An entropy-based methodology for detecting Online Advertising Fraud at scale

Anonymous Author(s)

ABSTRACT

Programmatic online advertising allows advertisers to diversify their campaigns dynamically, set the desired context of publishers' content, and target a specific audience. Besides, it is accessible to all budgets. Despite these notable benefits, programmatic advertising has the drawback of being highly exposed to low quality ad traffic, mainly associated to fraudulent activities.

Detecting low quality ad traffic at scale is a difficult task due to the large volume of ad requests delivered by the Ad Tech ecosystem every day as well as the lack of collaboration among different Ad Tech vendors. Indeed, identifying sources of traffic (i.e., IP addresses or websites) involved in ad fraud (and thus generating low quality traffic) requires an important effort due to the reported scale of the problem.

In this paper, we propose an efficient and scalable methodology, based in the concept of entropy, to identify sources presenting anomalously deterministic traffic patterns, so that the traffic from these sources can be safely classified as low quality traffic. We have applied this methodology to a set of large-scale datasets including up to 2.5 B ad requests associated to 1.5 M referrers and 150 M IP addresses. The obtained results indicate that low quality (anomalously deterministic) traffic represents roughly 15 % of the total ad requests in the considered datasets. This may lead programmatic advertising to generate estimated annual losses over $3 B for advertisers only in US. This figure is comparable to the losses generated by any major cyber-crime worldwide.

Finally, note that we have implemented the proposed methodology and released it as open source code for its wider use by Ad Tech vendors and the research community in the context of programmatic advertising.

1. INTRODUCTION

Advertising drives consumer spending, and as a result the global economy. In US for example, 2/3 of the total GDP is a result of consumer spending [4]. Depending on the category of product or service, typically advertising attributes 1 % to 90 % of gross sales for the product [17]. According to GroupM, the world's largest media agency group, US$160 billion are expected to be invested on digital media in 2016 [14], with World Federation of Advertisers reporting total annual media investment globally at close to $700 billion [1]. In addition, next year investment on digital will be overtake investment on TV in US [25].

Online advertising claims better access even for smaller advertisers, and more capabilities together with an alleged lift in comparison to traditional media. At the same time, it has been shown that digital media investment is burdened by lack of transparency, various forms of wastage and fraud [6, 19]. Low quality traffic, significantly represented by ad fraud, is the focus of the research and solution outlined in this paper. In a recent report titled Business of Hacking, HP Enterprise ranked ad fraud the lowest complexity and highest yielding cyber-crime [12]. A literature review shows that no well established measurement methodology is available to mitigate ad fraud, with estimates of total exposure ranging from 2 % to over 90 % of all ad buys in question [23, 33]. Even by moderate estimates, ad fraud has grown into a cyber mega-crime. Because of the central role the ad delivery chain has with the Internet's energy footprint, webpage associated end-user payload often being over 80 % related to ads, fraud is creating a negative effect in the ad delivery chain from data center energy economics to losses in national economy. In summary, ad fraud is a critical problem threatening the Internet age economy.

Detection and prevention of ad fraud and other types of low quality traffic is currently dependent on the propriety capabilities provided by various vendors (e.g. WhiteOps, Trustmetrics, Sentant, Integral Ad Science). This is in stark contrast with well known success stories of overcoming internet security related challenges, where industry-wide adoption of open-source technology has been the norm. For example Snort [29], an open-source Intrusion Detection System used widely by cybersecurity companies of all sizes, has become synonymous with network security.

In its recent guidance to its members, the World Federation of Advertisers (WFA) Ad Fraud Workgroup called for industry-wide adoption of open source solutions for countering ad fraud [1]. The group ranked open-source adoption as one of the critical success factors on a list of 20 strategic objectives for the online ad industry jointly fighting ad fraud.

This paper outlines how we created such a solution, for detecting both general and sophisticated low quality traffic using as a reference the definition of invalid traffic provided by MRC in Invalid Traffic Detection and Filtration Guidelines Addendum [9] in display, mobile, video and app media buys.

In particular, we define a data-driven methodology that measures the level of randomness (or determinism) in the ad requests pattern of a given source, including referrers[1] and IP addresses. To this end, we use the Shannon entropy, which summarizes the mentioned level of randomness in an

[1] Note that in this paper we use the term referrer for websites and mobile apps indistinguishably.

single value. Then, we apply a normalization process to obtain our final metric that we refer to as the Normalized Entropy Score (NES). The NES ranges between 0 and 100, where 0 indicates full determinism in the traffic pattern of a source and 100 indicates full randomness (for instance, in the case of the ad request received by a website, they would be distributed homogeneously among the IPs connecting to the website and this number of IPs would be sufficiently large in comparison with the number of ad requests). Finally, we use outlier identification techniques to find sources presenting a (statistically) anomalous NES value. These sources have an anomalous traffic pattern with an excessive level of determinism, which is typically a strong indicator of low quality of traffic.

We have applied our data-driven methodology using ad requests log files provided by a major AdTech company for 9 different days in 2015. These log files include information of up to 2.14 B ad request associated to 150 M IP addresses and 1.5 M referrers. Using a computing infrastructure consisting of a single Linux server with 48 GB memory, we can run our methodology at the described daily scale to identify referrers and IP address with low quality traffic in matter of a few hours. The application of our methodology to these datasets indicate that 15 % of the traffic in programmatic media buying is indeed low quality traffic. A ball-park estimation translates this fraction of low quality programmatic ad traffic into potential annual losses over $3.3 B for advertisers in US. This figure is comparable to the revenue losses generated by major cyber-crimes worldwide [28].

Note that following the example of the Internet security industry and the recommendation of the WFA, we deliver the code of the developed methodology as open-source[2] for its wider use by the Adtech industry and possibly other fields of research focused on analyzing large-scale Internet data.

In summary, the main contributions of the paper are:

- A novel and scalable methodology to identify websites, mobile apps and IP addresses generating low quality ad inventory.
- A software version implementation of the methodology available as open-source code through Github.
- The application of the methodology to large-scale datasets, which provides worrisome insights on the exposure of advertisers to a significant fraction of low quality traffic in the programmatic media ecosystem.

The rest of the paper is organized as follows: Section 2 shows the dataset as well as the data processing infrastructure used in the paper. Section 3 details the methodology for identifying sources delivering low quality traffic. Section 4 describes the approach used for validating the performance of the methodology. Section 5 shows the results obtained when we apply the detection methodology to our large-scale datasets. Finally, Section 6 presents the most relevant related work and Section 7 concludes the paper.

2. DATASET DESCRIPTION

The dataset used in this paper is formed by a random sampling out of the sell-side ad requests from tens of sources processed by a major vendor from the Adtech industry on 9 different days of 2015. Each entry in the log-file represents one ad request coming from a referrer and initiated by an IP address. Next we provide details about the fields of information for ad request records present in our dataset.

Timestamp: It indicates the time instant when the ad request was issued.

IP address: This field represents the IP address of the device that originated the ad request. Nearly all the entries correspond to IPv4 addresses. Only 30 entries have IPv6 addresses within all the datasets. Moreover, 8 % of the entries do not have an associated IP and then they are mapped to a ”null IP”. Note that we do not filter out these entries since they may be useful for identifying anomalous behaviors. For instance, a referrer with a major fraction of associated ad requests with null IPs.

Referrer: It represents the domain of the ad request’s referrer. Usually, it correspond to the website url (for web pages) or the app ID[3] (for apps). However, the referrer field can also indicate the domain of an ad-network that has prebought the advertisement slot. Note that the referrer value is missing in 15 % of the entries in our dataset.

day	# of ad requests	# of page visits
July 05, 2015	390 M	293 M
September 30, 2015	520 M	512 M
October 07, 2015	558 M	551 M
October 14, 2015	553 M	536 M
October 21, 2015	560 M	551 M
November 12, 2015	1.94 B	1.68 B
November 19, 2015	2.11 B	1.80 B
November 26, 2015	1.70 B	1.57 B
December 02, 2015	2.14 B	1.94 B

Table 1: Summary of datasets information: date, number of ad requests and number of page visits (aggregate together concurrent ad requests to a given referrer).

Table 1 summarizes the size of the datasets associated to each one of the 9 considered days. In particular, the sizes of the datasets range between 390 M and 2.14 B ad requests per day.

2.1 Overview of Data Processing Infrastructure

Given the large volume of the dataset, we need to define a scalable processing infrastructure to conduct the statistical analyses described in Section 3. To this end we process the original log files (with a Comma Separated Values (CSV) format) to store the information in a PostgreSQL database [24]. PostgreSQL provides a versatile environment rich in datatypes, having data types for IPv4 and IPv6, built-in functions, enhanced SQL syntax, and plugin facilities; while being constrained in resource requirement. We use a relational database schema with a main table for the ad request entries. In addition, we create look-up tables to codify the string fields. We store in the main log table an integer mapping to the corresponding string in its respective look-up table. This serves to save memory and disk space since each string is stored just once, as well as to accelerate common computation operations such as grouping or filtering

[2] https://github.com/apastor/nameless-postgresql

[3] App IDs are strings with the following format com.Company.ProductName.

the data (e.g., checking if two integers are equal is computationally more efficient than comparing two strings). Finally, to compute the fundamental metrics of our methodology directly on the database in an efficient manner, we implemented a plugin aggregate function for postgreSQL in the C programming language.

Using this data processing infrastructure in a standalone LINUX server with 48 GB RAM and 16 cores, we are able to compute the referrers' entropy for the data sample of December 02 including 2.14 B ad requests in less than 20 hours. An alternative solution using a Python Script with parallel computing needed more than 60 hours for the same process.

2.2 Data Preprocessing

When a user visits a page, multiple ads (e.g., embedded in different iFrames) can concurrently be shown to her. Each one of these ad request appears as an independent ad request entry in our dataset. For the purpose of analyzing the quality of traffic of a source (referrer or IP address), we would like to merge together in a single entry all the concurrent ad requests from a specific IP address to a specific referrer since they all correspond to the same page visit. To this end we process the dataset to merge ad requests, which share the same (or very similar) time-stamp, IP address and referrer, which are likely concurrent ad requests.

3. METHODOLOGY FOR THE DETECTION OF LOW QUALITY AD TRAFFIC

Our goal is to define a data-driven methodology which is able to assess the quality of the traffic associated with an individual source of traffic (referrer or IP address) at scale, and then identify sources delivering low quality traffic. This methodology would be tremendously useful for the Adtech industry for complementing existing proprietary filtering solutions [15, 32]. In the rest of the section we will first explain the rationale behind our methodology. Then, we will provide a detailed description of the technical details of the methodology.

3.1 Rationale

The ad traffic of a source (i.e., a referrer or an IP address) is likely to present different patterns when coming from ordinary human activity compared to non-human activity. Some examples of anomalous traffic patterns are as follows: (*i*) A fraudster owning a website where ads are displayed can set-up a few bots. These bots would visit her website and eventually click in some of the displayed ads. The fraudster may receive an economic compensation for ad views, clicks or other monetizable user events associated with the bots. Most of the traffic received by the fraudster's website will be concentrated in the bots' IP addresses resulting in an uncommonly deterministic traffic pattern coming from few IP addresses. (*ii*) A web scraper is a program used to extract content of websites–like auction sites, bet portals, or news websites–simulating the navigation of a human. Some of these web scrappers may visit a site millions of times per day using a headless browser technology for anti-scraping evasion, as a result also triggering ads on the pages the scraper bots visit. The traffic generated by the IP address running such type of web scrapper would present an unusually deterministic pattern concentrated in a single webpage. (*iii*) A major website buying a significant amount of sourced traffic from unreliable sources in the traffic market may show an anomalous traffic profile.

Hence, our goal is to define a method that is able to characterize unusual traffic patterns associated with an ad traffic source, e.g., a referrer. To this end, our methodology relies in a fundamental concept from thermodynamic theory, the *entropy*. The entropy based metric is capable of summarizing the traffic pattern of a referrer in a single value, so that we can compare traffic patterns across millions of referrers in an efficient manner. In addition, it can be computed in a scalable manner[4].

Once we have computed the entropy for every referrer in our dataset, we use statistical analyses such as outlier detection techniques, in order to identify those referrers presenting anomalous traffic patterns. Based on our initial assumption (which will be validated later in the paper), these anomalous traffic patterns are likely associated with low quality referrers delivering low quality traffic. Therefore, we believe that this methodology is an essential addition to the traffic filtering tools currently implemented by vendors in the Adtech industry.

Note that, as indicated above, our methodology can be applied to different types of ad traffic sources: referrers (representing websites and mobile apps) and IP addresses. For the shake of clarity, in the previous example as well as in the rest of the section, we consider the case of referrers and make explicit mention to the case of IP addresses only when needed.

3.2 Methodology Description

Our methodology is divided into three main components: First, we compute the entropy for each referrer in our dataset. Second, we define a normalization process of the entropy value based on the traffic volume associated with each referrer. Finally, we define statistically meaningful thresholds, to separate referrers in different groups based on the quality of their traffic (represented by the normalized entropy score).

3.2.1 Entropy computation

The entropy is a metric originally defined in the context of thermodynamic theory. In this paper, we specifically use the Shannon entropy, extensively used in Information Theory [26]. In essence, the entropy measures the level of randomness (or determinism) of a given process. In the context of this paper, our measurement of entropy in ad exchange data will capture the level of determinism in the pattern of ad requests received by a referrer from a set of IP addresses.

In particular, the traffic pattern associated to a referrer is represented by the ad requests in our dataset and then, it can be defined as a discrete random variable $X = \{x_1, x_2, ..., x_n\}$. X represents the set of the n different IPs sending ad requests to the referrer. The expression to compute entropy measure for a discrete random variable is as follows:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2(P(x_i)) \qquad (1)$$

where $P(x_i)$ is the probability of receiving an ad request from IP x_i.

[4] As described in Section 3, we can compute the entropy of 1.5 M referrers receiving overall 2.1 B of ad requests from 150 M IP addresses in less than 20 hours.

Based on this formulation, the maximum possible value of entropy, $log_2(n)$, is achieved by variables with a uniform probability mass function, where all the values have the same probability. In our case, this correspond to cases in which all IPs sending ad request to a referrer send exactly the same number of requests. In contrast, the minimum entropy value is 0, when the probability mass function is a Kronecker delta, meaning that there is no uncertainty in the expected value as only one of them can actually happen. In our case, this occurs when all the ad requests associated to a referrer come from a single IP address.

We estimate the probabilities of the elements in the dataset by maximum likelihood, $P(x_i) = C(x_i)/C(X)$, where $C(x_i)$ is the number of occurrences of the element x_i and $C(X)$ is the total number of occurrences of all the elements in the set. If we consider the case of a referrer in our dataset, $C(x_i)$ would represents the number of ad requests received by the website from IP$_i$, and $C(X)$ would represent the total number of ad requests associated with the website. By doing so, the entropy measure can be computed directly from the counts of occurrences of the elements and the total count as follows:

$$H(X) = log_2(C(X)) - \frac{\sum_{i=1}^{n} C(x_i)log_2(C(x_i))}{C(X)} \qquad (2)$$

Note that the resulting formula can be easily parallelized for its computation with a map-reduce process, where the operations inside the summation can be implemented in a map function with a final reduce function for adding the values resulting from each element map computation. Besides, storing in advance the joint occurrences of IPs and referrers from the ad-requests stream, we can design a system to compute the entropy from the occurrences count in linear time, which makes our implementation of entropy calculation efficient and highly scalable as reported in Section 2.

3.2.2 Normalized entropy score

As indicated above, the entropy value for a website ranges between 0 and $log_2(n)$. Hence, it is difficult to compare two referrers with a different volume of ad request or receiving ad request from a different number of IPs (represented by n).

Table 2 presents a toy example to illustrate this. In particular, it shows the traffic pattern associated with three different referrers and their entropy values. We observe that *Referrer 2* and *Referrer 3* have the same entropy value because the ad requests in both cases are homogeneously distributed across 5 IP addresses. *Referrer 2* is an unpopular referrer receiving 5 visits from 5 IPs. However, *Referrer 3* is receiving 5 K visits from just 5 IPs, which seems highly suspicious. This simple example demonstrate that we cannot define the quality of a referrer's traffic based exclusively on entropy measurement, without taking into account its volume of traffic.

To address this limitation, we perform a normalization process to obtain a *normalized entropy score* (NES), which takes into account the volume of traffic associated with a referrer. In particular, the NES of a given referrer is computed as the ratio of its entropy and the binary logarithm of the total number of associated ad requests. The formal expression to compute NES is as follows:

	IP_1	IP_2	IP_3	IP_4	IP_5	entropy	NES
Referrer 1	5	0	0	0	0	0	0
Referrer 2	1	1	1	1	1	2.32	100
Referrer 3	1000	1000	1000	1000	1000	2.32	19

Table 2: Entropy of different websites according to the distribution of their visits.

$$NES(X) = 100\left(1 - \frac{\sum_{i=1}^{n} C(x_i)log_2(C(x_i))}{C(X)log_2(C(X))}\right) \qquad (3)$$

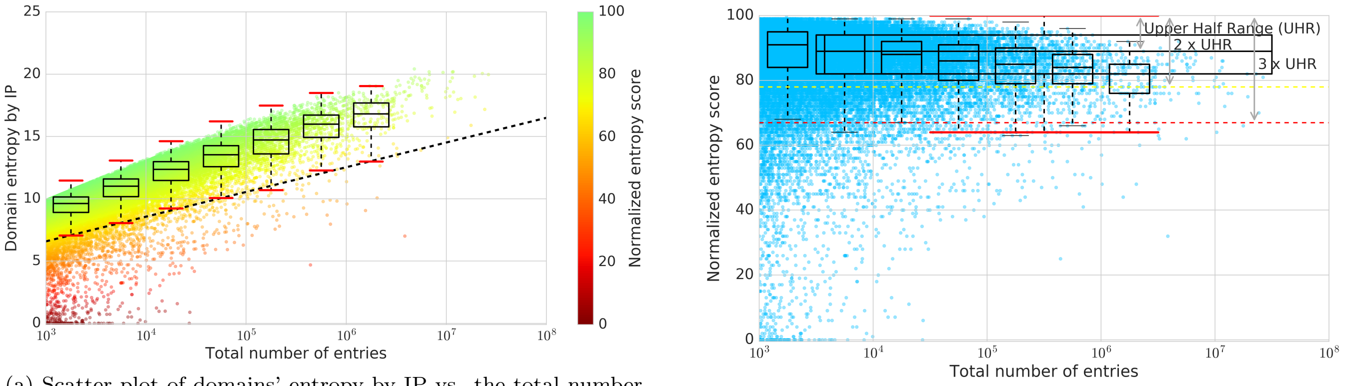
Note that the defined normalization process generates a unique range of values for the NES of any referrer between 0 (min value) and 100 (max value). If we consider the toy example from Table 2, we observe how the proposed normalization process penalizes *Referrer 3* without affecting *Referrer 1* and *Referrer 2*, which still present the minimum and maximum values, respectively.

The presented toy example provides a first intuition to understand the reason and the result of the normalization process. Figure 1 shows the results of the normalization process for the December 02 dataset. In particular, Figure 1a shows the total number of ad-requests (x-axis) vs the entropy-value (y-axis) for each referrer in the dataset. We have grouped the referrer in buckets based on their number of ad requests[5] and for each bucket we compute the distribution of the entropy value and present it in the form of a boxplot in the same figure. The boxplot represents the 25, 50 and 75 percentile values of the distribution as the bottom, mid and top lines of the box, respectively. Moreover, the top and bottom whiskers represent an extensively used outlier threshold (1.5 times the interquartile range) [34]. Hence, any website below or above the whiskers is an outlier with an unusual traffic pattern. In particular, outliers below the bottom whisker present anomalously deterministic traffic patterns whereas outliers above the top whisker show anomalously random patterns. We observe that there are outliers only below the bottom whisker. Then, we configure the threshold to distinguish referrers with anomalously deterministic traffic patterns as the logarithmic regression of the bottom whiskers' values of the seven boxplots. This threshold is represented by the dashed black line in Figure 1a. We observe that the threshold to identify suspicious referrers varies as function of the number of ad request associated with the referrers, then as our toy example illustrated, it may occur that two referrer have the same entropy value, one of them being suspicious and the other one not, which is counter-intuitive.

Figure 1b shows an equivalent to Figure 1a substituting the entropy measure by the NES in the y-axis. We observe how the normalization process equalize the distribution across the different buckets, confirming that NES is a metric independent from the volume of ad requests. Hence, NES in a much more intuitive and practical metric, especially if it has to be used by non technically-skilled users.

Finally, remind the explanation of the entropy computation and normalization processes has been based on referrers.

[5] The defined buckets are evenly spaced in the logarithmic scale: 1 K-3.16 K, 3.17 K-10 K, 10 K-31.62 K, 31.63 K-100 K, etc. Note that the figure does not consider referrers with less than 1 K ad-requests. Moreover, those buckets formed by less than 100 samples are discarded.



(a) Scatter plot of domains' entropy by IP vs. the total number of visits received by the domain. The colorbar shows the corresponding normalized score.

(b) Scatter plot of domains' normalized entropic score by IP vs. the total number of visits received by the domain.

Figure 1: Scatter plots with results of domains with more than 1 thousand visits the December 02. The boxplots show the distribution of the entropy and the score dividing the domains in buckets in function of the number of visits.

The process is exactly the same for IP addresses, considering that an IP address would distribute its ad requests across n different referrers.

3.2.3 *Threshold selection*

The goal of our methodology is to identify referrers and IP addresses with anomalous traffic patterns which are due to low quality traffic. To this end, we have defined above a normalized score, NES, which defines the level of randomness in the ad traffic pattern of a referrer or an IP. In this subsection, we will focus on defining statistically supported thresholds for NES, which allow to identify referrer with anomalous traffic patterns. Note that to compute these thresholds we will rely on outlier detection as well as information theory to reveal referrers with traffic patterns that appear with low probability in our dataset and thus can be considered suspicious. In particular, we will define different thresholds associated with different levels of suspicion. Note, that we will describe the threshold definition methods for referrers and then discuss on the suitability of each of them for the case of IP addresses.

- Outlier detection method: This method has been presented in Section 3.2.2. In particular we compute the distribution of the NES values for the referrers in our dataset and identify the outliers as those referrers with a NES value lower than the **25 percentile - 1.5 IQR**. These websites can be considered extreme outliers and thus have an extremely anomalous deterministic traffic pattern. Hence, we consider them as *highly suspicious*. Note that we do not have any outlier in the upper part of the distribution and thus there is no referrer with extremely anomalous random traffic pattern.

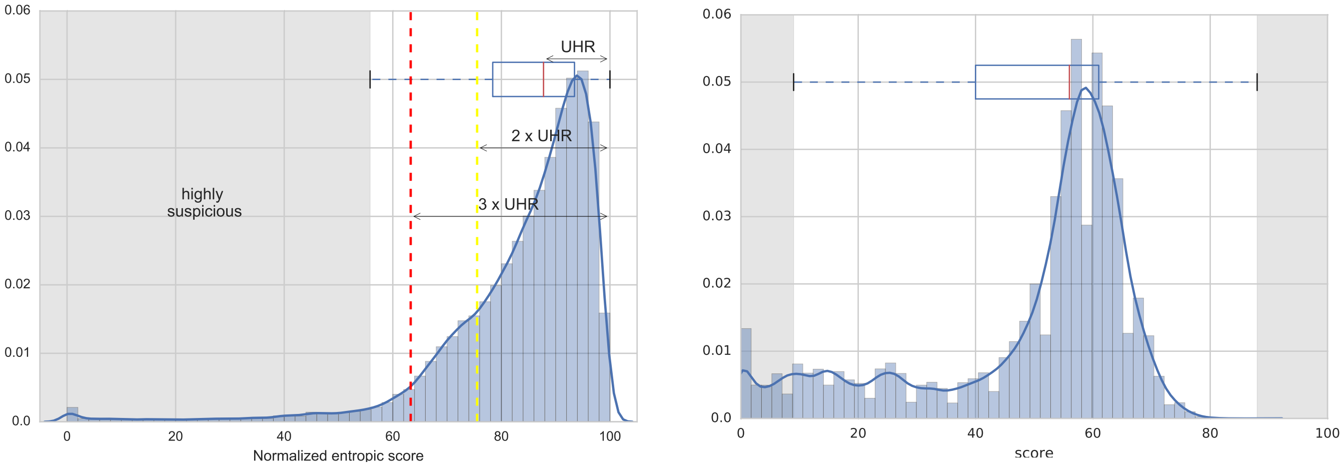
- Dispersion method: The outlier detection method identify *highly suspicious* websites with very low NES scores. We would like to define less conservative thresholds, which still identify websites with relative deterministic traffic patterns. Figure 2a shows the histogram for NES across the domains in our December 02 dataset. We observe a distribution skewed towards high values of NES with a long tail towards low values. In addition, the representation of this distribution in the form of a boxplot, depicted by the large boxplot in Figure 1b shows that there is no outliers on the upper side of the distribution. Hence, we can safely assume that sites with high NES scores present common traffic patterns which are more likely to correspond to high quality human traffic. Then, to define the new thresholds we use as a reference the upper half range (UHR) of the distribution, which is measured as the distance between the median and the maximum value of the NES distribution. Using the UHR as a dispersion metric, we define two different thresholds as **max(NES) - 2 UHR** and **max(NES) - 3 UHR**.

Figures 1b and 2a show graphically the three described thresholds. We observe that the threshold based on outlier detection techniques is the most restrictive followed by the threshold using 3 UHR and 2 UHR as dispersion distance, respectively. The definition of these three thresholds allows us to classify the websites in our dataset in 4 categories as: *highly suspicious* when the NES value of the website is below the outlier detection threshold; *suspicious* when the NES value falls between the outlier detection and the 3 UHR threshold; *likely suspicious* when the NES value falls between the 3 UHR and 2 UHR thresholds; *legit* when the NES value is over the 2 UHR threshold.

Finally, Figure 2b shows the histogram of the NES value across the IP addresses in our December 02 dataset. We observe that contrary to the referrers' distribution concentrated in high NES values, the distribution of NES values of IP addresses concentrates in central values of the range with a heavy tail towards low values and a much less significant tail towards high values. This shape of NES distribution for IP addresses avoid us to apply the two thresholds obtained from the described dispersion method. Hence, in the case of IP addresses we will consider *suspicious* only those IP addresses showing a NES value below the threshold defined by the outlier detection method.

4. VALIDATION

The validation of the performance of a detection methodology, as the one proposed in this paper, is ideally conducted based on ground truth dataset, which allows computing performance metrics including the recall, the precision or the F-score. Unfortunately, in digital advertising we do not find such ground truth dataset. Therefore, to validate our methodology, we will rely in external metrics used in the



(a) Domains score distribution.

(b) IPs score distribution.

Figure 2: Distribution of the score of the domains and IPs with more than 1 thousand entries the December 02.

Adtech industry and that are associated to the quality of traffic of a referrer or an IP address.

4.1 Metrics for Quality of Referrer's traffic

There exist several metrics which are associated to the traffic quality of a referrer. In particular, for the validation of our methodology we consider the following ones:

Bounce Rate: This metric measures the fraction of sessions in a referrer, which visit a single page. A low bounce rate is a strong indication of low quality traffic.

Traffic from popular publishers: This metric represents the percentage of upstream traffic coming to the website from popular publishers. In particular, the two publishers contributing a larger fraction of traffic to referrers are Google and Facebook. Then, for our validation we will compute the fraction of upstream traffic coming from Google and Facebook to a referrer. A very low fraction of traffic coming from Google and Facebook may reveal the presence of low quality traffic.

Search Traffic: This metric measures the percentage of traffic coming from search engines. A very low search traffic percentage is often an indication of low quality traffic.

Direct Traffic: This metric measures the percentage of traffic that visit the referrer directly without being redirected from other website. In this case, a large fraction of direct traffic is usually linked to low quality traffic.

Number of sites linking to a referrer: Interesting referrer attracting high quality traffic would typically be linked from a large number of other sites. Contrary, referrers associated to ad fraud or other malicious practices, which offer low quality traffic, would typically be linked from a lower number of sites.

We have queried two well-known services, Alexa [2] and SimilarWeb [27], to obtain these metrics for the referrers in our December 02 dataset with more than 1K associated ad requests. Note that not all the metrics were offered by both services. Table 3 present the median and IQR values for the distribution of each one of these metrics for the different groups of referrers identified by our methodology: *highly-suspicious*, *suspicious*, *likely suspicious* and *legit*. In addition, the table shows the relative difference of the median values of these metrics for each one of the suspicious groups in comparison to the legit group.

We can observe substantial relative differences (over 100% in some cases) between the *highly suspicious* and *suspicious* groups and the *legit* group for all the considered metrics. The relative differences are smaller, but still significant for most metrics, when comparing the *likely suspicious* and the *legit* groups.

4.2 Metrics for Quality of IP address' traffic

Associations responsible for defining the guidelines to fight fraud such as the MRC (US) or the JICWEBS (UK) include data centers' traffic as a common source of invalid traffic (with some exceptions as servers providing VPN services) and recommend to filter such traffic [9, 16]. Hence, to get some insight about the performance of our methodology to identify low quality traffic, we have computed the fraction of IP addresses from the *suspicious* and *legit* groups belonging to data centers. To this end we use two public available lists of cloud and hosting providers' IP prefixes [5, 13]. This validation exercise reveals that, in average, 8.4 % of IP addresses in the *suspicious* group belong to data centers and thus are considered to generate invalid traffic. This value shrinks to the 3.6 % for the *legit* group. Moreover, IPs belonging to data centers are responsible of 11.9 % of the traffic of the *suspicious* group, while in the *legit* group this value decrease to 5.49 %.

Finally, we have conducted a cross validation exercise between suspicious IP addresses and suspicious referrers. In particular, we have computed the fraction of ad requests generated by data center IP addresses directed to the different groups of *suspicious* referrers as well as the group of *legit* referrers. We observe that 99% of ad requests generated by data center IPs are associated to the *suspicious* referrers groups, and, in particular, 84% of these requests are directed to referrers in the *highly suspicious* group.

In summary, we can conclude that the results reported in this section provide solid evidences that our methodology is able to properly classify referrers and IPs based on the quality of their traffic in meaningfully defined groups.

		highly suspicious		suspicious		likely suspicious		legit
Alexa Upstream traffic from Google and Facebook (%)	median	10.3	(-185%)	12.4	(-137%)	21.5	(-37%)	29.4
	IQR	28.0		27.3		28.2		29.9
Alexa Bounce rate (%)	median	48.7	(-16%)	54.3	(-4%)	57.0	(+1%)	56.6
	IQR	35.0		40.9		39.5		30.1
Alexa Search traffic (%)	median	9.2	(-45%)	8.2	(-62%)	10.2	(-30%)	13.3
	IQR	15.5		12.4		12.5		18.6
Alexa Total sites linking to the domain	median	87	(-300%)	131	(-166%)	256	(-36%)	348
	IQR	616		371		800		1,198
SimilarWeb Bounce rate (%)	median	45.6	(-23%)	48.6	(-16%)	53.3	(-5%)	56.2
	IQR	34.1		33.9		31.5		27.7
SimilarWeb Direct traffic (%)	median	34.4	(+22%)	39.5	(+32%)	27.0	(+0%)	26.9
	IQR	42.6		41.1		32.8		32.6
SimilarWeb Search traffic (%)	median	15.2	(-95%)	13.9	(-114%)	22.8	(-30%)	29.6
	IQR	42.0		34.8		33.8		45.3

Table 3: Value of external metrics associated to quality of referrers' traffic quality for the different *suspicious* groups and the *legit* group of referrers reported by our methodology for the December 02 dataset.

5. RESULTS

Our large-scale datasets include a representative fraction (approximately between 0.2 and 1 %) of the total number (∼200 B) of daily ad-requests delivered in the programmatic advertising ecosystem [1]. Therefore, applying our detection methodology to these datasets will provide statistically significant insights about the volume of low quality traffic present in this ecosystem. In particular, we will first present results for referrers and IP address separately and conclude the section providing an overall estimation of the low quality traffic volume.

5.1 Referrers' results

Table 4 shows the fraction of referrers and their associated traffic belonging to the three defined *suspicious* groups and the *legit* group for our 9 daily datasets. We observe that (in average) 18.49 % of referrers are considered suspicious. In addition, these sites are responsible (in average) for 40.7 % of the daily programmatic ad-requests. Moreover, if we focus exclusively on the *highly suspicious* referrers, they represent 6 % of all referrers and 9.92 % of all ad requests.

5.2 IP addresses' results

In the case of IP addresses, we only have two groups, *suspicious* and *legit* IPs. Note that in this case the *suspicious* group correspond to highly suspicious IPs since they have been computed using the threshold defined with the outlier detection technique. In addition, we assume that Adtech vendors follow the recommendations from trade organizations in the assessment of traffic quality (e.g., the MRC (US) or the JICWEBS (UK)) and filter all data center traffic. Indeed, following the guidelines of these organizations, we have filtered out all data center IPs since their traffic is considered invalid. From the remaining IP addresses, we have computed the fraction of IPs belonging to the *suspicious* and the *legit* groups as well as the fraction of ad requests associated to IPs in each of these groups for each of our daily datasets. The results indicate that (in average) 8.7 % of IP addresses are classified as suspicious by our methodology. In addition, these IPs are responsible (on average) for 8.3 % of the daily ad requests.

		highly susp.	suspicious	likely susp.	total
July 05	referrers	7.38 %	7.67 %	11.76 %	26.82 %
	traffic	17.27 %	13.44 %	45.06 %	75.77 %
September 30	referrers	5.66 %	4.10 %	12.98 %	22.75 %
	traffic	14.97 %	11.76 %	24.85 %	51.57 %
October 07	referrers	6.18 %	0.88 %	10.68 %	17.74 %
	traffic	12.51 %	3.74 %	25.49 %	41.74 %
October 14	referrers	5.95 %	0.82 %	10.96 %	17.72 %
	traffic	14.01 %	2.00 %	22.57 %	38.58 %
October 21	referrers	6.11 %	0.81 %	11.09 %	18.01 %
	traffic	13.26 %	3.94 %	21.16 %	38.36 %
November 12	referrers	6.49 %	0.56 %	10.94 %	17.99 %
	traffic	5.15 %	0.57 %	25.29 %	31.01 %
November 19	referrers	5.38 %	1.16 %	10.39 %	16.93 %
	traffic	3.31 %	2.03 %	20.41 %	25.74 %
November 26	referrers	6.18 %	2.53 %	11.95 %	20.65 %
	traffic	4.74 %	2.49 %	26.14 %	33.37 %
December 02	referrers	4.71 %	1.32 %	10.79 %	16.82 %
	traffic	4.08 %	2.54 %	23.54 %	30.16 %
average	referrers	6.00 %	2.21 %	11.28 %	19.49 %
	traffic	9.92 %	4.72 %	26.06 %	40.70 %

Table 4: Percentage of referrers in each of the suspicious groups and the their associated traffic of ad-requests.

5.3 Overall results

To estimate the overall suspicious traffic, we perform an AND operation between the ad requests associated with suspicious websites and IPs that are either suspicious or from a data center. By doing so, if an ad request appears associated to a suspicious referrer and a suspicious IP we will just count it once, since it is actually the same request.

A conservative estimation of the overall suspicious traffic would consider only as suspicious the referrers in the *highly* *suspicious* group. In this case, our methodology indicate that ∼15 % of all ad requests can be considered low quality traffic. Using this figure and given that the investment in programmatic advertising is estimated to be $22 B for 2016 only in US [11], we can make a ball park estimation that advertisers would face losses over $3 B associated to low quality programmatic media buys in 2016 only in US. This figure is comparable to the revenue losses generated by major cyber-crimes such as tax-refund fraud or corporate account takeover worldwide [28]. Note that a less conservative estimation considering referrers in all the defined *suspicious* groups indicate that low quality traffic represents ∼40 % of ad requests in our dataset leading to a ball park estimation of losses for advertisers in US over $8.5 B. These results are in line with previous studies which estimate enormous economic losses for advertisers due to low quality traffic and ad fraud [23, 33].

6. RELATED WORK

The obvious performance and economic implications associated to low quality and fraudulent advertising traffic, has motivated the research community but also the industry to propose solutions to mitigate this problem. Adtech vendors have opted for proprietary solutions [15, 32], which contribute to the lack of transparency of the digital advertising ecosystem and avoid the possibility of conducting proper performance evaluations. Instead, the works we found in the research community openly describe the proposed solutions [21, 30], which contribute to build a body of knowledge to fight fraud. However, this initial effort has not lead yet to a collaborative open source development, which as demonstrated in other areas as network security [29], contributes to fight against malicious activities more efficiently. Next, we briefly discuss some of the most relevant related work.

In the recent years we observe an increasing body of work relying on data analysis and pattern recognition techniques for fraud mitigation in online advertising [10, 18]. As a consequence, the methods used by ad fraud cyber-criminals, have evolve in sophistication as the industry and the research community identify attacks and develop mitigation techniques [3, 20, 31].

Despite the existence of sophisticated attacks, the fraction of fraud using simple techniques is still relevant. In a recent paper [6], Callejo et al. analyze the presence of bot traffic from data centers on different campaigns they run in Google AdWords. The authors identify up to a 10 % of this type of fraud, depending the targeted keywords.

Besides, as cost per click (CPC) advertisements usually report higher profits than cost per mille (CPM) impressions, fraudsters simulate click events in websites and mobile apps. Miller et al. [22] and Cho [8] propose techniques for click fraud mitigation. For instance, serving a percentage of transparent or unattractive ads and analyzing the click through rate to identify the publishers committing artificial clicks.

The solutions proposed in these studies focus on specific types of fraud attacks such as click fraud, or propose solutions to be applied at the level of individual campaigns. Conversely, we propose a solution to address the problem at scale in the programmatic ecosystem, transversely to any type of advertisement.

Finally, from a methodology perspective we find a previous work by Chen et al. where entropy has been used in the area of ad fraud research, but in different context applied, to detect fake views in online video services [7]. The authors of this paper propose to use entropy as the final metric to assess the traffic quality and semi-supervised classification that rely on manually labeled samples to differentiate valid and invalid traffic. As we describe in Section 3 a limitation of using directly entropy as a metric of quality is that its interpretation depends on the volume of events associated. To overcome this limitation, we propose a novel normalization process to obtain a normalized score. In addition, our methodology uses statistical supported outliers detection methods without the need of manual labeling of suspicious traffic. Hence, although both papers are based on the same fundamental concept of entropy, the methodology built on top of this concept is significantly different.

7. CONCLUSIONS

In this paper we present a methodology for the identification of low quality ad traffic at scale. This methodology relies on the concept of entropy to classify the ad traffic associated to websites, mobile apps and IP addresses showing anomalously deterministic patterns as low quality ad traffic. The application of this methodology to one of the largest ad traffic datasets used in the research community to date suggests that low quality traffic represents (at least) 15 % of programmatic media buys. Based on this, the associated economic losses for US advertisers are estimated to be comparable to revenue losses generated by some major cyber-crimes worldwide.

To assist the Ad Tech industry in the adoption of the described methodology as well as to help researchers interested in re-using it, we have implemented a scalable version of our methodology and make it available as open source code. In addition, we are currently working in the implementation of a system prototype that will use the described methodology to deliver an updated list of websites, mobile apps and IP addresses and their associated normalized entropy score as well as their category based on the threshold scheme defined in this paper. We believe this list may be of high value for a variety of players (advertisers, DSPs, Ad Exchanges, SSPs, publishers, etc.) from the Ad Tech ecosystem.

References

[1] World Federation of Advertisers. *Compendium of ad fraud knowledge for media investors.* 2016.

[2] *Alexa: actionable analytics for the web.* URL: http://www.alexa.com (visited on 10/01/2016).

[3] Sumayah A Alrwais et al. “Dissecting ghost clicks: Ad fraud via misdirected human clicks”. In: *Proceedings of the 28th Annual Computer Security Applications Conference.* ACM. 2012.

[4] The World Bank. *The World Bank: Household final consumption expenditure, etc. (% of GDP).* URL: http://data.worldbank.org/indicator/NE.CON.PETC.ZS (visited on 10/24/2016).

[5] Botlab. *Deny Hosting IP repository.* URL: https://github.com/botlabio/deny-hosting-IP (visited on 09/01/2016).

[6] Patricia Callejo et al. *Independent Auditing of Online Display Advertising Campaigns.* 2016.

[7] Liang Chen, Yipeng Zhou, and Dah Ming Chiu. "Fake view analytics in online video services". In: *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop*. ACM. 2014.

[8] Geumhwan Cho et al. "Combating online fraud attacks in mobile-based advertising". In: *EURASIP Journal on Information Security* (2016).

[9] Media Rating Council. *Invalid Traffic Detection and Filtration Guidelines Addendum*. 15, 2015.

[10] Vacha Dave, Saikat Guha, and Yin Zhang. "Viceroi: Catching click-spam in search ad networks". In: *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM. 2013.

[11] eMarketer. *More Than Two-Thirds of US Digital Display Ad Spending Is Programmatic*. 5, 2016. URL: http://www.emarketer.com/Article/More-Than-Two-Thirds-of-US-Digital-Display-Ad-Spending-Programmatic/1013789 (visited on 10/24/2016).

[12] Hewlett Packard Enterprise. *The Business of Hacking*. 2016.

[13] Nick Galbreath. *ipcat: datasets for categorizing IP addresses*. URL: https://github.com/client9/ipcat (visited on 09/01/2016).

[14] GroupM. *Interaction 2016: Integrity in the Digital Marketplace*. URL: https://www.groupm.com/news/interaction-2016-integrity-in-the-digital-marketplace (visited on 10/24/2016).

[15] *Integral Ad Science: Products*. URL: https://integralads.com/capabilities/ad-fraud (visited on 10/20/2016).

[16] JICWEBS. *UK Traffic Taxonomy for Digital Display Advertising*. 15, 2015.

[17] John Philip Jones. *When ads work: New proof that advertising triggers sales*. ME Sharpe, 2006.

[18] Brendan Kitts et al. "Click fraud detection: adversarial pattern recognition over 5 years at Microsoft". In: *Real World Data Mining Applications*. Springer, 2015.

[19] Miriam Marciel et al. "Understanding the Detection of View Fraud in Video Content Portals". In: *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016.

[20] Wei Meng et al. "Your Online Interests: Pwned! A Pollution Attack Against Targeted Advertising". In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014.

[21] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. "Detectives: detecting coalition hit inflation attacks in advertising networks streams". In: *Proceedings of the 16th international conference on World Wide Web*. ACM. 2007.

[22] Brad Miller et al. "What's Clicking What? Techniques and Innovations of Today's Clickbots". In: *Proceedings of the 8th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer-Verlag, 2011.

[23] Adrian Neal, Sander Kouwenhoven, and Oxford BioChronometrics SA. *Quantifying online advertising fraud: Ad-click bots vs humans*. Tech. rep. tech. rep., Oxford Bio Chronometrics, 2015.

[24] *PostgreSQL*. URL: http://www.postgresql.org (visited on 08/11/2016).

[25] PwC. *Global entertainment and media outlook 2016-2020: US edition*. URL: http://www.pwc.com/us/en/industry/entertainment-media/publications/global-entertainment-media-outlook.html (visited on 10/01/2015).

[26] C. E. Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* (1948).

[27] *SimilarWeb: Website Traffic & Mobile App Analytics*. URL: https://www.similarweb.com (visited on 10/01/2016).

[28] Tommie Singleton. "The Top 5 Cybercrimes". In: *American Institute of CPAs* (2013).

[29] *Snort - Network Intrusion Detection & Prevention System*. URL: https://www.snort.org (visited on 10/21/2016).

[30] Ori Stitelman et al. "Using co-visitation networks for detecting large scale online display advertising exchange fraud". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013.

[31] Brett Stone-Gross et al. "Understanding Fraudulent Activities in Online Ad Exchanges". In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*. ACM, 2011.

[32] *White Ops: Products*. URL: http://www.whiteops.com/products (visited on 10/20/2016).

[33] ANA & WhiteOps. *2015 Bot Baseline: Fraud in Digital Advertising*. 2016.

[34] Rand R Wilcox. *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Springer Science & Business Media, 2010.