the spot, x is the access frequency. $f(x)$ is supposed to meet the following rules:

- Higher access frequency results in larger node. So $f(x)$ is an increasing function.
- The smallest node (accessed only once) should also be visible in the image. Thus we have $f(1) = r_{min}$.
- The largest node should not occupy too much space of the image. Thus we have $f(+\infty) = r_{max}$.
- The gradient of $f(x)$ should be gentle when x is relatively small. We can use $f(x_{gate}) = r_{gate}$ to restrict it. The $gate$ is a chosen value.

Inspired by the sigmoid function, we design our own $f(x)$ as follows. Given $r_{min}, r_{max}, x_{gate}, r_{gate}$, the parameters of a, b, c can be determined by solving the constraints above.

$$f(x) = \frac{c}{1 + e^{b-ax}} \quad (1)$$

4 Evaluation

4.1 Dataset

To evaluate BotGraph, we use datasets from real world web server logs including Bing search engine and several sites

Dataset	Type	Requests		Sessions		Precision (%)	Recall (%)	Accuracy (%)
		#Total	BoR (%)	#Total	BoS (%)			
Search engine (d: 0, hr: 0)	Train	8212838	81.8	163811	73.5	99.5	99.6	99.3
Search engine (d: 5, hr: 0)	Test	7541850	79.8	153464	68.8	95.0	95.6	93.5
News site (d: 0-3)	Train	48606	78.4	4723	49.2	98.4	98.4	97.4
News site (d: 4-7)	Test	52343	79.2	5070	51.6	98.0	96.9	95.7
University page (hr: 0-5)	Train	250000	6.7	3837	7.5	97.8	93.8	99.4
University page (hr: 6-15)	Test	250000	1.4	3754	4.7	95.3	69.1	98.4

Table 4: Performance on different datasets. *BoR* indicates the bot’s percentage of the total requests. *BoS* indicates the bot’s percentage of the total sessions.

Device	Type	TH (session/s)	LA (ms)
Intel Xeon E5620	CPU	3.67	103.60
Intel i7-8750H		8.76	50.44
Intel i7-7700		9.06	48.81
GTX 1050 Ti	GPU	132.25	5.97
Tesla K40		220.84	1.62

Table 5: Efficiency on the news site dataset under different CPUs and GPUs. *TH* is training throughput. *LA* is inference latency.

Scheme	Precision (%)	Recall (%)	Accuracy (%)
SVM	82.4	32.6	97.9
XGBoost	71.0	92.5	68.9
AdaBoost	71.0	92.5	68.9
DT	68.8	100.0	68.8
RF	68.8	100.0	68.8
MLP	73.8	87.8	70.2
LSTM+	35.0	81.4	95.1
SVM+	75.8	94.5	75.5
XGBoost+	86.8	98.9	88.8
AdaBoost+	85.4	96.7	86.3
DT+	85.9	98.7	87.9
RF+	83.8	96.5	84.8
MLP+	84.0	88.6	80.6
BotGraph	95.1	95.5	92.5

Table 6: Comparison with other bot detection methods on the search engine dataset. The trailing + indicates the *clientIp* field is used as a feature.

from different industries like news site, university homepage, etc. As shown in the *Dataset* and *Type* columns, the training set of the search engine dataset is collected for a hour on day 0. The testing set is for the same hour on day 5. These logs have already been sessionized by tracking the *SessionId* cookie of the client. Each session is labeled as *bot* or *non-bot*. The dataset only includes sessions the images of which have more than 3 spots, which will be explained in Section 5. The labeling is performed as such: a team of 30+ professional engineers manually analyzed the traffic and used various ways including JavaScript support checking, mouse movement and click tracking, IP reputation, UserAgent blacklisting to label the traffic. In this paper, the labels are assumed to be accurate and used as ground truth.

4.2 Setup

The sitemaps for each dataset are shown in Table 2. The search engine dataset provides a *PageName* field, e.g., *Home*, *Page.Serp*, *Page.NoResults*, *Page.Image.Results*, etc. So we directly use it as the node in sitemap. Some random edges are generated to make most of the nodes connected in the sitemap. For other datasets, we used a crawler to actively scrape their sitemaps. It is notable that the sitemap edges have nothing to do with the subgraph edges. The sitemap edges are site-wise and only used as input of Verlet algorithm to generate the nodes’ coordinates. However, subgraph edges are session-wise and used as a feature in the generated trace images.

We use 256×256 as the image size. A padding of 5% is added to the image’s four sides to ensure the spots cannot exceed the canvas easily. We empirically use $r_{min} = 4$, $r_{max} = 80$, $x_{gate} = 50$, $r_{gate} = 50$ in the access frequency function $f(x)$. The parameters of $f(x)$ can be solved accordingly.

To classify the trace images, we tried different CNN models, including LeNet-5, AlexNet, ResNet, etc. They all get similar precision and recall. Thus we choose the fastest LeNet-5, a 7-level CNN to train and inference. The trace images are used as input. The output is a scalar, indicating bot or non-bot. We use the following hyperparameters in our experiments: *BatchSize* = 64, *Epoch* = 100, *LR* = 0.01, *SGDMomentum* = 0.5. Our training code is open sourced at: <https://github.com/botrainer/botrainer>.

4.3 Performance

Although a bot can easily modify its *UserAgent* to pretend a normal browser, A client with *UserAgent* claiming to be a bot















Category	Trace Images						
False Positives							
False Negatives							

Table 7: Trace images of false positives & false negatives.

is usually for real. So we use the claimed *UserAgent* of each trace image (aka session) as a group key and show several groups of randomly selected trace images in Table 3. We find these images have described the client’s behavior pretty well in the following two aspects:

1. The trace images of the same bot share high similarity. Different bots have distinct image patterns.
2. The trace images of normal users have different shapes, which are usually not the same as those of bots.

The performance of BotGraph on different datasets is shown in Table 4. besides precision, recall and accuracy, we also present two interesting metrics which are related to the bot detection result: bot’s percentage of requests (*BoR*) and bot’s percentage of sessions (*BoS*). *BoR* and *BoS* are usually not the same value but highly related. When $BoS \geq 49\%$, we usually have $BoR > BoS$. This is because the session length (number of requests) of a bot is larger than that of a normal user on average. So when bots are significant in a traffic, this pattern is more obvious. However, when $BoS < 10\%$, the existing bots are usually unorganized and with no harmful intention, like Googlebot, Bingbot, etc. They do not crawl very large number of pages in a session, which is not significantly different from normal users. For the datasets with $BoS \geq 49\%$, BotGraph achieves higher than 95% precision and recall. It means when a site is heavily affected by bots, BotGraph can effectively detect those bot traffic. For the datasets with $BoS < 10\%$, BotGraph still gets higher than 95% precision with nearly 70% recall. We think it is because the bot traffic in such sites are scattered and have no stable patterns, which influences our effect.

Our CNN-based model for image classification is implemented in PyTorch. We benchmarked the training and inference performance in different circumstances on the previously mentioned news site’s dataset, as shown in Table 5. We can see that under an ordinary GPU, BotGraph does not cause obvious latency compared to the common round-trip delay of 50~100 ms on Internet.

4.4 Comparison

We compared BotGraph with other bot detection methods like long short-term memory (LSTM), SVM, XGBoost [Lagopoulos *et al.*, 2017], AdaBoost, decision tree (DT), random forest (RF), multi-layer perceptron (MLP), etc. Some methods are proposed in previous work and some others are

implemented by ourselves. The experiment is done on the dataset of news site, as shown in Table 6. we also present the request fields that each method uses. We use the following feature engineering approach for each field: *UserAgent* is parsed into tuple (*Browser*, *OS*, *Device*). *ClientIp* (IPv4) is directly used in its 32-bit integer form. For fairness, we also add the session length as a feature, which is similar to the access frequency of BotGraph. We provide both results with and without the *ClientIp* feature. It shows that most methods have no more than 75% precision without *ClientIp* and 87% precision with *ClientIp*. It is also notable that BotGraph uses neither *ClientIp* nor *UserAgent* as features, but achieves ~95% precision and recall.

5 Discussion

A weakness of BotGraph is when the number of spots in trace images is pretty small (e.g., < 3), the detection result can be wrong. It is because the normal users and bots are more likely to have similar page browsing behaviors when only one or two page patterns are accessed. We show several randomly selected false positives and false negatives in Table 7. We can see some images of false positives are nearly the same as the images of false negatives. We believe this drawback is not severe, considering bots are harmful largely due to their large amount of traffic caused to the site. In fact, we found that for more than 95% of the sessions with 1~2 spots, their access frequency is also minimized by accessing each page only once. We think a bot which only makes one or two requests in total will do no harm to the site.

6 Conclusion and Future Work

Website bots have been proven to be a severe threat for Internet these years. In this paper, BotGraph provides a novel scheme to describe bot behaviors in 2-dimensional images. Then state-of-the-art image classification methods like CNN can be used to determine whether a session is a bot. The experiments on real-world 35-day datasets show that BotGraph is a very effective model to detect bots: it achieves ~95% both in precision and recall. BotGraph leverages the client’s behavior instead of its identity as features, it is a promising way to detect advanced bots that frequently change their identities. Currently, we are working on more generic graph-based classification method to take advantage of underlying general graph related feature to better describe the characteristics of bots.

References

- [Akamai, 2018] Akamai. Akamai bot manager. <https://www.akamai.com/us/en/products/security/bot-manager.jsp#features>, 2018.
- [Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, pages 785–794, New York, NY, USA, 2016. ACM.
- [CSIC, 2009] CSIC. Csic 2010 http dataset in csv format. <https://petescully.co.uk/research/csic-2010-http-dataset-in-csv-format-for-weka-analysis/>, 2009.
- [Forrester, 2018] Forrester. The forrester new wave™: Bot management, q3 2018. <https://www.forrester.com/report/The+Forrester+New+Wave+Bot+Management+Q3+2018/-/E-RES143516>, 2018.
- [Google, 2005] Google. Sitemaps protocol 0.84. <https://googlepress.blogspot.com/2006/11/major-search-engines-unite-to-support-16.html>, 2005.
- [Google, 2019] Google. Google analytics. <https://marketingplatform.google.com/about/analytics/>, 2019.
- [Incapsula, 2016] Incapsula. Bot traffic report 2016. <https://www.incapsula.com/blog/bot-traffic-report-2016.html>, 2016.
- [Jaafar *et al.*, 2019] Ghafar A Jaafar, Shahidan M Abdullah, and Saifuladli Ismail. Review of recent detection methods for http ddos attack. *Journal of Computer Networks and Communications*, 2019, 2019.
- [KDD, 1999] KDD. Kdd cup 1999 data. <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>, 1999.
- [Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- [Lagopoulos *et al.*, 2017] Athanasios Lagopoulos, Grigorios Tsoumakas, and Georgios Papadopoulos. Web robot detection in academic publishing. *CoRR*, abs/1711.05098, 2017.
- [Networks, 2018] Distil Networks. Block bot detection. <https://www.distilnetworks.com/block-bot-detection/>, 2018.
- [Schölkopf *et al.*, 1999] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS'99*, pages 582–588, Cambridge, MA, USA, 1999. MIT Press.
- [Shah, 2016] Harsh Shah. A simple e-commerce web-site using flask. <https://github.com/HarshShah1997/Shopping-Cart>, 2016.
- [ShieldSquare, 2019] ShieldSquare. Detect bots in real-time with shieldsquare bot mitigation solution. <https://www.shieldsquare.com/bot-traffic-detection/>, 2019.
- [UNB, 2014] UNB. Botnet dataset. <https://www.unb.ca/cic/datasets/botnet.html>, 2014.
- [Verlet, 1967] Loup Verlet. Computer” experiments” on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical review*, 159(1):98, 1967.
- [Yuan *et al.*, 2017] X. Yuan, C. Li, and X. Li. Deepdefense: Identifying ddos attack via deep learning. In *2017 IEEE International Conference on Smart Computing (SMART-COMP)*, pages 1–8, May 2017.
- [Zenati *et al.*, 2018] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *CoRR*, abs/1802.06222, 2018.
- [Zhai *et al.*, 2016] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16*, pages 1100–1109. JMLR.org, 2016.
- [Zong *et al.*, 2018] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. 2018.