Impression Fraud in Online Advertising via Pay-Per-View Networks

Kevin Springborn
Broadcast Interactive Media
kspringborn@bimlocal.com

Paul Barford
Broadcast Interactive Media
University of Wisconsin-Madison
pbarford@bimlocal.com

ABSTRACT

Advertising is one of the primary means for revenue generation for millions of websites and mobile apps. While the majority of online advertising revenues are based on pay-per-click, alternative forms such as impression-based display and video advertising have been growing rapidly over the past several years. In this paper, we investigate the problem of invalid traffic generation that aims to inflate advertising impressions on websites. Our study begins with an analysis of purchased traffic for a set of honeypot websites. Data collected from these sites provides a window into the basic mechanisms used for impression fraud and in particular enables us to identify *pay-per-view (PPV) networks*. PPV networks are comprised of legitimate websites that use JavaScript provided by PPV network service providers to render unwanted web pages "underneath" requested content on a real user's browser so that additional advertising impressions are registered. We describe the characteristics of the PPV network ecosystem and the typical methods for delivering fraudulent impressions. We also provide a case study of scope of PPV networks in the Internet. Our results show that these networks deliver hundreds of millions of fraudulent impressions per day, resulting in hundreds of millions of lost advertising dollars annually. Characteristics unique to traffic delivered via PPV networks are also discussed. We conclude with recommendations for countermeasures that can reduce the scope and impact of PPV networks.

1. INTRODUCTION

Advertising is one of the primary methods for generating revenues from websites and mobile apps. A recent report from the Internet Advertising Bureau (IAB) places ad revenues in the US for the first half of 2012 at $17B, which represents a 14% increase over the previous year [15]. While the majority of that revenue is search-based, ad words advertising, display and video advertising have been growing. Indeed, a recent report places display and video advertising in the US at $12.7B for FY2012, growing at 17% annually [27]. At a high level the basic notion of selling space on web pages and apps for advertising is simple. However, the mechanisms and infrastructure that are required for online advertising are highly diverse and complex.

The online ad ecosystem can roughly be divided into three groups: *advertisers, publishers* and *intermediaries*. Advertisers pay publishers to place a specified volume of creative content with embedded links (*i.e.,* text, display or video ads) on websites and apps. Intermediaries (*e.g.,* ad servers and ad exchanges) are often used to facilitate connections between publishers and advertisers. Intermediaries typically place a surcharge on the fees paid by advertisers to publishers for ad placements and/or ad clicks. What is immediately obvious from this simple description is that publisher and intermediary platform revenues are directly tied to the number of daily visits to a website or app. Thus, there are strong incentives for publishers and intermediaries to use any means available to drive user traffic to publisher sites.

There are certainly legitimate methods for traffic generation for publisher sites. The most widely used are the text-based ad words that appear in search results *e.g.,* from Google or Bing. However, it can be quite difficult and expensive to drive large traffic volumes to target sites using ad words alone.[1] Thus, other methods for traffic generation have emerged, many of which are deeded as fraudulent by advertisers and intermediaries. Google defines invalid (fraudulent) traffic as follows:[2]

> *Invalid traffic includes both clicks and impressions that Google suspects to not be the result of genuine user interest [21].*

Standard methods for generating invalid traffic includes *(i)* using employees at publisher companies to view sites and click on ads, *(ii)* hiring 3rd parties to view sites and click on ads, *(iii)* click/view pyramid schemes and *(iv)* using software and/or botnets to automate views/clicks [21]. The challenges for advertisers and intermediaries focused on offering trustworthy platforms are to understand these and potentially other threats so that effective countermeasures can be deployed.

In this paper, we investigate a relatively new threat for display and video advertising called Pay-Per-View (PPV) net-

[1] This has led to the emergence of a large number of Search Engine Optimization companies in recent years.

[2] While Google is not the only company in this domain, we refer to them as an authoritative source of information due to their size and experience in online advertising.

works. The basic idea for PPV networks is to pay legitimate publishers to run specialized JavaScript when users access their sites that will display other publishers websites in a camouflaged fashion. This will result in impressions and potentially even clicks that are registered on the camouflaged pages without "genuine user interest" *i.e.,* invalid traffic generation. Legitimate publishers view this as another way to monetize their sites without impact to their users. PPV networks sell their traffic generation capability by touting real and unique users, geolocation and context specificity among other things. The fact that pages are appearing on real users' systems makes detecting and preventing PPV traffic generation challenging.

To study PPV networks, we employ a small set of honeypot websites that we use as the target for traffic generation. These sites were constructed to include what appears to be legitimate content and advertising. We then use search to identify a wide variety of traffic generation offerings on the Internet. We purchased impressions for our honeypot sites in various quantities from a selection of different traffic generation services over the course of a 3.5 month period. By engaging with traffic generation services directly, we were able to uncover the basic mechanisms of PPV networks and initiate additional measurements to characterize their deployments.

The characteristics of the traffic purchased for our honeypot sites is dictated at a high level by the service offerings, which enable volume, time frame and geographic location, etc. of users to be specified. Our results show that impressions are typically spread in a somewhat bursty fashion over the specified time frame and that user characteristics are well matched with specifications. By considering the referer field of the incoming traffic, we were able to identify the fact that our honeypot sites were being loaded into a frame (along with as many as ten other sites) for display on remote systems. By considering names of a small selection of traffic generation services, we use a recent, publicly-available, Internet-wide web crawl to identify the scope of PPV networks. We find tags from these services are, in fact, widely deployed – on tens of thousands of sites. By appealing to MuStat [29], we conservatively estimate the number of invalid impressions that are generated from this small set of PPV networks to be on the order of 500 million per day. Assuming a modest quality level for sites that are part of PPV networks, we estimate the annual cost to advertisers for this invalid traffic to be on the order of $180 million annually.

Finally, we offer three different methods to defend against PPV networks. First, observing viewport dimensions of ad requests can determine if the end user can possibly view the advertisement. In an effort to increase traffic, PPV networks commonly display destinations in zero sized frames. Second, blacklists of websites that participate in PPV networks can potentially be used. The idea is to block advertising on websites that commonly receive PPV traffic until the publisher discontinues purchasing PPV traffic. Such blacklists can be compiled through programmatic enumeration of PPV destinations. Finally, referer fields can be queried at the time of advertisement load in order to identify traffic originating from known PPV domains.

The remainder of this paper is organized as follows. In Section 2 we provide a description of the online advertising ecosystem and an overview of invalid traffic generation threats. In Section 3, we describe the details of our honeypot websites and our traffic purchases for these sites. In Section 4, we describe the details of the evaluations that we conduct on our data including analyses of additional data sets and measurements that enable us to project some of the broader characteristics of PPV networks. We provide recommendations for counter measures that can be employed to reduce the impact of PPV networks in Section 5. We discuss prior studies that inform our work in Section 6. We summarize, conclude and discuss future work in Section 7.

2. ONLINE ADVERTISING ECOSYSTEM

In this section we provide an overview of the online advertising ecosystem including both the business framework and technical framework for delivering advertisements to publisher websites and apps. Some prior studies have provided similar overviews including [16, 34, 41]. We also provide an overview of invalid traffic generation threats and the challenges they pose in the ecosystem.

2.1 Business Framework

As mentioned in Section 1, there are three main participant groups in ad networks: advertisers, intermediaries and publishers. As shown in Figure 1 there are two other important groups: brands and users. Brands pay advertisers to help them sell their products and services. Internet-based campaigns are attractive to brands and advertisers since consumers/users spend a growing proportion of their time online. An important appeal of online advertising (especially for consumer goods) is that it offers the opportunity to tie ad campaigns and associated costs directly to sales *e.g.,* by tracking clicks from online ads to purchases on a brand's ecommerce site.

Advertisers are companies that create and manage advertising campaigns for brands. Advertisers pay publishers to make ad placements on websites and apps using one of several different models. One is the widely used Pay-Per-Click (PPC) model, where an advertiser only pays a publisher for an ad when a user clicks on it. PPC campaigns are typically associated with ad words (short, text-based ads) campaigns. An alternative payment method that is common in display and video advertising is Cost Per Mille/Thousand (CPM), where advertisers pay publishers whenever users *view* an ad (CPM prices are given per thousand impressions). The CPM-based payment model is the primary focus for this paper. The goal for advertisers is to place ads on sites that they believe attract a brand's target demographic in a cost-effective fashion. Thus, their challenge is in identifying these

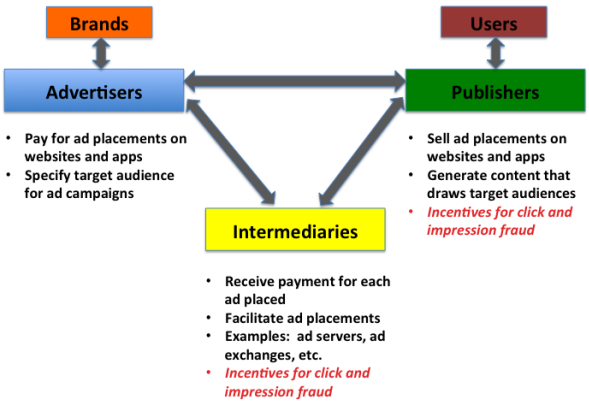

Figure 1: Key participants in the online advertising ecosystem. Payments flow from brands to advertisers to intermediaries and publishers.

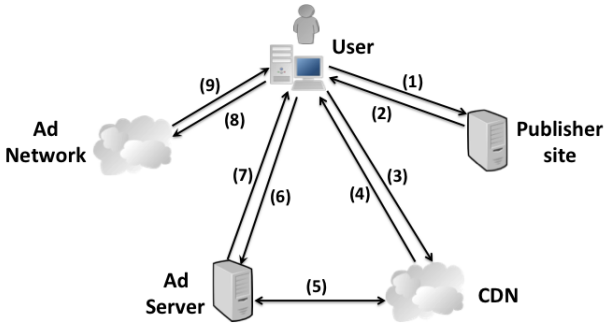

Figure 2: Typical data exchanges required to render an ad in a user's browser. (1) User request to publisher page. (2) Base page delivered. (3) Ad tag request to CDN. (4) JavaScript delivered. (5) Update to JavaScript in CDN if necessary. (6) Request to ad server. (7) Redirected request delivered. (8) Request to exchange or 3rd party ad server. (9) Ad creative delivered.

sites and facilitating ad placement.

In addition to working with publishers directly, advertisers often work with intermediaries in order to actually place ads on websites and apps. The two main reasons for this are the complexity of Internet advertising's technical landscape (see below) and the enormous and growing diversity of websites and apps. Among other things, intermediaries offer "one-stop shopping" for advertisers, and competitive CPM rates to publishers who may not be able to fill all of their placements via direct campaigns.

The scope of intermediaries is quite broad. The most common offerings include targeting services, ad servers and ad exchanges to facilitate placements. One of the most widely used intermediaries in the display advertising space is Google AdExchange (AdX) [20,30]. The revenue model that is most commonly used by intermediaries is to take a small CPM payment for each ad that they participate in serving and then to pass the remainder of the CPM paid by the advertiser to the publisher.

Internet publishers are companies that create content that is of interest to users. Publishers display ads on their pages using standard sized creatives that typically appear in an iframe. A publisher's goal is to maximize their revenue yield by attracting *(i)* premium advertisers that pay high CPM's and *(ii)* a high volume of users, some whom will click through on ads. It is important to note that while ad words-based advertising (*e.g.,* through AdSense) is widely available, display and video ads are typically only available to sites that have somewhat higher volumes of users.

2.2 Technical Framework

Displaying an advertisement on a publisher's page includes potentially a large number of data exchanges between participants in the advertising ecosystem. A simple example is depicted in Figure 2. The process begins with the placement of an *ad tag* in a section of a publisher page. Ad tags (often supplied by intermediaries that manage ad servers) are simple HREF strings that typically reference JavaScript code hosted in a CDN infrastructure.

The JavaScript typically gathers context keywords and other information from the publisher page or user browser and then sends an ad request to the target ad server infrastructure. Ad servers process the ad request and either respond with an ad directly (*e.g.,* from a direct advertiser campaign) or send a redirect to a third party such as an ad exchange. The redirect is forwarded by browser to the target server or exchange, which will respond with an ad that is rendered in the browser. The redirect usually includes sufficient information for ad targeting and billing. This entire process must take place quickly (typically on the order of tens of milliseconds) in order to ensure a good user experience. When the ad is delivered, an impression is registered for the ad serving entity. Click tracking is typically managed by directing clicks to the ad server, which then redirects to the advertiser.

2.3 Invalid Traffic Generation Threats

Impression-based advertising has a number of potential threats. The focus of this paper is on traffic generation that causes invalid impression and thereby inflates publisher and (some) intermediary revenues. Specifically, we focus on invalid traffic generation via PPV networks, which we describe in detail in Section 4.

Valid methods for traffic generation include search and ad words-based advertising. However, web search reveals that there is a wide variety of other traffic generation offerings available. Many offer a specified volume of traffic at a target site over a specified time period. Many also include guarantees of specific features in the traffic such as geographic locations of host systems. Most do not describe their method-

ology in detail if at all. One of the important objectives of traffic generation is that it appear to come from real users. Appealing to the definition of invalid traffic given in Section 1 above, there are many ways in which such traffic might be generated.

Common methods for invalid traffic generation have been borrowed directly from click generation services that have been offered for some time. Examples include hiring people to view pages, bots of various types, and using expired domains to divert users to 3rd-party pages.

PPV networks are sites that load 3rd-party pages in an obfuscated fashion when accessed by users. Publishers become part of a PPV network simply by placing a tag on their site that looks very much like a standard ad tag. We define a "network" as a series of sites that run tags from the same PPV service. Participating publishers are paid on a CPM basis for something that appears to be low or no impact on their site.

Since the third party pages that are rendered via PPV networks are clearly not the interest of the users, all of the resulting impressions are invalid. Beyond laking the intent necessary to qualify as valid traffic, we show that PPV network traffic has characteristics unlike organic traffic. For example, natural traffic displays a diurnal traffic pattern, while the PPV traffic we observed often showed highly artificial delivery patterns.

3. DATA COLLECTION ON HONEYPOT WEBSITES

To begin our investigation of traffic generation and impression fraud we established a set of honeypot websites. We then purchased traffic from a number of different services and captured a diverse set of data from the resulting hits on our sites. In this section we provide details on our honeypot websites and traffic purchases. The results of these activities are described in detail in Section 4.

3.1 Honeypot Websites

We created three websites as the starting point for our investigation of traffic generation service providers. The sites differed only in styling, formatting, and deployment. The content on each site was identical. The reason for creating three different sites was to enable us to conduct A-B comparisons between different traffic generation services.

The design objective for our honeypots was to create sites that looked relatively "legitimate". To that end, they have a standard layout, content changes regularly and the deployment is standard. A second objective was that the sites were instrumented to gather as much data as possible on arriving traffic.

Each site consisted of a base landing page and four subpages. Three of the pages displayed RSS content from the news feeds of *topwirenews.com* or *espn.com*. One page listed links to popular news sites. The final page was a nonfunctional search result. Every page contained four advertisement placements, identical to standard CPM placements except they contained dummy creatives instead of displaying paying advertisers' placements. All of the ads have embedded links to dummy landing pages that we also monitor.

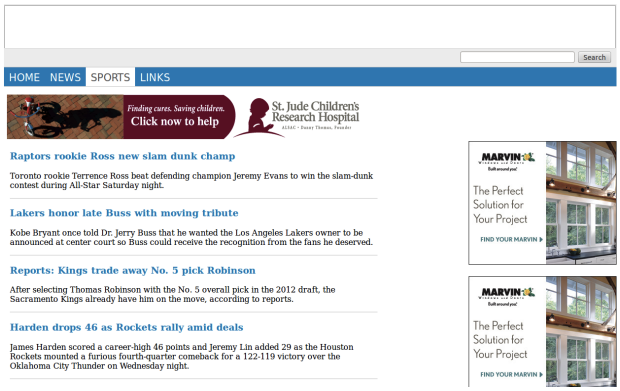

Figure 3: Screenshot of one of the honeypot websites that was a target for traffic generation purchases.

Domain names were registered for each site with GoDaddy using their anonymous registration option. We attempted to give the sites names that sounded interesting and connoted the news-related content of the sites. The sites were created using dotCMS inside Amazon EC2. Amazon's CloudFront CDN was enabled for the sites in order to handle larger bursts of traffic. We used a "noindex, nofollow" meta tag and a robots.txt file to attempt to prevent inclusion in search engine results.

Instrumentation was facilitated in several ways. Google Analytics tags were deployed on all pages for general monitoring. Logs from the serving infrastructure were used to understand the details of individual connections. A series of JavaScript blocks collected information about the site visitors. The instrumentation reported viewer characteristics (See Table 1) using 1x1 pixels. Each advertisement on the sites was instrumented with code that reported the three key events in the life cycle of every ad: *(1)* JavaScript load *(2)* JavaScript execution and *(3)* successful delivery. Finally, the pages contained JavaScript that tracked user interaction on the site. Simliar to [41] the interaction metrics reported mouse movements and clicks. The mouse position was collected every time the cursor moved at least 20 pixels.

3.2 Purchased Traffic

We identified and reviewed 34 traffic generation service providers for this study. These service providers were identified using web search. We manually reviewed each service provider's site to catalog available purchasing options. Details of the sites and options are given in Table 3. We make no claims on the completeness of this list of traffic generation service providers. However, given the commonality of their offerings, we believe that they are a representative cross section.

We also investigated the service provider websites themselves to gain some insights on their legitimacy. Their do-

main names were checked with McAfee SiteAdvisor [6]. The DNS record was inspected using Network Solutions' Whois tool [8]. Finally, a tool available from SameID.net [9] was used to search for sites sharing the same IP address or Google Analytics tag.

Table 1: Visitor information collected from honeypot websites.

Timestamp	Client IP
URL	User Agent
User UID	Page Load UID
Viewport Dimensions	Referer

From the set of 34 traffic generation services, we selected 5 from which we made purchases. Services were selected to get a diversity of delivery rates and price points. The characteristics of our purchased traffic indicated the selected services were independent networks. The purchased traffic was directed to the honeypot sites between November 11th, 2012 and February 18th, 2013, resulting in over 69K delivered impressions. We used target URL's including Google Analytics campaign parameters [5] to help to differentiate overlapping purchases.

Our purchasing strategy was oriented around diversity and not volume. Details of the purchased traffic can be found in Table 2. With the exception of BuildTraffic all traffic purchased was designated as only traffic from United States and labeled as news and information. The intended delivery rate of purchased traffic varied between 333 visitors per day to 25,000 visitors per day. We intend to investigate further diversity and higher volume purchases in future work.

Table 2: Traffic purchases made for this study.

Vendor	Amount	Runtime	Price
MaxVisits	10,000	5 days	$11.99
BuildTraffic	20,000	60 days	$55.00
AeTraffic	10,000	7 days	$39.95
BuyBulkVisitor	20,000	5 days	$53.00
TrafficMasters	50,000	2 days	$70.00

3.3 Pay-Per-View Publisher Signup

In addition to traffic generation itself, PPV service providers also offer publishers the opportunity to participate as a traffic source in their network (this was our initial indication of PPV networks). To further investigate the mechanisms of traffic generation, we enrolled as a website owner willing to display content with a PPV service provider called InfinityAds. The signup was completed using InfinityAds' fully automated publisher signup system on their website. Upon signup we were given a block of JavaScript to load on our site. In return for running this tag, the website owner is assured of a relatively attractive CPM (quoted and qualified at $1.80) and that "...pop under ads will not block any of your site content and do not lead to actions where users might be led to leave your site." [23]. In this case, pop-under windows are the method that InfinityAds uses to generate traffic. We describe these in more detail below.

4. PAY-PER-VIEW NETWORK CHARACTERISTICS

In this section we report the results of our analysis of purchased traffic at our honeypot sites. This analysis reveals the mechanisms used to drive traffic to target sites and opens the door to a broader analysis of PPV networks, which is also reported below.

4.1 Traffic Generation Offerings

We reviewed the details of the 34 traffic generation/ecommerce sites that we identified via web search using strings like "website traffic", "buying web traffic", "web trafficking", etc. Features such as traffic characteristics, pricing, timing, reseller information, and DNS entries were noted for each site. Details are listed in Table 3.

4.1.1 Pricing

There is no uniform pricing for traffic providers. The pricing given in Table 3 was normalized to the cost of delivering 25,000 visitors from the United States for comparison. Of the 34 traffic generation services that we investigated, five of them did not allow purchasing traffic originating exclusively from the United States. One site was deemed fraudulent because it did not have a space to enter a traffic destination prior to checkout completion. The remaining 28 sites charged between $29.99 and $200 to purchase 25k visitors.

4.1.2 Overlap/Reselling

There were significant similarities between many of the traffic purchase sites. Many of the providers made multiple copies of their site in order to target different publisher segments or to simply use another attractive domain name. All of the provider domains were assessed using the `sameid.net` domain investigation tool [9]. Seven of the providers appeared to be repackaging another site (handytraffic, cmkmarketing, visitorboost, revisitors, buybulkvisitor, highurlstats, xrealvisitors). Four of the repackaged sites shared a Google Analytics account with another traffic provider site (handytraffic, cmkmarketing, visitorboost, revisitors). Three of the repackaged sites shared an IP address with another traffic purchase site (buybulkvisitor, highurlstats, xrealvisitors). Shared website hosting could cause IP overlap, but it is unlikely that 3 sites in our 34 site sample are randomly hosted on the same IP. Furthermore an implementation error caused *highurlstats.com* to load *buybulkvisitor.com*, making it plausible that these sites are related.

Four of the PPV sellers investigated offered the ability to become a traffic reseller (hitpro, ineedhits, toptrafficwholesaler, traffic-masters). A reseller sells traffic without having

Table 3: Traffic provider details.

Site	Price[2]	Geotargeting	Category	Pacing	Adult	Allow Pop-up/Sound
aetraffic.com	$75	Yes	Yes	Option	Option	Yes[2]
allseostar.com	NA	No	No	No	Opion[2]	No
bringvisitor.com	NA	No	No	No	?	Yes[2]
buildtraffic.com	$119	Yes	Yes	30 days	?	No
buybulkvisitor.com	$53	Yes	Yes	Option	?	No
buyhitscheap.com	$110	Yes	No	No	?	Yes
cheapadvertising.biz	NA	No	No	No	Option	?
cmkmarketing.com	$82	Yes	Yes	No	?	No
cybertrafficstore.com	$70	Yes	Yes	30 days	Option	?
easytraffic.biz	$100	Yes	Yes	60 days	?	No
fulltraffic.net	$220	Yes	No	No	?	?
getwebsitetraffic.org	$75	Yes	Yes	Option	Option	Yes[2]
growstats.com	$84	Yes	Yes	Option	?	Yes[2]
handytraffic.com	$99	Yes	Yes	Option[2]	?	Yes
highurlstats.com	$200	Yes	Yes	30 days	?	?
hitpro.us	$60	Yes	Yes	30 days	?	No
ineedhits.com	$120	Yes	Yes	30 days	?	No
masvisitas.net	No information, nowhere to enter website URL					
maxvisits.com	$30	Yes	Yes	Option	?	Yes
meantraffic.com	$30	Yes	Yes	No	Option[2]	?
perfecttraffic.com	$43	Yes	Yes	Option	?	?
plusvisites.com	$30	Yes	Yes	Option	?	?
purchasewebtraffic.net	$99	No	No	No	?	No
realtrafficsource.com	$55	Yes	Yes	No	?	?
revisitors.com	$119	Yes	Yes	Option[2]	Option[2]	Yes[2]
source4traffic.com	$88	Yes	Yes	30 days	?	No
thewebtrafficdominator.com	$32	Yes	Yes	No	Option[2]	?
toptrafficwholesaler.com	$111	Yes	Yes[2]	30 days	Option[2]	No
traffic-masters.com	$35	Yes	Yes	Option	Option[2]	No
trafficchamp.com	$89	Yes	Yes	30 days	No	No
trafficelf.com	$55	Yes	Yes	Option	Option[2]	Yes
trafixtech.com	$35	Yes	Yes	Option	Option[2]	No
visitorboost.com	$116	Yes	Yes	30 days	?	No
xrealvisitors.com	NA	No	No	No	?	?

[1] Cost to purchase 25,000 United States visitors (normalized where needed)
[2] Extra cost

to manage traffic delivery infrastructure or payment processing. The reseller acts only as an intermediary forwarding orders along to the true traffic provider. As per the descriptions, the reseller is charged a fixed rate for the traffic and can resell the traffic at the price of their choosing. Two of the reseller packages offered prepackaged websites where the reseller only needs to supply their branding and marketing.

4.1.3 Provider Site Analysis

Given the potentially fraudulent nature of traffic generation, we were interested in a general measure of the trustworthiness of providers sites. McAfee's SiteAdvisor [6] rated most of the provider websites as safe. Specifically, out of the 34 sites investigated 22 were labeled as Safe, 11 had not yet been reviewed by SiteAdvisor, and 1 was labeled as suspicious.

4.1.4 DNS Registration

A Whois lookup was performed on each of the traffic providers websites to gain insights on deployments. 14 out of the 34 sites listed a DNS anonymization service as their primary contact. Four of the sites were registered or renewed in the previous 12 months. Expiration and creation dates give the period the domain registration. On average the sites were registered for 5.71 years. The longest registration was for 16 years. Six sites are registered for only 1 year.

Looking at the contract information of the sites not us-

ing anonymization gave the following breakdown of country residency: 10 United States, 2 Australia, 2 Canada, 2 Spain, 1 France, 1 Italy, 1 Singapore, 1 China.

4.1.5 Features

Providers offer a variety of options for purchased traffic. Many provide assurances that only "real" traffic will be delivered and no "black hat techniques" are used. Every site promises unique views, such that the same user will not be directed to the site multiple times in 24 hours. Six sites were more precise, specifying that a user's IP address will only be directed to the destination once in a 24-hour period. Typical traffic volumes range between 10K and 1M visitors per campaign. Direct email was required for campaigns larger than 1M visitors. See Table 4 for other options offered by the traffic providers that we evaluated.

Table 4: Traffic provider features.

Adsense Safe	Safe to use with Google AdSense
Adult Traffic	Deliver users interested in porn
Alexa Boost	Traffic to increase Alexa ranking
Allow Pop-ups/Sound	No restrictions on destination
Campaign Pacing	Select length of campaign
Geo-targeting	Deliver users from a region
Clicks	Deliver clicks on target website
Mobile Traffic	Deliver users of mobile devices
Traffic Classes	Deliver users with specific interest

4.2 Purchased Traffic Characteristics

One of our purchases did not deliver any appreciable volume of traffic. The reason for the failure of traffic delivery is not clear. The provider may have decided not to deliver due to the instrumentation of the destination site. The provider still collected payment for the traffic which was not delivered. See Tables 5 and 6 for a summary of our measurements. Of the target of 110,000 visits that we purchased, we received 69,567. At the time of writing AeTraffic was still delivering visitors beyond the campaign end. The BuildTraffic purchase stopped delivering visitors abruptly at the end of January, 28 days into the 60-day campaign.

We analyzed traffic delivered to our honeypot websites for a variety of characteristics. Before processing, the data was filtered to remove any events originating from our honeypot server's IP address. Also any user agent containing case-insensitive 'bot' was excluded. This was done to remove the effects of web crawler traffic from our results. All of the traffic observed appeared to originate from our purchases. We did not see any indications of natural traffic.

4.2.1 Blacklist Comparison

The IP addresses of the purchased traffic showed some overlap with public IP blacklists. Every morning at 7 GMT IP blocklists were pulled from DShield.org [3] and UceProtect [10] as points of comparison. The count of blacklisted IP addresses from these sources averaged 303,968 (or 0.007% of the entire IP space) for January 2013. On average, source IP addresses of the purchased data matched the blacklists 0.97% of the time. This is perhaps more than would be expected by chance, but too low to draw a strong conclusion about overlap between the set of sources from traffic generation services and malicious sources.

4.2.2 Interaction

Each of our honeypot pages tracked four JavaScript events: onmousemove, onmousedown, onblur, onfocus. There was an extremely small number of activity events (190) reported for all purchased traffic. There are a few explanations for such low interation: *(i)* it may be an accurate reflection of reality, *(ii)* the site was 0 sized and the user could not interact with it (see 4.2.7) or *(iii)* it could be the result of JavaScript events not firing as expected. Unfortunately we cannot rule out JavaScript failure. We cannot draw strong conclusions from the lack of interaction events other than the fact that we did not pay for anything other than impressions.

4.2.3 Temporal Distribution

The pacing of visitor delivery varied greatly depending on traffic service provider. As is described below service providers traffic millions if not billions of visitors a day, but individual purchases can require delivery of less than 100 visitors a day to a destination. Furthermore, the network throughput is not guaranteed. So the deliveries need to be slightly front-loaded to ensure full delivery in the case of lower than expected throughput. The problem of pacing manifested itself in both the time of arrivals within a day and the arrival distribution over the entire campaign.

The daily arrival patterns of visitors showed some unusual artifacts. AeTraffic delivered consistently though the entire day as can be seen in Figure 4. It is well known that typical user traffic follows a diurnal cycle, reaching the high peak during the day and low peak overnight when users are sleeping. A more obvious example of artificial delivery is BuildTraffic, which delivered only during the first 10 minutes of the hour, as can be seen in Figure 5.

The arrival of users throughout the campaign was quite bursty in some cases. With periods of high delivery followed by periods of low delivery. MaxVisists delivered traffic primarily in the first half of every day as can be seen in Figure 6. Meanwhile, TrafficMasters delivery primarily consisted of two large spikes with little delivery between, as can be seen in Figure 7.

4.2.4 Incomplete Loads

Every page on our honeypot sites contained four JavaScript blocks which loaded advertising creatives. Each creative was independently instrumented to report when it had been loaded. Four blocks of JavaScript need to complete in order to successfully load all of the ads on the pages. Using this information, we can calculate the percentage of page loads

Table 5: Purchased traffic delivery.

Vendor	Expected Visitors	Delivered Visitors	Expected Duration	Actual Duration	% Loading all 4 Ads
AeTraffic	10,000	17,205	7 days	8 days[1]	16.40
BuildTraffic	20,000	1,086	60 days	29 days	60.75
BuyBulkVisitor	20,000	1	5 days	1 day	Unknown[2]
MaxVisits	10,000	9,635	5 days	5 days	12.80
TrafficMasters	50,000	41,640	2 days	3 days	58.34

[1] Still sending traffic at the time of submission
[2] User failed to load JavaScript

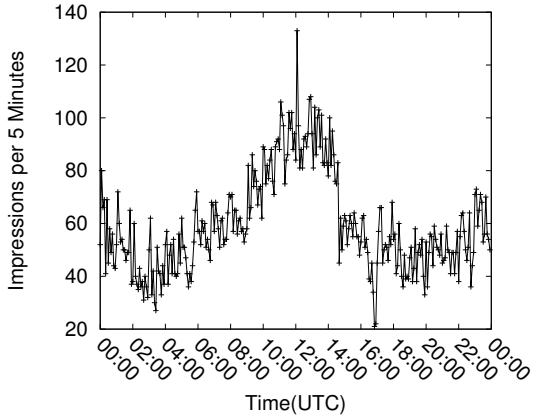

Figure 4: Traffic distribution from AeTraffic.

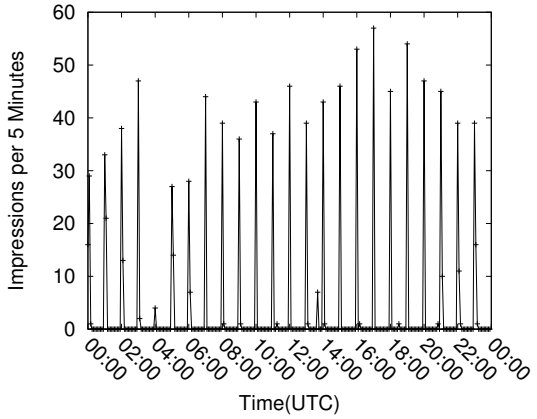

Figure 5: Traffic distribution from BuildTraffic.

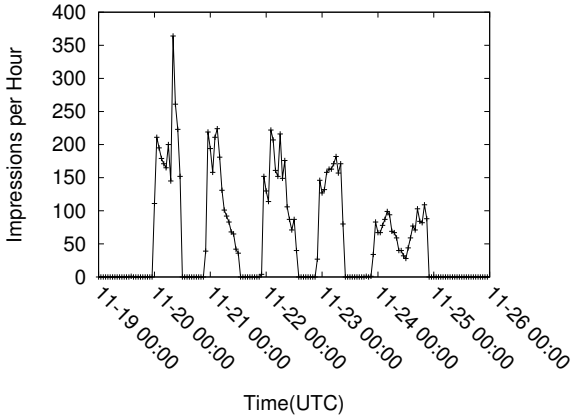

Figure 6: Traffic distribution from MaxVisits.

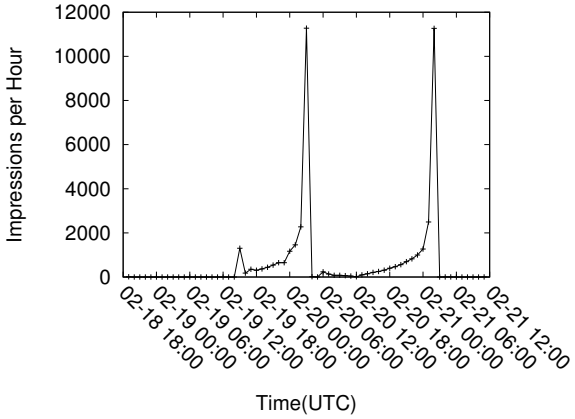

Figure 7: Traffic Distribution from TrafficMasters.

that completed for all four ads.

Traffic from BuildTraffic and TrafficMasters resulted in ads completely loading approximately 60% of the time. AeTraffic and MaxVisits only loaded approximately 15% of the time. Reasons for failure to load all the ads include: JavaScript blockers, JavaScript errors, JavaScript execution timeout, and navigation away from the page.

4.2.5 IP Address Distribution

IP addresses from an entire traffic generation campaign where checked for duplicates to get an idea of the distribution of traffic sources. According to the advertised 24-hour-unique policy an IP address could be used once per day. For small purchases our data showed very little overlap of IP address for the campaigns: AeTraffic reused 0.75% of IP addresses, BuildTraffic reused 0.64% and MaxVisits reused 11.25%. The larger purchase from TrafficMasters showed significantly more IP address overlap with 65% of IPs reused. The majority of the IPs geolocated inside of the US, with the exception of the BuildTraffic IPs.

4.2.6 User Agents

The number of unique user agents across the purchases shows the traffic came from a diverse set of browsers. An alternative explanation could be that artificial traffic generators utilized a large set of User Agent strings. However, combined with the diverse set of IP addresses, it appears the traffic could well be generated from genuine viewers.

Table 6: Purchased traffic characteristics.

Vendor	IP Sources	% US IPs	% Blacklisted	Unique User Agents	% Mobile User Agents	% Zero Sized
AeTraffic	17,075	93.14	.99	3,331	5.44	34.08
BuildTraffic	1,079	.17	1.75	312	14.42	NA[1]
BuyBulkVisitor	1	100	0	Unknown[2]	0	Unknown[2]
MaxVisits	8,551	98.83	.47	1,883	4.03	NA[1]
TrafficMasters	14,489	99.29	1.65	3,096	4.59	47.34

[1] Detailed tracking not implemented at time of purchase
[2] User failed to load JavaScript

About 5% of the traffic from AeTraffic, MaxVisists and TrafficMasters had the User Agent signature of a mobile device. BuildTraffic traffic contained a much higher percentage of mobile device User Agents. Possibly due to the increased geographic diversity of the traffic.

4.2.7 Viewport Size

Halfway through our purchases we instrumented the code to record the element height and width.[3] Overall 46.51% of ad views had a height or width of 0, meaning that the advertisement could not possibly be viewed by the user. 13.42% of views had both a height and width of zero. These results corroborate the BuildTraffic delivery technique of zero-sized frames described in 4.3.1.

4.3 Pay-Per-View Networks

By examining the JavaScript provided by traffic generation services and the referer fields from traffic on our honeypot sites, we were able to identify the fact that traffic was generated primarily from pop-under windows. Interestingly, while we did see evidence of traffic from expired domains, we saw no evidence of traffic from botnets. This observation led to our deeper investigation of the use of pop-unders for traffic generation and our characterization of PPV networks.

As noted above when publishers participate in a traffic generation service *i.e.,* a PPV network, they are given a block of JavaScript to place on their site, which looks very much like a standard ad tag. In the case of PPV networks, when a user accesses a PPV network publisher page, the JavaScript opens a new window (typically behind the active browser window, hence a pop-under) and loads the PPV server URL. The publisher running the tag gets a share of the revenue for every PPV URL that is subsequently loaded. The PPV network solves two problems with respect to marshaling users: *(i)* it delivers the JavaScript which creates the pop-under window and *(ii)* it determines the site to display in the window.

In response to prevalent pop-up advertising, web browsers give users the option to prevent pages from opening unsolicited windows. PPV networks need to circumvent this restriction. One option is the PPV code can explicitly bypass browser protections. A review of the issue trackers for Chrome or Firefox does not list many bugs related to the browsers' pop-up blockers, thus this is likely to be a difficult coding challenge. Our empirical data did not show any PPV network tags that attempt to bypass the pop-up blocker directly. The common approach is to tie pop-under creation to a user action since browsers typically allow creation of new windows on these events. Typically the pop-under action is attached to the onclick event of the body of the page. This causes the pop-under action to fire whenever the user clicks *anywhere* on the site.

After creation, the pop-under window is directed to load a specific URL pointing to the network's ad server. The ad server URL contains a number of parameters describing targeting and attribution of the visitor. The parameters always include an identifier for the originating site so that the publisher can get paid for the traffic. The list of parameters is clearly dependent on individual implementations, but some of the more common targeting parameters are: *(i)* userToken, *(ii)* indication if adult sites are allowed, *(iii)* user IP/geolocation, and *(iv)* viewport size. Using these parameters the ad server selects and returns the most profitable 3rd-party web sites (*i.e.,* the publishers that have purchased traffic) available. This is presumably the point where the 24 hour unique user guarantee is enforced.

Manually loading a publisher's PPV network tag often showed multiple redirections through a network of PPV servers. This mimics what is seen in standard advertising networks where an individual ad can be redirected across many networks in order to optimize the return from each user. For example, repeatedly loading the InfinityAds publisher tag showed network connections being made to ads.lzjl.com,
cpxcenter.com, and 199.21.148.39. Whois and reverse IP lookups on these all indicate YesUp eCommerce Solutions Inc. for the contact information. YesUp is located in Ontario Canada and has a host of eCommerce offerings.

Ideally we would have identified the referer to the main pop-under page in our purchased traffic. This would enable us to identify the sites hosting pop-under tags. Un-

[3] Using *document.documentElement.clientHeight*, *document.body.clientHeight*, *window.innerHeight* depending on browser type.

fortunately, the sandboxing of child frames (especially child frames with different domains than the parent) protects the Document Object Model of the parent frame. Therefore, the *document.referrer* node of the parent is inaccessible to the child frame. None of our traffic purchases had a value for *parent.document.referrer*. The best we can do is the referer value of the frame loading our honeypot sites. This referer points to the origin of the pop-under window code (originating from the PPV service provider).

4.3.1 Delivery Analysis

In order to gain a better understanding of how traffic is delivered to purchasing sites, we reviewed the pages listed in the referer fields for the traffic arriving at our honeypot sites. A closer examination of two of the referer sites (BuildTraffic and RealTrafficSource) showed methods for increasing the number of "page views" for every user delivered.

Loading the referer of traffic purchased from BuildTraffic resulted in a simple HTML page, including 11 frames (see Appendix for example code). The main frame loads the primary target destination in 100% of the browser viewport. Following the primary frame there are 10 frames with a height of 0 pixels. Each of these frames loads the URL of a PPV network customer. Eight of the frames load paths from a link shortening service (itsssl.com), which resolve directly to a number of sites (presumably those purchasing traffic). One of the frames loads another targeting link from the same network. The final frame loads a targeting link from yet another domain. Resulting in a total of 11 "page loads" each time the PPV network URL is loaded. Ten of those page loads are invisible to the end user because they are loaded in a frame 0 pixels high.

The page listed as the referer in traffic from RealTrafficSource also used a frame to load the final destination. In this case only a single frame covered the entire viewport, but the outer page reloaded itself every 15 seconds. When the page is displayed as a pop-under it will continue to load a different site every 15 seconds even if the pop-under window is not visible to the user.

4.4 PPV Network Throughput

Based on our evaluation of the pop-under mechanisms used by PPV networks, we endeavored to assess the broader issues of the scale of these networks (*e.g.,* number of publisher sites and number of users) and the potential volume of impressions that are being delivered on a daily basis. While all of this analysis is approximate and is based on certain assumptions, we take a conservative approach and argue that our results provide a meaningful depiction of this threat.

4.4.1 Self Reported Network Data

Many of the PPV providers list the throughput details (unique users and page views) of their network in advertising materials. Clearly, when self reporting these numbers, PPV network providers have incentive to over state in order to make their network appear larger than their competitors. Nonetheless, the self reported numbers give an insight into at least the approximate size of the networks. None of the providers publish throughput numbers broken down by features or delivery mechanisms. Thus, the numbers include pop-unders, expired domains and any other generation techniques. As shown in Table 4, 8 of the providers offer throughput information. An average of 17.16M unique visitors and 6.29B page views per provider per day are claimed. While the self report by TrafficMasters on page views is much higher than others and could be false, it may be due to an extensive affiliate network. Indeed, the use of affiliate networks means that simple summation of throughput to assess scope is unlikely to be accurate. However, the self reported numbers still point to a sizable capacity for PPV networks.

Table 7: Self reported network throughput from PPV providers.

Site	Daily Visitors	Daily Deliveries
CMK Marketing	2M	25M
HitPro	40.5M	112M
TrafficElf	20M	45M
BuildTraffic	3.3M	-
FullTraffic	20M	-
TopTrafficWholesaler	-	30M
BringVisitor	-	26.6M
TrafficMasters	-	37.5B

4.4.2 Volume Estimation

In order to estimate throughput of the networks we investigated the scope of the deployment of the PPV network tags across publisher sites. Given the publisher sites where the PPV network tags are present along with the estimated traffic for those sites we create a conservative estimate for the daily traffic across PPV networks.

The first step in determining where the PPV network tags are deployed is identifying the tag URLs. The PPV networks we considered commonly used a domain name for their delivery infrastructure that was different from the public facing websites that market to publishers. We used three techniques to identify 10 active PPV network tag URLs: *(i)* subscribing to a PPV network as a publisher, *(ii)* investigating referer fields and *(iii)* searching for ad code on public forums.

Where possible we utilized automated signup processes to harvest PPV tags directly from the publishers. This is a trivial case where the code to be run on the publisher site is directly supplied.

Using referer fields to identify PPV tags was more challenging. Typically the destination is loaded inside a frame, so the referer references the outer page hosting the frame. The display page is typically not loaded directly from the publisher site. The publisher loads JavaScript which handles

Table 8: Estimated pop-under window loads per day.

Network	Tag Count	Domains	Domain Traffic	Subdomain Traffic	Total Estimate (Views/Day)
adsrevenue.net	1,797	21	802,815	128	802,943
adversalservers.com	93,060	269	1,185,769	14,168	1,199,937
clicksor.com	855,268	2,801	24,741,249	909,649	25,650,898
edomz.com	21,750	62	971,409	11,244	982,653
ero-advertising.com	2,691,930	5,830	100,664,523	69,110	100,733,633
flagads.net	36,382	102	2,294,143	2,023	2,296,166
lzjl.com	195,406	1,192	17,427,379	425,839	17,853,218
popadscdn.net	245,302	1,029	17,016,554	124,463	17,141,017
poponclick.com	28,521	164	2,651,188	2,467	2,653,655
visit-tracker.com	90	38	623,344	0	623,344
Total	4,169,506	11,508	168,378,373	1,559,091	169,937,464

the pop-under creation and then calls the display page to fill the newly created window. In some cases, both the display page and the pop-under JavaScript are hosted on the same infrastructure. Searching the Common Crawl [2] database for the infrastructure domain lead to the identification of a number of PPV tags.

Finally, entering PPV network names into search engines resulted in a number of forum posts discussing pop-under tags. Many of the tags collected this way were no longer in use, but there were a few that were still active.

The next step is identifying the publisher sites that have deployed PPV network tags. To do this we used the Common Crawl repository of web crawl data. The August 2012 (see Table 4 for details provided by [11]) dataset included derived metadata about all of the crawled URLs. The metadata dataset contained a list of all outgoing links for each crawled page (including loading of JavaScript files). Amazon's Elastic MapReduce was used to list all paths with egress links pointing to the serving domains. The egress links were then manually reviewed to identify JavaScript files resulting in pop-under advertising. Selecting only pop-under tags from the MapReduce results gives a list of domains running those tags. We argue that this results in a conservative estimate of PPV networks that use pop-unders and an even more conservative estimate of PPV networks in general.

Estimates on traffic volumes on the identified publisher sites was done using public web analytics data. Alexa and Compete did not have traffic estimates for many of the domains. Thus, mustats.com was used to estimate domain traffic. A script was used to programmatically query mustats.com for traffic estimates on the identified PPV sites. We collected issued queries for 11,629 domains. MuStats returned an estimate for 10,737 of the queries. 2,635 of the returned queries estimated 0 views per day for the domain.

Subdomains posed an additional problem for traffic estimation. The web analytics products did not estimate traffic per subdomain. They only gave an estimate for the entire domain. For example, it is clear that just because *blogsofnote.blogspot.com* hosts a PPV network tag, not every domain on *blogspot.com* hosts that same ad tag. Attributing all of the traffic for *blogspot.com* to a PPV network would be inaccurate.

Table 9: August 2012 CommonCrawl database summary.

Crawl Date: January-June 2012
Data Size: 40.1TB (compressed)
Parsed URLs: 3,005,629,093
Domains: 40,600,000

To estimate the impact of subdomains on PPV networks, we again utilize the Common Crawl database. Our analysis counts the total number of URLs crawled for each domain that lists PPV tags. URLs with file extensions jpg, png, gif, js were removed from the total count. The final total count approximates the number of webpages and page fragments crawled for a given domain. Dividing the link count by the total crawled pages results in the percentage of pages in a domain containing links to the PPV code. This is likely a significant underestimation of reality for two reasons First, many of the URLs crawled were page fragments (where a full page is the combination of many fragements). Second, each path is given even weight despite the fact that tags are more likely to be found on high traffic pages. In any case, subdomain traffic is estimated by taking the estimated traffic for the whole domain and multiplying that by the percentage of pages inside the domain linking to the tag.

$$domains = \{123lyrics.info, serverhk.net, ...\} \quad (1)$$

$$subDomains = \{site1.blogspot.com, site2.blogspot.com, ...\} \quad (2)$$

$$estimate = \sum_{domains} domainTraffic + \sum_{subDomains} \frac{linkedPages}{totalPages} * domainTraffic \quad (3)$$

Our final algorithm for calculating PPV network throughput is then the estimated traffic for domains hosting PPV tags plus the proportional estimated traffic for subdomains containing PPV tags as shown in Equation 3. Our estimates only include the traffic expected from pop-under tags. Obviously, by including traffic from expired domains and typo squatting domains and bots would likely increase the estimated throughput substantially.

Table 8 shows throughput estimates for a selection of 10 PPV networks using our algorithm. As is expected from our conservative approach, the dominant portion of estimated traffic was to full domains with subdomain estimates making up a small portion of the total estimate. The PPV tags from ero-advertising.com, which is the largest PPV network, were displayed predominantly on publishers hosting adult content. It is possible that visitors browsing adult content are more tolerant of pop-under advertising.

So far we have estimated the number of times that pop-under code is executed per day. In reality many users have browser add-ons that prevent the creation of the pop-under window. One such popular extension for Firefox and Chrome is Adblock Plus [1]. The Firefox add-ons page for Adblock Plus lists 15.6M users [4]. Firefox claims 450M users [7], giving an install rate of 3.5% for Adblock Plus on Firefox. We conservatively estimate one quarter of all page loads prevent pop-up/pop-under creation due to plugins. Given this, we still would expect 75% of the estimated loads to result in a pop-under window. Our investigation of delivery mechanisms shows that PPV networks can load up to 11 destinations or more (in the case of auto refresh) in a single pop-under window. To maintain our conservative approach we assume four destinations loaded per pop-under window. Combining the effect of pop-up blockers and multiple loads we expect each view of a page hosting pop-under code will deliver 3 (0.75 * 4) impressions to the PPV network.

Our calculation of throughput for just 10 publisher networks resulted in more than 160M estimated tag loads per day, thus more than 500M visitor deliveries per day. Assuming a modest price of $25 per 25k visitors, the PPV providers make a minimum of $15M in sales of targeted traffic per month. Those 15B page views per month are delivered to purchasing websites. Assume the purchasing websites contain an average of 4 ads and each of those ads pays a $0.25 CPM. Advertisers spend $15M a month advertising to pop-under viewers on these 10 networks alone.

5. PAY-PER-VIEW NETWORK COUNTERMEASURES

In this section, we describe three potential counter measures to address the problem of invalid impressions generated by PPV networks. Each method offers a different perspective on the threat and each offers a different capability in terms of what can be done about the threat. While there could certainly be other viable counter measures, the following methods can be implemented by participants in the ad ecosystem who would benefit by detection and/or prevention of invalid impressions via PPV networks.

5.1 Viewport Size Filters

Advertisers who run their own ad server or intermediaries who run ad servers who are interested in removing impressions from PPV networks can filter ad requests based on viewport size. An advertiser or intermediary could implement a viewport size check countermeasure by augmenting their current JavaScript tag to include code that ensures a minimum sized viewport. This simple check code would prevent display of the advertisement for viewports which are too small to reasonably be seen by users on target platforms. In addition to reducing invalid impressions, this approach would save advertisers the bandwidth costs of delivering creatives in PPV networks.

JavaScript that detects zero-sized viewports could prevent a large amount of invalid impressions. Over 46% of the impressions in our data corpus are delivered via zero-sized viewports. Assuming this approach is used by PPV networks writ large, we estimate that a zero size viewport filter could block impressions from loading on over 200M pages per day from just the 10 PPV networks we investigated.

5.2 Referer Blacklist

Participants in the ad ecosystem could also use blacklists to identify and block traffic originating from PPV networks. We found that the referer field identifies a source in the majority of the traffic that we purchased. Over time, a blacklist of referers could be built that identifies traffic originating from a large number of PPV networks. This is similar to browser ad-blocking add-ons or in-network solutions that utilize a blacklist to remove undesired traffic. The difference with the referer blacklist is that the advertiser or intermediary implements the list directly. One limitation of this approach is that it will only work if no iframes are in use since iframes would prevent the advertiser code from accessing the referer.

Similar to viewport size filters, an advertiser/intermediary could incorporate the blacklist into their ad tags in order to prevent display to questionable viewers. As a passive alternative an advertiser could simply log the referers and compare them against the blacklist at a later time. Then the advertiser can use the information in negotiations with their advertising network.

The blacklist will need continual tuning as new PPV networks emerge and old networks disappear. One drawback of this approach is that a savvy adversary can trivially defeat this method by clearing or altering the referer field. There is some evidence that this is already happening. A few of

the referer strings in our data corpus contained direct IP addresses instead of DNS names, possibly to thwart existing or suspected blacklist methodology or simply to obfuscate their behavior. Even so a referer blacklist based on domain names would have prevented 99.51% of our purchased traffic.

5.3 Publisher Blacklists

An alternative approach is to create and maintain a blacklist of publishers that participate in PPV networks. Similar to countermeasures described above, this list could be used by advertisers to avoid running their display advertising on sites sourcing traffic from the PPV networks. This somewhat strong-armed approach would be likely to get the attention of publishers very quickly since we assume at least some percentage may not be aware of the negative aspects of their participation. Even if a publisher was aware, such an approach would discourage them from engaging with invalid traffic. Thus, this method could have potential benefits to the entire advertising ecosystem.

Publisher blacklists can be implemented by the advertiser in their tag as either preventative or informative, similar to the referer blacklist. Again this list will need continual updates as publisher behavior changes. One method of generating a publisher blacklist is to isolate and repeatedly call the PPV destination selection code block. This would enumerate all possible destinations for that PPV network over time.

6. RELATED WORK

General aspects of online advertising have been discussed in a large number of studies over the past decade. These studies have focused on wide variety of issues including the economic aspects of advertising *e.g.,* [17, 18], theoretical or analytical evaluations of sponsored search and ad auctions *e.g.,* [13, 35, 37] and more recently ad exchanges *e.g.,* [14, 30]. However, there are relatively few examples of empirical characterization studies of online advertising, most likely due to the private nature of advertising data. Relatively recent empirical studies include [19, 26, 31, 32, 39], which provide informative insights on key assumptions made in theoretical studies as well as recommendations that improve the effectiveness of online advertising.

Google, Microsoft, Yahoo and other large industry players have online documentation about their invalid traffic monitoring activities (although no significant technical details are disclosed) [21, 24, 38]. This is given to raise trust for advertisers. However, many platforms offered by intermediaries have almost no documentation on fraud. What is clear is that detecting and preventing fraud in advertising networks presents significant challenges [33, 36].

The problem of fraud in online advertising has been the subject of many different studies over the years. The majority of these studies have focused on fraud in PPC-based environments. Botnets are well known to be used for click fraud. One example of a large-scale botnet focus on click fraud was the Bamital botnet, which was recently dismantled [25]. Similarly, the ZeroAccess botnet can generate fraudulent clicks estimated to cost advertisers over $900K/day in lost revenue [12]. Other studies have focused on developing methods for detecting click-fraud *e.g.,* [28, 40]. Haddadi describes bluff ads as a means for measuring click fraud activity and creating blacklists for IP addresses to reduce click fraud [22]. Dave *et al.* [16] developed a novel measurement methodology to gather data on click fraud in ad networks. Their work informs our measurement efforts. Another recent empirical study by Zhang *et al.* is perhaps most similar to our work in terms of measurement methods [41]. In that study, the authors purchased traffic aimed at a honeypot website, and reported on a range of characteristics. Our findings on the characteristics of purchased traffic are in line with theirs, although we only purchased impression traffic and did not focus on click-through in our study.

Finally, several recent studies have included brief discussions of impression fraud. In particular, Stone-Gross *et al.* use logs from a large online ad exchange to investigate a variety of characteristics that relate to invalid activity, including behaviors related to impression spam [34]. Our work differs from prior studies principally in its focus on impression fraud. To the best of our knowledge there are no prior studies that investigate impression fraud in depth from an empirical perspective, or that investigate PPV networks and their characteristics.

7. SUMMARY AND CONCLUSIONS

Internet-based advertising is a large and growing industry. Search-based advertising still dominates in terms of annual expenditures, however display and video advertising have seen significant growth over the past several years. While publishers have always been motivated to use diverse methods to drive users to their sites, the fact that payments for display and video ads are often based on impressions motivates new offerings from 3rd-party traffic generation services.

In this paper, we investigate the problem of invalid traffic generation that is aimed at inflating impressions on publisher websites and apps. We address this problem empirically by setting up several honeypot websites that were used as the targets for traffic generation purchases, which we made over the course of several months. This traffic provides the baseline from which we were able to identify a particular form of impression generation that we call pay-per-view networks. A PPV network is a series of legitimate publisher sites that include a common embedded reference from a particular traffic generation service. When users access publisher sites that participate in PPV networks, 3rd-party websites are rendered in an obfuscated and often invisible fashion. By evaluating the JavaScript associated with PPV networks, we find that the predominate mechanism used is pop-under windows. We also find that PPV networks place multiple 3rd-party pages on pop-unders using frames or use periodic refresh to leverage every user access. This approach

preserves the user experience on the publisher's site and generates invalid impressions on the 3rd-party sites in a way that is difficult to detect.

Next, we investigate aspects of the broader scope of PPV networks by gathering information from a small selection of ten traffic generation services. We search for tags from these services in a publicly available Internet-wide crawl database to estimate deployments on publisher sites. We couple these estimates with estimates for daily unique page views from those sites and find tag throughput above 150M per day. Combined with conservative estimates of 3rd-party displays per tag and ad placements per page, this easily pushes the number of invalid impressions above 500M per day from these ten PPV networks alone. Based on the fact that our sampling is so small, the impact of PPV networks is likely to be much larger.

To address the threat of PPV networks, we describe three different counter measures. Each offers a different constituency an opportunity to block the display of the unwanted 3rd-party content. In future work, we plan to focus on developing implementations of the proposed counter measures as well as developing other techniques to address this threat. Our measurement and characterization work are ongoing and will soon focus on traffic generation services outside of North America.

Acknowledgements

The authors would like to thank our shepherd, Chris Grier, for his input and multiple reviews of the paper. This work was supported in part by NSF grants CNS-0831427, CNS-0905186, ARL/ARO grant W911NF1110227 and the DHS PREDICT Project. Any opinions, findings, conclusions or other recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, ARO or DHS.

8. REFERENCES

[1] Adblock Plus. http://adblockplus.org, 2013.

[2] Common Crawl. http://commoncrawl.org/, 2013.

[3] Dshield Daily Sources. http://www.dshield.org/feeds/daily_sources, 2013.

[4] Firefox Add-Ons - Adblock Plus. https://addons.mozilla.org/en-US/firefox/addon/adblock-plus/, February 2013.

[5] Google Analytics URL Builder. http://support.google.com/analytics/bin/answer.py?hl=en&answer=1033867&topic=1032998&ctx=topic, 2013.

[6] McAfee SiteAdvisor. http://www.siteadvisor.com/, 2013.

[7] Mozilla Press Center. http://blog.mozilla.org/press/ataglance/, February 2013.

[8] Network Solutions Whois. http://www.networksolutions.com/whois, 2013.

[9] SameID.net. http://sameid.net/, 2013.

[10] UCEPROTECT Blacklist. http://rsync-mirrors.uceprotect.net/rbldnsd-all/ips.backscatterer.org.gz, 2013.

[11] Web Data Commons - Extraction Results from the August 2012 Common Crawl Corpus. http://webdatacommons.org/#toc2, 2013.

[12] ZeroAccess is Top Bot in Home Networks. http://www.infosecurity-magazine.com, February 2013.

[13] G. Aggarwal, J. Feldman, S. Muthukrishnan, and M. Pai. Sponsored Search Auctions with Markovian Users. *Internt and Network Economics, Lecture Notes in Computer Science*, 5385, 2008.

[14] S. Balseiro, J. Feldman, v. Mirrokni, and S. Muthukrishnan. Yield Optimization of Display Advertising with Ad Exchange. In *Proceedings of the ACM Electronic Commerce '11*, San Jose, CA, June 2011.

[15] Internet Advertising Board. IAB Internet Advertising Revenue Report 2012 First Six Months Results. http://www.iab.net, October 2012.

[16] V. Dave, S. Guha, and Y. Zhang. Measuring and Fingerprinting Click-Spam in Ad Networks. In *Proceedings of ACM SIGCOMM '12*, Helsinki, Finland, August 2012.

[17] B. Edelman, M. Ostrovsky, and M. Schwarz. Internet Advertising and the Generalized Second Price Auctions: Selling Billions of Dollars Worth of Keywords. American Economic Review, 2007.

[18] D. Evans. The Economics of the Online Advertising Industry. *Review of Network Economics*, 7(3), 2008.

[19] A. Ghose and S. Yang. An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets. *Management Science*, 55(10), July 2009.

[20] Google. Adexchange. http://www.google.com/adwords/watchthisspace/solutions/ad-exchange/, 2013.

[21] Inc Google. Ad Traffic Quality Resource Center. http://www.google.com/ads/adtrafficquality/index.html, 2013.

[22] H. Haddadi. Fighting Online Click-fraud Using Bluff Ads. *ACM SIGCOMM Computer Communications Review*, 40(2), 2010.

[23] InfinityAds. Publisher Signup. http://www.infinityads.com, 2013.

[24] T. Kelleher. How Microsoft Advertising Helps Protect Advertisers from Invalid Traffic. http://community.bingads.microsoft.com, November 2011.

[25] J. Kirk. Microsoft, Symantec Take Down Bamital

Click-fraud Botnet. http://www.infoworld.com, February 2013.

[26] S. Lahaie and P. McAfee. Efficient Ranking in Sponsored Search. In *Proceedings of the Seventh Workshop on Ad Auctions*, San Jose, CA, July 2011.

[27] I. Lunden. Forrester: US Online Display Ad Spend $12.7B In 2012, Rich Media and Video Leading The Charge. http://www.techcrunch.com, October 2012.

[28] A. Metwally, D. Agrawal, and A. El Abbadi. Using Association Rules for Fraud Detection in Web Advertising Networks. In *Proceedings of the International Conference on Very Large Databases*, Trondheim, Norway, August 2005.

[29] MuStat. MuStat. http://www.mustat.com, 2013.

[30] S. Muthukrishnan. AdX: A Model for Ad Exchanges. *ACM SIGEcon Exchanges*, 8(2), 2009.

[31] M. Ostrovsky and M. Schwarz. Reserve Prices in Internet Advertising Auctions: A Field Experiment. In *Proceedings of the Sixth Workshop on Ad Auctions*, Cambridge, MA, July 2010.

[32] N. Vallina Rodriguez, J. Shah, A. Finamore, Y. Grunenberger, K. Papaginnaki, H. Haddadi, and J. Crowcroft. Breaking for commercials: characterizing mobile advertising. In *Proceedings of ACM Internet Measurement Conference (IMC '12)*, Boston, MA, November 2012.

[33] B. Schwartz. Google: Investigating Invalid AdSense Traffic is Extremely Difficult. http://www.seroundtable.com, April 2012.

[34] B. Stone-Gross, R. Stevens, R. Kemmerer, C. Kruegel, G. Vigna, and A. Zarras. Understanding Fraudulent Activities in Online Ad Exchanges. In *Proceedings of ACM Internet Measurement Conference (IMC '11)*, Berlin, Germany, November 2011.

[35] C. Tucker and A. Goldfarb. Search Engine Advertising: Pricing ads to context. In *Proceedings of the Fourth Workshop on Ad Auctions*, Chicago, IL, July 2008.

[36] A. Tuzhilin. The Lane's Gifts v. Google Report. http://googleblog.blogspot.com/pdf/Tuzhilin_Report.pdf, 2006.

[37] H. Varian. Position Auctions. *International Journal of Industrial Organization*, 25, 2007.

[38] Yahoo. Traffic Quality: We Work to Protect You in a Variety of Ways. http://advertisingcentral.yahoo.com, 2013.

[39] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How Much can Behavioral Targeting Help Online Advertising? In *Proceedings of WWW '09*, Madrid, Spaing, April 2009.

[40] L. Zhang and Y. Guan. Detecting Click Fraud in Pay Per Click Streams of Online Advertising Networks. In *Proceedings of the International Conference on Distributed Computing Systems*, Beijing, China, June 2008.

[41] Q. Zhang, T. Ristenpart, S. Savage, and G. Voelker. Got Traffic? An Evaluation of Click Traffic Providers. In *Proceedings of WebQuality '11*, Hyderabad, India, March 2011.

APPENDIX

Traffic Delivery Code

```
...
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='header' src='about:blank' scrolling='no' noresize>
<frame name='main' src='' +rurl+ '' scrolling='auto'>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='raaj1' src='http://itsssl.com/37kt' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='house2' src='http://stats.itsssl.com/?VFJDSz0zNA==' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='house3' src='http://stats.itsssl.com/?VFJDSz0zNA==' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='usnopop' src='http://stats.itsssl.com/?VFJDSz00' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='usnopop2 src='http://stats.itsssl.com/?VFJDSz00' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='usnopop3 src='http://stats.itsssl.com/?VFJDSz00' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='geo1' src='http://www.itsssl.com/georedirect/main.html' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='geo2' src='http://www.itsssl.com/georedirect/main.html' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='raaj2' src='http://stats.buildtraffic.com/?VFJDSz01OA==' scrolling='no' noresize>
<frameset rows='0,*' framespacing='0' border='0' frameborder='0'>
<frame name='georedirect' src='http://adzay.com/redirect.php' scrolling='no' noresize>
...
```