VegaStar: An Illegal Domain Detection System on Large-scale Video Traffic

Xiang Tian[1,2,3], Yujia Zhu[2,3], Zhao Li[1,2,3], Chao Zheng[2,3], Yong Sun[2,3] and Qingyun Liu[2,3]
[1]University of Chinese Academy of Sciences
[2]National Engineering Laboratory for Information Security Technologies
[3]Institute of Information Engineering, Chinese Academy of Sciences
Email: {tianxiang,zhuyujia,lizhao,zhengchao,sunyong,liuqingyun}@iie.ac.cn

***Abstract*—Today online video is growing and becoming increasingly popular on the Web. It is no secret that illegal content is now a one-click-away from everyone, including children and minors. Intelligent video analysis methods can help to automatically detect and isolate questionable content in media. Unfortunately, these methods are hugely costly, and affecting public privacy. In this paper, we present an illegal domain detection system on large-scale video traffic, VegaStar. Using metadata of over 5 million URLs of video, VegaStar: (i)provides lexical and behavior characteristics of video domain names, (ii)proposes a model to detect illegal video domains constructed by twelve feature sets, (iii)detects website domains hosting illegal video content even before the videos are being downloaded, and (iv) understand different CDNs and cloud providers that host content for a particular resource. We conduct extensive experimental analysis and the result shows that the proposed model can classify domains with accuracy approximately 90% by cross validation experiments on Random Tree. We argue that VegaStar represents an important development in the field of video traffic identification, and it can be significantly improve the efficiency of former methods.**

I. Introduction

Today the Internet is a large multimedia delivery infrastructure [1], [2]. "Content is King" is a current meme when organizing or building a website. Video content is particularly popular: By 2021, 82 percent of all consumer IP traffic will be video.

Going beyond being popular among adolescents, video content has now evolved into a much-debated public concern because of excessive or maladaptive use. In fact, a growing collection of the literature demonstrates a consistently positive relationship between adolescents Internet consumption and Internet risks [3]. For decades, parents and others have been consistently concerned about the potentially harmful influences of exposure to pornographic and violent content, being targeted for harassment, cyber-bullying, sexual solicitation, and Internet addiction [4].

Censorship of video images has become an important field of science nowadays. However, the concern about privacy on the content-based censorship is increasingly becoming an issue of dispute. Besides,current video content detection algorithms are not suitable for large-scale traffic applications. The content inspection is inefficient. Steps such as assembling, decoding, extracting the key frames of the video, analyzing the context structure or motion information of the video, often require a huge cost of time and human resources [5].

In this paper, we focus on the website domain names hosting video content rather than the content itself. We describe a class of domain names hosting pornographic and violent content as *illegal domains*. Therefore, this article is different from malicious domain name detection which often referred to phishing and pharming attacks [6].

The goal of this work is to show that there is a significant portion of illegal video traffic passing through Internet gatewaysnamely, flows with domain names that can be immediately classified, simply by looking at the HTTP header and at the domain name. We also show interesting dependencies between IP and domains on CDN-like HTTP flows, demonstrating an innovative perspective for flow-based video traffic analysis.

In this paper, we present VegaStar: a novel method for classifying IP flows by inspecting HTTP and DNS traffic transmitted in a computer network. First, CDN-like HTTP and DNS flows from the ISP are passively collected. VegaStar identifies video traffic through file suffixes in URL *(e.g. dl.stream.qqmusic.qq.com/C4000041Xpkq3XMxIp.m4a)*. Second, a filter modular is applied to VegaStar. This filter divide video domain names into three parts. That is (1)high reputation-benign domains, (2)well known illegal domain name, and (3)suspicious domain names. Thirdly, by running classification on suspicious domain name sets, VegaStar reveals illegal video domain names by proposing a new model containing twelve feature sets.

The novelty of our method is explained in the fact we believe this is the first classifier employing HTTP header and DNS features of domain names on large-scale video traffic mainly from a passive way. Specifically,our main contributions are summarized as follows.

- An illegal domain detection system on large-scale video traffic, VegaStar, is presented. Based on HTTP traffic, discover candidate illegitimate domain names and then correlate DNS traffic to verify illegitimacy. The method is mainly passive, supplemented by active Verification.
- We provide lexical and behavior characteristics of video domain names. Totally twelve video domain features for online traffic are constructed. We argue that IP diversity,

TABLE I
Nov.2017 and Feb.2018 Datasets

Datasets	Start	Duration	Src.IP	Dst.IP	Log Bytes	Benign Log Bytes	Illegal Log Bytes
Nov.2017	2017-11-17 15:30	20h	596636	23299	4.49GB	1.71GB	2.78GB
Feb.2018	2018-02-28 9:20	7h	395159	19522	2.23GB	1.76GB	0.47GB

TABLE II
URLs Distribution in Datasets

	Benign URLs	Illegal URLs	Benign Files	Illegal Files	Benign File Types	Illegal File Types
Nov.2017	988933	2542880	314024	164182	59	43
Feb.2018	1079581	428590	2997886	58539	65	27

IP count, IP location diversity, Subdomain count, TTL values have a greater impact on the classification results.

- This article compared several popular classification algorithms. The measurements show, VegaStar can classify domains with accuracy approximately 90% by cross validation experiments on Random Forest. Thus it can be significantly improve the efficiency of former methods.

The rest of the paper is organized as follows. In Section II, we provide background and data collection process. In Section III, we define illegal domains and examine their key properties. In Section IV, we provide details of our illegal domain detection system. In Section V, we discuss experiments. We comment on related works in Section VI. Finally, we conclude the paper in Section VII.

II. Background and Datasets

CDN-like HTTP traffic: On HTTP traffic, if the same file relates to multiple IPs, it is regarded as CDN-like HTTP traffic. CDN-like means that the same file is stored in several different network locations. Nowadays, video content providers usually use CDN to distribute content across POPs(Points of Presence) around the world in order to achieve load balancing [7]. CDN-like content distribution strategy, on the one hand ensuring that clients get faster access to content and availability, on the other hand, is a way of avoiding censorship. We focus on the domain names extracted from visual CDN-like HTTP traffic, which greatly reduces the amount of analysis requirement and limits the incoming traffic to visual related traffic.

Illegal Domains and Benign domains: The illegal domains in our work focuses on are domain names that hosting illegal content. Initially, we collected illegal domains from multiple sources. Specifically, based on the public illegitimate domain names list and video content censorship, matching the domain name and labeling illegal. Our whitelist was obtained through Alexa and Baidu list of video websites, removing from some of the domains appeared in the list but actually illegal by research or statement. Labeling these domain names as benign.

Datasets: This paper investigates video streaming traffic from an ISP perspective. The input of the system is based on HTTP and DNS log, sniffed passively. Compared with active probing domain names [8] , active probing method may be detected by the attackers, who often controls the authoritative name servers responsible for responding to DNS queries about domain names or modify HTTP traffic content. In another word, our passive detection system is able to detect video services in a stealthy way. The URL in the HTTP log is the location where the video content is stored, not only the accessing websites. We obtain the URL about the resource really storage. The starting point of our entire system and the basis for determining video traffic.

For this study, HTTP logs firstly recorded with URLs containing file name and type of video content. We judge whether the file is a video traffic by file suffix. We sniffed 24-hours long traffic in November 17, 2017 and February 28, 2018 within a large ISP. To protect privacy, client IP addresses and other sensitive information were anonymized. Table I summarizes the datasets. The two datasets contains 5039984 URLs, 909304 individual URLs, contains 809304 files and 89 file types, the composition in Benign and Illegal domain names are shown in Table II.

In two HTTP datasets, we count the proportion of different file type, and in Table III displays six of the most video file types, including MP4/TS/M3U8/MP3/RMVB/F4V. In addition, it is worth noting that the largest percentage of file types between Illegal and Benign domain names is completely disparate. The files is dominated by TS type in Benign domains, up to 58% and 57% in 2017 and 2018 datasets.

III. Feature Selection

In this paper we refer to the first sub-domain after the TLD as second level domain(2LD); it generally refers to the organization that owns the domain name. Finally Fully Qualified Domain Name (FQDN) is the domain name complete with all the labels that unambiguously identify a resource [9].

We conduct an in-depth analysis of the illegal features. These features reflect the behavior patterns of a given domain name, shown in Table IV. Fig. 2 and Fig. 3 shows the CDF for all twelve features in the ALL dataset. Each plot has two curves, the dashed line shows the CDF of the feature values for the benign domain names and the solid line shows the CDF for the illegal domains.

1) ***Independent Characteristics:*** Independence refers to the independence of the domain name and IP address corresponding to the domain name, DNS behavior characteristics of a given domain name will be considered separately to extract

TABLE III
NOV.2017 AND FEB.2018 DATASETS FILE TYPE

	Nov.2017 Illegal	Nov.2017 Benign	Feb.2018 Illegal	Feb.2018 Benign
MP4	35%(149735/426986)	27%(115002/426986)	11%(48027/426986)	27%(114222/426986)
TS	0.1%(379/353075)	52%(182442/353075)	0.001%(5/353075)	48%(170249/353075)
M3U8	0.1%(22/15914)	52%(8241/15914)	48%(7641/15914)	0.06%(10/15914)
MP3	0.7%(65/8914)	56%(5002/8914)	0.28%(25/8914)	43%(3822/8914)
RMVB	44%(2657/6057)	31%(1866/6057)	33%(1971/6057)	22%(1303/6057)
F4V	8%(501/5974)	44%(2629/5974)	9%(552/5974)	38%(2292/5974)

the corresponding characteristics of benign and illegal domain name.

We exploit the fact that illegal domains are required to be stealthy and available at the same time. The web-sites providing illegal video content need to be well-reachable by the users at anytime, and responding the requested content efficiently to gain profit. High availability of illegal domains implies some features sharing with fast-flux domains by the following:

ϕ_1. **IP count.** Number of distinct IP addresses per domain name. Illegal video content hosted by several IPs. The IP corresponding to the domain name is composed of the resolved IP returned from the A record in the DNS log and the IP generated by accessing the domain name in the HTTP log.

ϕ_2. **Subdomain count.** Number of subdomains per domain name. Illegal video websites comprised by several subdomains.

ϕ_3. **Subdomain length entropy.** The entropy of distinct subdomains per domain name.

ϕ_4. **Max DGA ratio.** The maximum value of the DGA ratio of the domain name pointed to by IPs per domain name.

ϕ_5. **IP location diversity.** The number of distinct countries per domain name.

ϕ_6. **IP diversity.** The number of distinct /16 network prefixes per domain name. Illegal domain names are scattered several different networks.

$$p16\_entropy = \frac{-\sum_x p(x) log_2 p(x)}{log_2 |P|} \tag{1}$$

ϕ_7**. TTL values per domain name.** We record three values of TTL, including min TTL, max TTL and domain activity period. domain activity period computed as follows:

$$lifetime = max_{ttl} - min_{ttl} \tag{2}$$

*2) **Relevant Characteristics**:* Domain names that provide illegal video services are strongly related to each other to a large extent. If a given domain name is similar to a domain name already present in the illegal domains blacklist, the domain is likely to be illegal. Statistical analysis of existing blacklist which contains 2210 domain names. 1388 of them have the identical 2LD domain name. More than half of them have the semblable 2LD domain names.

Although the domain name of each website is different, the subject of the domain name under the same topic may be the same. Usually, some domain-specific words such as ”sport” in the sports category, ”finance” in the economic category, ”auto” in the automobile category, etc. will be concentratedly used. In other words, the domain name subject word in the Internet has been given the wisdom of people. This article define the lexical words to classify domain names as domain-key.

On the other hand, correlation includes the same domain name is resolved between multiple IP is related to resolve to the same IP between multiple domain names are related. Often, a group of IP addresses for web-sites that host benign video content, such as legitimate CDNs, will have more than one domain name pointing to them, and the associated domain names will be more. The illegal domain name, in order to evade censorship will often change the IPs and domain names, so the associated domain name will be less.

ϕ_8. **Score of similarity with the illegal domain names.**

ϕ_9. **Number of relevant domain names.**

ϕ_{10}. **Ratio of distinct 2LD in relevant domain names.**

ϕ_{11}. **Ratio of non-shared IP in relevant domain names.**

ϕ_{12}. **Ratio of non-shared /16 network prefixes in relevant domain names.**

IV. SYSTEM DESIGN

In this section we describe how VegaStar works. Fig. 1 shows a high-level overview of this system. The VegaStar is composed of three procedures:(1)Traffic Preprocessing, (2)Filter,and (3)Classification. Firstly, by special strategies, we obtained the video CDN-like HTTP traffic. The second procedure is matching the domain names in the *Domain Bucket* with benign and illegal domains. Thirdly, To achieve further load shedding, VegaStar considers the suspicious for deeply analysis.

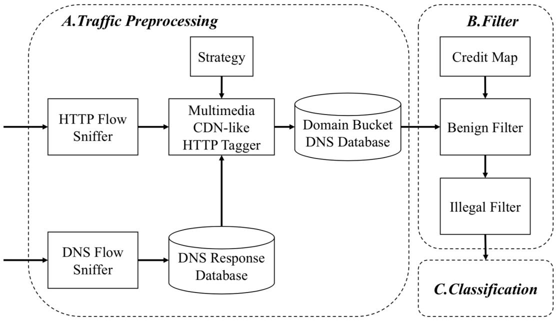
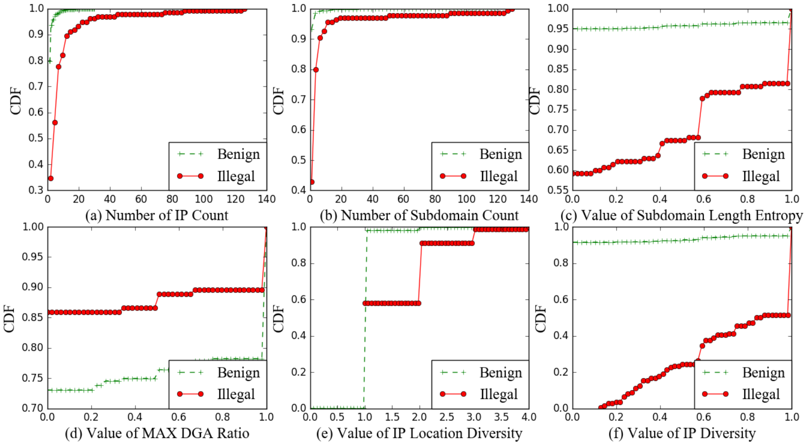


Fig. 1. VegaStar Architecture Overview

TABLE IV
FEATURES

category	#	feature
Independent	1	IP count
	2	Subdomain count
	3	Subdomain length entropy
	4	Max DGA ratio
	5	IP location diversity
	6	IP diversity
	7	TTL values per domain name
	8	score of similarity with the illegal domain names
Relevant	9	number of relevant domain names
	10	ratio of distinct 2LD in relevant domain names
	11	ratio of non-shared IP in relevant domain names
	12	ratio of non-shared /16 prefixes in relevant domain names



(a) Number of IP Count (b) Number of Subdomain Count (c) Value of Subdomain Length Entropy

(d) Value of MAX DGA Ratio (e) Value of IP Location Diversity (f) Value of IP Diversity

Fig. 2. CDF Distributions for the 12 Features(1)

A. *Traffic Preprocessing*

From a passive perspective, the module of Multimedia CDN-like HTTP Tagger aims to separate multimedia HTTP traffic. The first step is to identify the HTTP URL with video suffix. The second step is to get the CDN-like video resources with different cache IPs . Then, VegaStar extracts the resources domains as well as the well-known five tuples into *Domain Bucket*. Last but not the least, VegaStar selects the related DNS logs in DNS Response Database into *Domain Bucket*.

*1) **Multimedia CDN-like HTTP Tagger**:* The main strategy of identifying video files is the file suffix. In general, the audio file suffix includes MP3/WMA/WAV and so on, and the video file suffix contains WMV/AVI/FLV and so on.

All IPs corresponding to domain name of *Domain Bucket* in the HTTP log are recorded and stored in the *Domain Bucket* DNS Database as attributes of the domain name.

*2) **Domain Bucket DNS Database**:* We deploy Passive DNS in one of the monitored links to get a deeper insight into the features of different domains. DNS Flow Sniffer and DNS Response Database obtained all DNS record. And then VegaStar filters the corresponding DNS traffic of *Domain Bucket*. Since we monitored the passive traffic at the ISP vantage point, we could not find all of the DNS records corresponding to the *Domain Bucket* in the DNS log. For this part domain names, we actively dig to obtain the DNS traffic and added it to the database as a supplement.

B. *Filter*

The Filter module contains Benign Filter and Illegal Filter. Benign Filter filters DNS traffic which related to benign domain names. Illegal Filter filters illegal domain names. Based on two filtering steps, the traffic volume reduces significantly. To the purpose of reducing our computational burden, and making sure that the Filter does not discard illegal domains, we conservatively set the filtering rules. Only traffic related to domains with explicit labels will be discarded.

***Credit Map**:* The filter rules of Benign Filter are produced by Credit Map modules, and these rules are base on keyword checking, pumped-frame form video or manual review. Obviously, a domain name with a complete record of information and a well-defined organization usually provides valid service through formal registration. Credit Map module could access a website credit and justify whether a website is benign, according to: (a)legitimate domain name updates WHOIS information more often, while most illegal domain names almost never change WHOIS data, (b)legitimacy domain name

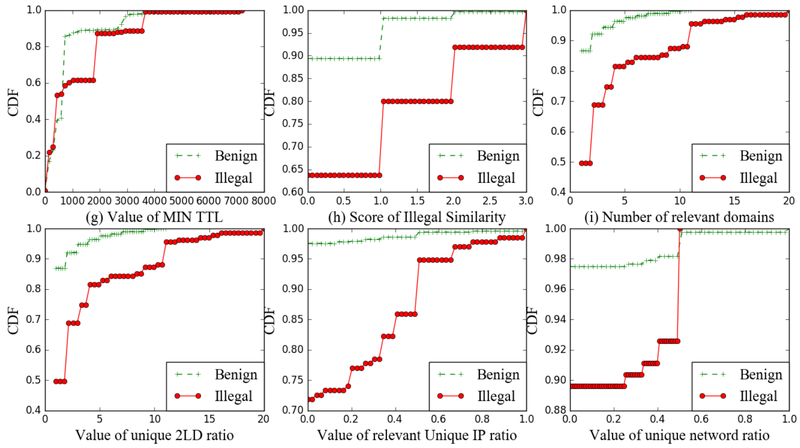


Fig. 3. CDF Distributions for the 12 Features(2)

always has a clear and specific organization information, while illegal domains don't have., and (c) If a domain has not an ICP(Internet Content Provider) Licensing or uses ICP Licensing privacy protection (WHOISGUARD, Domains By Proxy), this kind of domains are scored lower score.

C. Classification

Suspicious Domains will be classified in this procedure.

1) ***Feature Extraction****:* Feature extraction extracts the characteristics of the domain names which are described in Section III.C, and these are more suitable for domain names hosting illegal contents of video in the DNS traffic.

2) ***Classifier****:* The goal of classifier module is able to label domain as being benign, or candidate illegal. Thus, we require a training set that contains DNS traffic of a labeled and representative sample of benign and illegal domains. We used these domain names tagged in the *Domain Bucket* for constructing our training set.

We measure the 12 features described in Section II, and employ the popular random forest classifier to automatically classify a given domain name as either non-illegal or candidate illegal. First, By the training set, we learn the classifier module. Afterwards, the trained classifier is used to classify the unknown domain names in *Domain Bucket*.

3) ***Active Validation****:* The entire system aims to reduce the amount of video censorship of domain names stored video content. Since the goal of video checking and manual review is to discover illegal domain names and to monitor or regulate this part of the domain names, we are more concerned about the illegal domain name. The output of classifier module is non-illegal and candidate illegal domain names. For these candidate illegal domain names, we employ further proactive verification, filtering out the truly illegal domain name.

Illegal domain names often can not be accessed directly by posting HTTP requests, and the response status code of HTTP request is over 400, so we take the initiative to verify from another point of view.

Our active verification strategy refers to utilize the search engine to search the domain name, we will get the information directly or indirectly related to the domain names. In the returned results, matching the key words of malicious content, if the word hit the malicious keyword list, then determine the domain name is illegal.

V. Experiment

In this section, we discuss the experimental results of our VegaStar system. The system input is the domain name of the website storing the video content obtained from the HTTP log, including the illegal, legitimate and suspicious domain names. In our experiment, we tested online adult website resources as illegal domain samples, which accounts for most of the illegal video content in Internet. According to one estimate, there are at least 4 million adult websites on the Internet, which constitute approximately 12% of all websites [7]. Overall, these statistics indicate that online adult content attracts a large number of users and accounts for a substantial fraction of the global Internet traffic. The identification of online adult traffic is important for discerning illegal multimedia traffic.

Specifically, we evaluate the availability and correctness firstly. Second, we compare classification algorithms to certificate the advantages of Random Forest which we chooses in our system in this scenario. Finally, in order to verify the validity of the selected features, we analyze the classification effects of different feature subsets.

A. Input Data

We use the labeled domain names to evaluate our system. Two datasets Nov.2017 and Feb.2018 contains 7126 labeled FQDNs, 2017 labeled 2LD domain names in total. Additionally, we combine the two labeled datasets to acquire a more larger datasets, defined as ALL. ALL datasets include 4916 benign FQDNs, 2210 illegal FQDNs, 1195 benign 2LD and 822 illegal 2LD. Table V summarizes the domains names in three labeled datasets.

TABLE V
DOMAINS DISTRIBUTION IN THREE LABELED DATASET

	ALL domains	Benign domain	Illegal domain	Benign 2LD	ALL 2LD	Illegal 2LD
Nov.2017	5767	3708	2059	1666	904	762
Feb.2018	3969	2437	1532	1086	594	492
ALL	7126	4916	2210	2017	1195	822

Based on the TF-IDF, we statistics the word frequency in different domain names and found that there exist semantic features: string features of illegal domain name in the *Domain Bucket* and certain semantic patterns. The string feature refers to substrings such as ”*porn*”, ”*gay*” and so on. The string feature is domain-key which used to discern online adult websites. We have a dictionary containing 425 domain-key of adult websites. Once these substrings appear, the possibility of illegal is great. The specific mode refers to the 2LD consists of characters and numbers, the 2LD has a large number of repeated characters appear one after another, for example, the domain name is ”*11bubu.com*”.

B. Experiments with the Labeled Datasets

Methodology: To evaluate the accuracy of our system, we classified our training sets with 10-fold cross-validation and percentage split, where 66% of the training set is used for training, and the rest is used to check the accurateness.

Result: Fig. 4 shows the ROC(Receiver Operating Characteristic) curve results of the percentage split about three labeled datasets, confirms that our system can detect illegal domains with high accuracy. The mean area under the ROC curve(AUC) of Random Forest was able to achieve 0.902 in ALL datasets. Besides the AUC, we calculate the confusion matrix, including True Positives(TP), True Negatives(TN), False Positives(FP) and False Negatives(FN), then,we apply accuracy, precision, recall and f1 value to evaluate out system, the result is summarized in Table VI.

TABLE VI
CONFUSION MATRIX RESULT OF THREE LABELED DATASETS

	ALL Dataset	Nov.2017 Dataset	Feb.2018 Dataset
Accuracy	0.846	0.836	0.845
Precision	0.858	0.793	0.785
Recall	0.744	0.657	0.667
F1	0.797	0.719	0.719

Additionally, we exploit the 10-fold cross validation about three labeled datasets. The precision of classification is 88% with ALL Dataset, while it can achieve approximate 91% with Nov.2017 Dataset. Therefore, employ appropriate data preprocessing method for different datasets would improve the classification correctness.

C. Classification Algorithms Evaluation

In our system, we employ Random Forest. It is based on the decision tree, which is utilize to select all the variables in the classification. However Random Forest is to produce a lot of decision trees, and then each decision tree to select different variables for analysis, and finally select the decision tree mode as the final result. The reason of choosing this classifier is more sufficient in video services scenes compared to other classifiers.

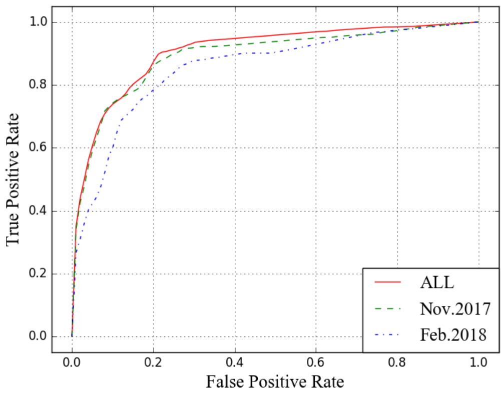


Fig. 4. Receiver Operating Curve(ROC) for the Random Forest, evaluated on the three labeled dataset.

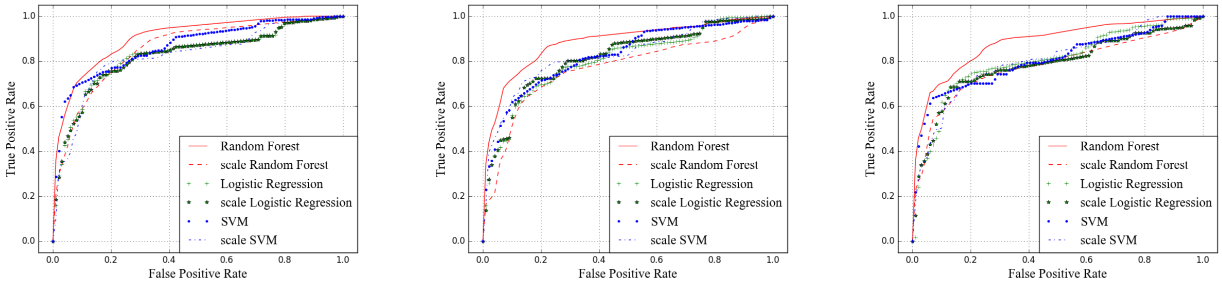
Classification Algorithms: The most commonly used classification methods is SVM [10]. Initially, we choose SVM, Logistic Regression and Random Forest. We also consider the regularization of features, based on the attributes of the eigenvalues, select the Z-Score for feature normalization, subtract the mean of the data features by attributes (by columns), and impose their variance.

Result: We performed 10-fold cross validation to evaluate the detection rate of SVM, Logistic Regression and Random Forest. Precision is used to assess the quality of classification algorithms. The average precision with Logistic Regression is 78%, Random Forest can achieve up to 89%, with an average of 84%.

Additionally, We performed 66% split validation to evaluate the detection rate in Fig. 5. The mean AUC with Logistic Regression, SVM, Random Forest is 78%, 84%, 90%. Instead, we choose a simple Random Forest, which can get effective results under the category of domain names, the method of implementation is unambiguous. After training, it can give out the more important features. And training between trees is independent of each other, the training speed is fast, so it is easy to adopt in parallel methods.

D. Feature Extraction Evaluation

In order to verify the validity of the classification features we selected, we evaluate the extracted features. The feature importance result shows that IP diversity, IP count, IP location diversity, Sub-domain count, TTL values, score of similarity



(a) ALL Dataset

(b) Nov.2017 Dataset

(c) Feb.2018 Dataset

Fig. 5. Receiver Operating Curve(ROC) for algorithms, evaluated on the three labeled dataset.

with the illegal domain names have a greater impact on the classification results.

By random forest before 6 important features as basic features, we compare the precision of two features sets to assess the classification effect, the combination of all features and basic features. The eigenvector group all features can achieve 89% precision rate. However, the accuracy rate of 80% can be achieved by using only 6 features. When the data volume is too large, in order to improve the classification efficiency of the system and balance the efficiency and accuracy, basic feature sets can be used for classification.

VI. Related Work

At present, we have not seen adequate research conducted on the use of passive way to help reducing the load of video content-based censorship via HTTP header and DNS logs on large-scale video traffic. As far as we know, mainly research work is about measurement and analysis adult sites like YouPorn and PornHub. The most relevant to our discern research work is the detection of malicious domain names. Prior word on malicious domains detection fall into two categories: active identification and passive identification.

A. Measurement Illegal Domain

TYSON et al. has presented a detailed measurement study of a large-scale study of one of the most popular Porn 2.0 websites: YouPorn [11]. Exploring its delivery infrastructure using Domain Name System (DNS) and HTTP probes, confirming a significant distribution scale, with servers spread across the globe. They inspected five key aspects of this system: the delivery infrastructure, the video upload characteristics, the nature and evolution of content popularity, the use of 2.0 features, and the impact of using categories. Ahmed et al. presented the first large-scale measurement study of online adult traffic using HTTP logs collected from a major commercial content delivery network [7]. They analyzed approximately 323 terabytes worth of traffic, revealed several unique characteristics of online adult traffic on adult website design and content delivery infrastructure management.

B. Detect Malicious Domain

Illegal is different from malicious, but there still are some common characteristics, so the first study of related research of malicious domain is the basis of our research. The second line of research focuses on the behaviors differentiation in HTTP traffic about malicious or benign domains.

*1) **Active Identification**:* These studies are based on probing the specific domains which occurred in publicly available blacklists or malicious domain sites actively. A number of approaches for detecting malicious domain names differ from each other in the number of features used to characterized domains, and the details of the classification algorithms. The main limitation of these works lies in the use of spam email as the primary information source, thus detecting malicious domains advertised through email spam [12], [13], [14], [8]. Active probing of malicious domain names may be detected by the attacker. Our detection system is able to detect flux services in a stealthy way. And the input traffic is the real traffic generated by passive monitoring and obtained by the user actively accessing the video resource service.

*2) **Passive Identification**:* In general, flux domain detection though passive monitor follows two lines of research: The first line of research tries to detect domains by monitoring DNS traffic, this type of research has proposed number of approaches that leverage the distinguishing features between malicious and benign DNS usage.

***Based on DNS traffic**:* Chiba et al. presented DomainProfiler [15], a system actively collects DNS logs, analyzes their temporal variation patterns, and predicts whether a given domain name will be used for malicious purposes. DomainProfiler can predict malicious domain names 220 days beforehand with a true positive rate of 0.985. Antonakakis et al. proposed a dynamic reputation system for domain names, called ”Notos” [16]. The system processes DNS query responses from a passive DNS database and extracts a set of 41 features from observed FQDNs and IPs. Notos uses historic IP addresses and historic domain names to extract effective features to discriminate malicious domain names from legitimate ones. In a similar spirit, Bilge et al. presented their ”EXPOSURE” system [17], which requires 15 features and 1 week training data. Perdisci et al. proposed a system, FluxBuster, which detects previously unknown fast-flux domain names by using large-scale passive DNS data [18]. It is capable of accurately detecting previously unknown flux networks days or even weeks in advanced before they appear in public blacklists.

Both systems employ the Alexa list for whitelisting popular domains. In addition, as our identify benign domains methodology relies on WHOIS record, Alexa list and Baidu list.

Based on HTTP traffic: Recently Hsu et al. proposed a real-time system for detecting flux domains based on anomalous delays in HTTP/HTTPS requests from a given client [19]. The assumption is that the malicious domains often provide the malicious web content with large latencies. Manadhata et al. [20] proposed a malicious domain name detection system. The system models the detection problem to graph inference problem by constructing the detection log as the domain name map of the host.

Some research apply machine learning techiques to analyze and classify URLs or web content based on static features and behaviors such as URL lexical patterns, HTML page content, Javascript features, or host attributes [21]. Mekky et al. developed a machine learning-based method for identifying malicious HTTP directions [22]. Our methodology only relies on the URLs generated by client access behavior, we do not need so-phisticated de-obfuscation mechanisms to untangle obfuscated websites. Furthermore, we obtain the URL about the resource really storage.

VII. Conclusion

In this work, we developed VegaStar, a system to unveil illegal domains on large-scale video traffic by constructing 12 features. We verified that it is possible to detect illegal video domains with accuracy approximately 90%. We collaborated lexical and behavior analysis based on HTTP and DNS traffic, employed file type association obtaining video domain names in HTTP logs. Based on the thesis of utilizing features in HTTP head and DNS traffic, we can detect website domains hosting illegal video content even before the content being downloaded.

Acknowledgment

The authors would like to thank the anonymous reviewers for their valuable comments, and hard work of MESATEAM (*www.mesalab.cn*). This work was supported by National Key R&D Program 2016 (Grant No. 2016YFB0801300)

References

[1] G. Maier, A. Feldmann, V. Paxson, and M. Allman, "On dominant characteristics of residential broadband internet traffic," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*. ACM, 2009, pp. 90–102.

[2] A. Rao, A. Legout, Y.-s. Lim, D. Towsley, C. Barakat, and W. Dabbous, "Network characteristics of video streaming traffic," in *Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*. ACM, 2011, p. 25.

[3] L. Leung and P. S. N. Lee, "The influences of information literacy, internet addiction and parenting styles on internet risks," *New Media & Society*, vol. 14, no. 1, pp. 117–136, 2012.

[4] L. Leung, "Predicting internet risks: a longitudinal panel study of gratifications-sought, internet addiction symptoms, and social media use among children and adolescents," *Health Psychol Behav Med*, vol. 2, no. 1, pp. 424–439, 2014.

[5] Y. Chen, W. He, Y. Hua, and W. Wang, "Compoundeyes: Near-duplicate detection in large scale online video systems in the cloud," in *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*. IEEE, 2016, pp. 1–9.

[6] M. A. Rajab, L. Ballard, P. Mavrommatis, N. Provos, and X. Zhao, "The nocebo effect on the web: An analysis of fake anti-virus distribution." in *LEET*, 2010.

[7] F. Ahmed, M. Z. Shafiq, and A. X. Liu, "The internet is for porn: Measurement and analysis of online adult traffic," in *Distributed Computing Systems (ICDCS), 2016 IEEE 36th International Conference on*. IEEE, 2016, pp. 88–97.

[8] M. Konte, N. Feamster, and J. Jung, "Dynamics of online scam hosting infrastructure," in *International conference on passive and active network measurement*. Springer, 2009, pp. 219–228.

[9] I. N. Bermudez, M. Mellia, M. M. Munafo, R. Keralapura, and A. Nucci, "Dns to the rescue: discerning content and services in a tangled web," in *Proceedings of the 2012 Internet Measurement Conference*. ACM, 2012, pp. 413–426.

[10] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[11] G. Tyson, Y. Elkhatib, N. Sastry, and S. Uhlig, "Measurements and analysis of a major adult video portal," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 2, p. 35, 2016.

[12] T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling, "Measuring and detecting fast-flux service networks." in *NDSS*, 2008.

[13] E. Passerini, R. Paleari, L. Martignoni, and D. Bruschi, "Fluxor: Detecting and monitoring fast-flux service networks," in *International conference on detection of intrusions and malware, and vulnerability assessment*. Springer, 2008, pp. 186–206.

[14] J. Nazario and T. Holz, "As the net churns: Fast-flux botnet observations," in *Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on*. IEEE, 2008, pp. 24–31.

[15] D. Chiba, T. Yagi, M. Akiyama, T. Shibahara, T. Yada, T. Mori, and S. Goto, "Domainprofiler: Discovering domain names abused in future," in *Dependable Systems and Networks (DSN), 2016 46th Annual IEEE/IFIP International Conference on*. IEEE, 2016, pp. 491–502.

[16] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, and N. Feamster, "Building a dynamic reputation system for dns." in *USENIX security symposium*, 2010, pp. 273–290.

[17] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi, "Exposure: Finding malicious domains using passive dns analysis." in *Ndss*, 2011.

[18] R. Perdisci, I. Corona, and G. Giacinto, "Early detection of malicious flux networks via large-scale passive dns traffic analysis," *IEEE Transactions on Dependable and Secure Computing*, vol. 9, no. 5, pp. 714–726, 2012.

[19] C.-H. Hsu, C.-Y. Huang, and K.-T. Chen, "Fast-flux bot detection in real time," in *International Workshop on Recent Advances in Intrusion Detection*. Springer, 2010, pp. 464–483.

[20] P. K. Manadhata, S. Yadav, P. Rao, and W. Horne, "Detecting malicious domains via graph inference," in *European Symposium on Research in Computer Security*. Springer, 2014, pp. 1–18.

[21] D. Canali, M. Cova, G. Vigna, and C. Kruegel, "Prophiler: a fast filter for the large-scale detection of malicious web pages," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 197–206.

[22] H. Mekky, R. Torres, Z.-L. Zhang, S. Saha, and A. Nucci, "Detecting malicious http redirections using trees of user browsing activity," in *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2014, pp. 1159–1167.