Got Traffic? An Evaluation of Click Traffic Providers

Qing Zhang, Thomas Ristenpart†, Stefan Savage and Geoffrey M. Voelker
Department of Computer Science and Engineering
University of California, San Diego
†Department of Computer Sciences
University of Wisconsin-Madison

ABSTRACT

Internet advertising has been a highly profitable means by which companies and organizations can pay to attract visitors to their Web sites. Over time, satisfying the demand for this service has evolved into a market of "click traffic" providers that use various models to direct visitors to customer sites. Well-known premium providers like Google AdWords use pay-per-click auctions, for instance, while a variety of bargain providers offer click traffic in bulk. In this paper, we evaluate the quality of purchased click traffic from a range of such traffic providers. Using multiple instances of a custom Web site, we purchase click traffic to our sites from nine providers. In each case, we characterize click traffic directed to the sites using a variety of metrics, including timing properties, access patterns on the site, network properties of the hosts accessing the site, correlation with blacklists, etc. We find that providers differ substantially, and that these characteristics correlate with click quality: the traffic you get is the traffic you pay for.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: On-line Information Service—*Commercial services, Web-based services, Online advertising*; K.4.1 [**Computing Milieux**]: Public Policy Issues—*Abuse and Crime involving computers, Click fraud*

General Terms

Click fraud, Online Advertising, Traffic measurement, Pay-per-click

1. INTRODUCTION

Much of today's free on-line Web is underwritten by an advertising business model that explicitly monetizes the traffic of Web site visitors. At its core this model supposes that a user's decision to click (e.g., on a sponsored search term, a banner ad, etc.) represents interest in the associated content, and thus, such a click can be sold as a low-level form of "sales lead". While much of this multi-billion dollar market is dominated by large advertising networks (e.g., sponsored search from Google, Yahoo and Bing, display advertising from Facebook, etc.) it is less-well appreciated that a secondary traffic-selling ecosystem — comprising traffic vendors who will contract to deliver clicks to a site in exchange for payment — has been engendered as well.

Indeed, in our investigations we have identified a broad range of secondary traffic vendors with various pricing mechanisms, asserted traffic sources, and targeting interfaces. For example, many services charge purchasers a flat fee for a promised amount of traffic or traffic rate ("bulk vendors") as contrasted with the per-click pricing of traditional advertising networks. Similarly, while some bulk vendors represent that their traffic originates from an ad network, typically they are quite ambiguous and evidence suggests that they use a variety of other means (e.g., recently acquired expired domain names). More commonly, vendors are simply ambiguous about the source of their traffic; the truth of which is further obscured by affiliate agreements that allow one provider to resell or repackage traffic streams for sale to further sellers or customers.[1] Finally, while established ad networks typically use a keyword-based traffic targeting model, many bulk vendors sell traffic that is simply targeted by market "category" or, completely undifferentiated traffic. While none of these practices are inherently problematic, they certainly give rise to questions about traffic quality.

Roughly speaking, the quality of a click corresponds to the potential for providing value to the site that has paid for it (typically in the form of sales conversions). Operationally, quality typically refers to the legitimacy of the source — the gold standard being real users with honest interest being directed to a site. However, not all traffic sources are legitimate and fraud can occur when traffic originates from malware bots ("click bots"), illegally installed adware, or illegitimate redirects of a real user during a browsing session (e.g., via pop-unders or "black-hat" search engine optimization schemes). Unfortunately, traffic does not identify its causal origin and thus traffic purchasers are not always aware of the quality associated with the traffic they have purchased.

In this paper we examine the relationship between traffic provider and traffic quality from the customer viewpoint. Specifically, we focus on understanding the extent to which the quality of purchased traffic differs between traditional ad networks and bulk traffic vendors. While there is considerable "received wisdom" about this question, we are unaware of any work explicitly evaluating it. To this end, we have taken an empirical approach in our study, measuring the results of traffic purchased from six different vendors over a one-month period. By evaluating a range of traffic features, both explicit (e.g., User-Agent string and referrer) and implicit (e.g., user mouse movement) we establish that there are highly distinct profiles in the traffic provided by different classes of vendors.

We see strong evidence that the bulk traffic vendors we measure are not directing organic Web traffic: their traffic has no mouse activity, no subpage visits, no link accesses, different browser distributions and so on. In one case, visiting the site identified in the referrer field revealed that the traffic originated from a "paid-to-read" site. These evidence show strong correlation between traffic quality and the price paid.

2. BACKGROUND AND RELATED WORK

Web traffic originates from various sources on the Internet today, enabling traffic vendors to make various claims regarding the

[1] On underground forums that deal with the sale of traffic it is common for purveyors to "slice and dice" their traffic, reselling pieces for maximum value. For example, redirected search traffic with high-value keywords originating from high-value countries might be resold at a premium via a quasi-legitimate traffic reseller, while untargeted "garbage" traffic might be resold into a bulk market.

WebQuality '11, March 28, 2011 Hyderabad, India.

origins and quality of traffic they provide. This section provides an overview of popular pricing models associated with Web traffic as well as a summary of prominent traffic sources observed and purported by traffic vendors.

2.1 Payment models

The pay-per-click (PPC) traffic model is widely adopted by traffic vendors such as major search engines. In the PPC model, a Web site agrees to pay the referrer site for each user who clicks through to the targeted site [1]. Search engines typically do this via hosting an auction on different keywords.

With a keyword-auction based pricing model such as the PPC, the traffic buyer does not know in advance the cost per click. Although traffic customers are given a price estimate at the time of purchase, they ultimately control the amount spent by specifying an allowance in a given time period, typically a day. When the allowance is exhausted, the traffic vendor removes the customer from the auction until the end of the time period.

An alternate payment method for Web traffic is buying traffic in bulk. Bulk traffic vendor offers a certain quantity of traffic for a set amount of money. The traffic is typically guaranteed to arrive in a given period, commonly a month. The traffic buyer pays the fees upfront. At the time of our experiments, for example, Revisitors.com offered 5,000 "clicks" over a month for US$28.95.

Other models of billing and accounting exist. However, due to limited space, we will focus on Web traffic accounted for by click-through rates, because this is the most prevalent method used.

2.2 Traffic Sources

Online advertising is dominated by large advertising networks [13], such as Google AdWords, which bring together publishers who display advertisements on behalf of advertisers seeking to direct traffic to their sites.

Search Engines. Search engines are the dominant method for directing Web traffic to sites [10, 14]. A search engine directs Web traffic through links returned from search results. Since search engines determine the order the search results are returned, they in turn directly affect the quantity of traffic.

Rather than relying upon the ranking algorithms used by search engines to direct users to them via search results, sites pay to have their ads displayed when users search for keywords, typically using the pay-for-click model. Search engines group sponsored ads around search results in blocks that are displayed before, after, or alongside the results, sometimes with a contrasting background to distinguish them. Since the major search engine companies also provide other free services, such as Web email, these companies also naturally integrate advertising into these services as well.

Pop-up, Pop-under and Banner Ads. Another method for diverting traffic to Web pages is through pop-up and pop-under ads. automatically load in a new window when a Web page loads: a pop-up ad opens a window over the existing page, whereas a pop-under ad opens the window behind the current window. In the pop-up scenario, a user browsing the Web is forced to look at the ad, minimize it or close it before they can continue to the page they intended to visit. Due to negative reaction from Web users to the disruptive nature of pop-up ads, pop-under ads were developed. Pop-under ads are designed to be less obtrusive than pop-ups, but they are still considered a nuisance by most [12]. Today most Web browsers block pop-up and pop-under ads by default due to their disruptive nature. However, advertising companies continue to develop new ways around pop-up blockers using JavaScript and Flash.

Another widely used technique, Banner ads, are considered more benign. Banner ads are embedded in a Web page usually in the form of an image or a multimedia object. When clicked, banner ads direct users to the targeted page. Much controversy surrounds the effectiveness of pop-up, pop-under and banner ads. Although users find pop-up ads irritable, yet click-through rates for pop-up ads were almost twice as high as banner ads [12] and, as a result, remain an enticing mechanism for advertisers. Other studies suggest that the relevance of the ad plays a great deal on the click through rate as well as the rate for which people find the ad annoying [11].

Expired Domains. Expired domains are valuable because existing links to them on Web pages and in search results will direct traffic to whomever obtains the domain next. Legitimate Web developers sometimes acquire these domains to jumpstart traffic to a site. Advertisers can use the domain to forward user traffic on to customer sites directly, or to create link aggregation pages with site advertisements. Expired domains are also exploited by spammers, who use them to direct traffic to scam sites [9].

Click Fraud. Online advertising via pay-per-click (PPC) is based on the assumption that users clicking on an advertisement have some interest in the site being advertised, which is ultimately why advertisers are willing to pay for such activity. Due to the large variety of traffic sources on the Web, though, it is often hard to determine the true nature of Web traffic visiting a site. This situation opens the door for malicious users to commit click fraud.

Click fraud takes place when clicks to online advertisements are generated with the goal of triggering payment for the click, rather than having any interest in the advertised site. Generally, the motivations for click fraud are profit-driven. Fraudulent clicks inflate revenue for sites publishing ads, and for advertising networks. As a result, given the large amount of money being invested in online advertising, click fraud has been a controversial issue. Companies such as Google that serve multiple roles—they publish ads and charge advertisers—can arguably profit from click fraud, and have been accused (and sued) for not sufficiently preventing it. Not surprisingly, there has been much interest in detecting click fraud [8], both by the advertising networks to maintain their reputations [15] as well as by third-parties who offer click-through auditing services (which then raises the question of who audits the auditors [4]). Other motivations for click fraud are more malicious; one company performing clicks on the advertising links of a competitor, artificially inflating their advertising bill.

Click fraud can manifest in a number of ways, including being automated, with techniques ranging from simple scripts on a single machine clicking on ads, to malware that generates clicks while piggy-backing on unsuspecting users [7], to large-scale botnets that generate traffic from many disperse hosts [5]. The bulk traffic vendors we purchased traffic from in this study exhibit many characteristics of automated traffic (Section 4). Individuals can also create their own sites, participate in an advertising network, and click on ads on their site to receive payment from the ad network [3]. To circumvent click fraud detection methods that identify automated click traffic, services have also emerged which employ cheap human labor to click on advertisements. Indeed, one of the vendors we used, Rent-a-list, appears to use this approach to generate traffic to their customers (Section 4.5).

3. METHODOLOGY

This section describes the Web site we created to advertise in our study, the kinds of traffic that we purchase to it, and the traffic vendor we purchase from as customers.

3.1 PeachySkin

We created a custom Web site called PeachySkin for purchasing traffic to it. PeachySkin is a cosmetic consulting site with various pages and links to other cosmetic related Web sites. Figure 1 shows a browser screenshot of the top-level page of PeachySkin, and lists

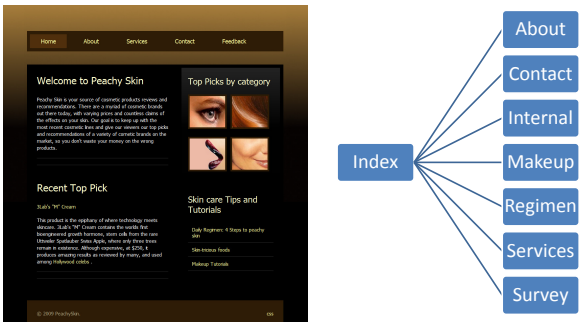


Figure 1: PeachySkin browser screenshot and the topics of second-level pages.

its second-level topical pages. In addition to the HTML and image content, the pages also include tracking scripts, such as JavaScript for tracking mouse activity, script for snapshotting referrer's page, scripts for storing all the web requests, JavaScript for Google Analytics as an additional method for tracking activity on the site, and instrumentation for recording clicks on all links to track user navigation of the site.

We host multiple replicated instances of PeachySkin in separate virtual machines in Amazon's EC2 service running the Apache Web server. We use EC2 for its ease of deployment and anonymity. We use separate replicas for assigning a unique instance of the PeachySkin site for each traffic vendor from which we purchase traffic, thus isolating traffic from each vendor.

We preprocess data to remove traffic that is not associated with a traffic vendor, such as Web crawlers. We filter crawler traffic by matching against crawler signatures in the User-Agent HTTP header (e.g., "Googlebot"). We generated this crawler list by manually examining User-Agent strings that did not match common browsers, and then verifying that they are known crawlers. The amount of crawler traffic varies across the different PeachySkin instances. While some have as little as dozens, others have hundreds of requests which would otherwise skew analyses. We also filter requests that we ourselves generate when testing the site by removing requests from IP addresses in our organization.

3.2 Traffic

We purchase traffic directed to the top-level PeachySkin page. Traffic vendors typically offer one of two types of traffic: traffic based on keywords, or traffic based on categories. The pricing model between keyword-based and category-based is very different. Keyword-based vendors typically employ a keyword auction where traffic buyers bid on certain keywords and set an allowance. Most search engines employ a pay-per-click payment model The price of each click in the keyword-based scheme is not known ahead of time. For the keyword-based vendors used in this study, we bid for keywords targeting "cosmetics" and "peachyskin".

Others provide a simpler category-based model. A customer selects a category, such as "cosmetic" or "health", from a given list of categories offered by the traffic vendor. Category-based traffic avoids the need for traffic buyers to do any bidding by associating a fixed cost with a category for a predetermined amount of traffic. Thus the price per click is known ahead of time. For the bulk traffic vendors used in this study, we purchased traffic targeting the category "cosmetics". From our experimental results in Section 4, we find the quality of category-based traffic vendors questionable.

Geographic regions.

A common option in buying Web traffic allows for the specification of a particular geographic location. For example, companies that want to sell merchandise in the US can choose to buy traffic only originating from the US. We purchased Web traffic from an additional geographic region—the United Kingdom—from the same vendor to validate this feature of their service. For traffic purportedly from a specific region, we geolocate the IP addresses of the HTTP requests using NetAcuity's IP lookup service [6].

Untargeted traffic.

A traffic customer buying category-based Web traffic can also choose to buy traffic that is untargeted. Untargeted traffic is not associated with any keyword or category. This type of traffic is the cheapest form of Web traffic. In an initial experiment, we found little difference purchasing untargeted traffic compared with targeted traffic from the same bulk traffic vendors. As a result, we did not experiment with untargeted traffic thereafter.

3.3 Traffic vendors

We purchased traffic to PeachySkin replicas initially from nine different vendors. From reviewing many vendors offering services on the Web, we categorized them into three tiers—low, middle, and high—ultimately based on the price per volume of traffic: the more expensive the traffic, the higher the tier. We then selected three popular vendors from each tier as a representative sample. In addition, we purchased traffic from the same geographic region from at least one vendor in each tier. We also logged traffic on two baseline sites for PeachySkin: one with the catchy domain name `peachyskin.com` and another "hidden" site that was unregistered and otherwise idle.

Although they accepted our payment, we did not receive traffic from two middle tier vendors (Bid4Keywords and Wpromote). We also advertised through Bing, but discovered after we started our experiment a technical issue in getting traffic from Bing.

Table 1 lists the traffic vendors we used, as well as their traffic guarantees, prices, and the amount of traffic we purchased from them. The lowest tier consists of bulk traffic sellers. These are traffic vendors that offer thousands of visitors for a set price. Their Web sites often contain testimonials and guarantees for delivering "quality" visitors to the target site provided by the traffic customer. From the lowest tier we purchased traffic from Revisitors, HandyTraffic, and Aetraffic; we placed them in this category because they sell bulk traffic for a low fixed cost.

Bulk traffic vendors make various claims such as redirecting visitors to their customer sites through their specialized Web sites. Revisitors have banners all over their website illustrating their commitment to bring "quality" Web traffic.

The middle tier consists of traffic vendors that sell traffic that is slightly more expensive or, from our own experiences interacting with the services, have better customer support such as live chat available for customer service. From the middle tier we selected Rent-a-list, Wpromote, and Bid4Keywords, but only received traffic from Rent-a-list. Rent-a-list uses a keyword-based auction, but is not nearly as well recognized as the high tier vendors.

Finally, the highest tier consists of well-known search engines based in the US that operate under a PPC model. We use Google AdWords because they are a market leader in online advertising, they are the most expensive, and they serve as a viable reference point for comparing the less well-known traffic vendors. In addition to Google AdWords, we also purchased traffic from Yahoo ads.

Additionally, we created two baseline sites that have no traffic explicitly directed to them. One has the eponymous domain name of `peachyskin.com`. It represents a baseline for a site that has a registered domain name, appears in search engine results, and receives traffic from both Web crawlers as well as an occasional

Traffic Advertised				Traffic Observed				
Vendor	Traffic claims (u=unique)	Cost	Clicks	Visits	Visits/day (avg)	Clicks 24-uniq	Distinct Visits	Requests
Revisitors	160/day 24hr u	$\$28.95/mo$	5000	3516	117	3395	3386	3535
Revisitors(UK)	160/day 24hr u	$\$28.95/mo$	5000	3781	126	3274	3260	3807
HandyTraffic	24 hr u	$\$24.95/mo$	5000	4021	134	3431	3414	4054
HandyTraffic(UK)	24 hr u	$\$24.95/mo$	5000	4016	133	71	56	4024
Aetraffic	$visitor$ u	$\$69.65/mo$	10000	13045	43	4019	3986	13066
Rent-a-list(UK)	$1/2$	$\$27.50/mo$	5000	43	1	34	34	68
Google	None	$\$10/click(max)$ $299.27	67	113	–	100	99	206
Google(UK)	None	$\$10/click(max)$ $283.17	104	143	–	126	124	239
Yahoo	None	$\$10/click(max)$ $408.28	132	205	–	36	32	95
peachyskin.com	None	–	–	34	–	30	26	46

Table 1: Summary of traffic vendors, the prices for which we paid for traffic, and the amount of traffic delivered.

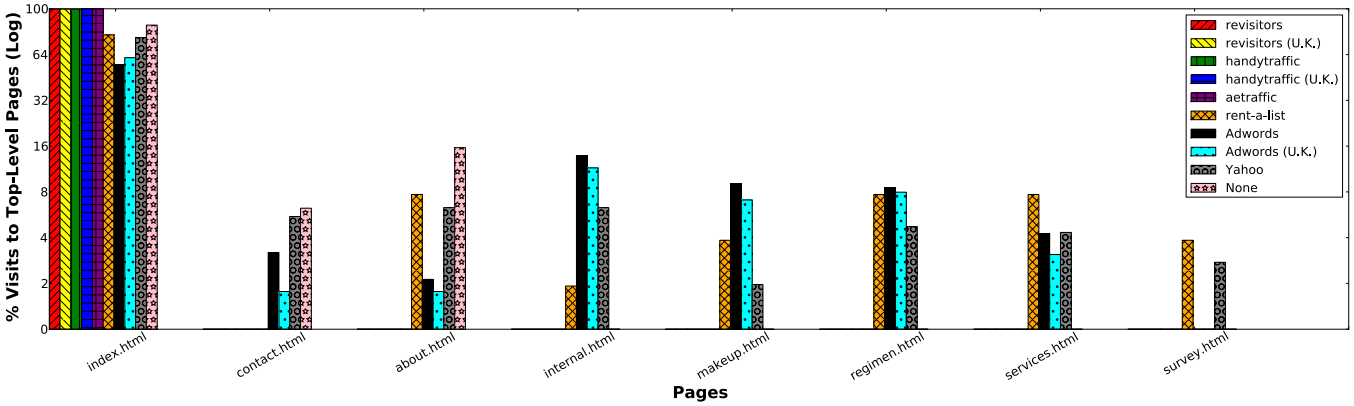


Figure 2: Traffic to PeachySkin pages, normalized by the total number of all clicks to a site. Note that the y-axis is log-scale.

visitor perhaps attracted by the domain name. With crawler traffic removed, this site represents a baseline of presumably organic traffic to PeachySkin. The other site has no proper domain name other than what was assigned by EC2.[2] This baseline represents a "hidden" site, and is primarily visited by Web crawlers.

4. CHARACTERIZING TRAFFIC VENDORS

In this section we characterize traffic purchased. We characterize traffic by evaluating traffic volume, mouse activity, link accesses, User-Agent, referrers, timing metrics, and finally blacklists. Our high-level goal is to arrive at a conclusion about the "quality" of the traffic from the vendor. We examine characteristics that might differentiate between so-called "organic" traffic from real users and "inorganic" automatically generated traffic. In general, we use characteristics from the established pay-per-click vendors, Google and Yahoo, as a baseline for comparison and assume that they represent useful, organic traffic. Of course, even organic traffic can be fraudulent; one of the vendors we used employs "paid-to-read" sites to generate organic but useless traffic to their customers.

4.1 Traffic Volume

When purchasing traffic, the first natural question is whether customers received the amount traffic they pay for. We compare the amount of traffic that we observed from each traffic provider with the amount of traffic that was expected by that provider in Table 1. Note that the traffic statistics in the table have all crawler and whitelisted traffic removed in a preprocessing step (Section 3.1).

[2] *ec2-79-125-60-211.eu-west-1.compute.amazonaws.com*

Since a "click" can have a number of definitions, we refer to traffic using different terms depending on how the traffic is counted. We define a "visit" as an HTTP request to the top-level page of a site. We define a "click" as an individual visit to a site according to the accounting definition of a traffic vendor. This definition depends on the type of vendor. For the pay-per-click vendors, "visits" and "clicks" are equivalent. Bulk traffic providers, on the other hand, sell traffic in terms of the number of *unique* users within a 24-hour period ("Clicks 24-uniq" in Table 1). According to their terms, then, customers are only "charged" one click for multiple visits from the same user on the same day. In addition, to get an overall sense of IP address dispersion, we also count the number of visits from unique IP addresses for the entire month as "distinct visits". Finally, we also count the number of HTTP requests to any HTML container page on the site, and refer to them as "requests".

Our two baseline sites, `peachyskin.com` and a "hidden" site on EC2, do receive some traffic. We do not purchase any traffic to either site, nor are there any external links that point to them that we are aware of. Even so, the `peachyskin.com` site receives a small amount of traffic each day on average, perhaps because of its domain name. The hidden site has no domain name associated with it. We expect, therefore, that it would receive little to no traffic. We did observe small amounts of HTTP traffic to the hidden baseline, and all of it was crawler traffic from bots.

For the bulk providers, we both purchased and received a large volume of click traffic from them. Even so, the bulk providers underdelivered. Revisitors delivered 65–68% of the amount purchased, HandyTraffic 69%, and Aetraffic 40%. Although HandyTraffic from the UK delivered a commensurate amount of traffic as

	% of Visits			
Vendor	U.S.	U.K.	Netherlands	—
Adwords	14.2	79.0	1.35	5.45
HandyTraffic	0	98.7	0	1.3
Revisitors	26.0	72.1	0	1.9

Table 2: Geographic regions of visits purchased to originate from the U.K.; the column "—" includes traffic from regions accounting for 1% or less of the traffic, or traffic the geolocation tool was unable to geolocate.

from the US, it did so using very few hosts: the amount of clicks (visits from 24-hour unique hosts) was only 71 even though we technically paid for 5,000. Rent-a-list has a 50% guarantee on the target, but does not reach even 1% of our target (Section 4.5 has more context for Rent-a-list). In subsequent analyses, we find other characteristics of the bulk provider traffic to be further suspicious.

The pay-per-click providers delivered the amount of traffic expected given our designated daily budget. Calculating the average cost per click on each day and the remaining funds that day, we found that the remaining funds were less than the average cost for one click. Given the price premium from the auctions for the PPC traffic, not surprisingly we received substantially less traffic than the bulk providers.

4.1.1 Subpage Traffic

One of the goals of attracting visitors to the main page of a site is to interest them into exploring other parts of the site. Figure 2 shows the distribution of clicks across the pages comprising the PeachySkin site (Figure 1) for each provider. Not surprisingly, the vast majority of clicks go to the main page `index.html`. Admittedly, the utility of our site is limited, so perhaps not many visitors will explore further. However, across all of the visitors to the site, we would expect at least some of the visitors to explore PeachySkin beyond the main page. Indeed, for the baseline version of the site `peachyskin.com`, we see that roughly 15% of clicks are to other pages on the site. We see similar behavior for traffic provided by the well-known PPC sites (Google, Bing, Yahoo) where 30–50% of clicks are to second-level pages, noticeably above the baseline. Rent-a-list appears encouraging, with 20% of the traffic to subpages. The remaining bulk vendors, though, had negligible or no visits to subpages.

4.1.2 Geographic Region

As discussed in Section 3.2, customers can purchase traffic to their sites from specific geographic regions. For three of our providers, we also purchased traffic from the United Kingdom to validate whether such traffic indeed appears to originate from that country. For these HTTP requests, we mapped the IP addresses of the client hosts to their country of origin (Section 3.2). For the cases where we requested traffic from the UK, Table 2 shows the distribution of countries originating traffic as determined by the IP geolocation tool. Notably, nearly all of the traffic from the bulk vendor HandyTraffic originates from the UK, although from just a small number of hosts. Revisitors, though, fares worse at just over 70%, failing to deliver on this feature for nearly a third of its traffic.

Interestingly, over 20% of the traffic from Adwords originates outside the UK (primarily the US). Recall that Adwords does not charge for every click-through, only those that meet its heuristics [8]. From our data collected from Adwords, the difference between the number of visits and the number of click-throughs charged to us varies between 18-33% of the total visits, which could easily account for the US-originated traffic.

4.2 Mouse activity

We expect organic traffic from real users who visit a site to exhibit some kind of mouse activity.[3] Even if a user quickly decides that they are not interested in the site, there generally should be some mouse movement involved to navigate away. Of course, automated traffic can be scripted using browsers to emulate mouse activity, but doing so adds additional complexity to the automation.

To record mouse activity from the visits to our sites, we used JavaScript in the top-level page to record mouse movement. (A not uncommon technique for detailed analytics on major sites.) To reduce logging overhead, we only record mouse moves beyond a small threshold. Experimenting with various thresholds, we found a movement delta of 20 pixels to be a good tradeoff. It substantially reduces logging overhead, yet is sensitive enough to record even short moves and will capture events such as navigating away from a page via clicking on the back button.

Visually, Figure 3 overlays user mouse movement on the main PeachySkin page for traffic from three representative vendors. For each mouse move we recorded, we draw a line segment on a screenshot of the page. Although Aetraffic had thousands of clicks to the page, we recorded only four mouse moves (circled at the top of the page) — even less mouse activity than recorded on the baseline site `peachyskin.com`, which had no traffic explicitly directed to it. Rent-a-list, again encouragingly, has substantial mouse activity. Even further, clicks from Adwords resulted in many mouse moves all over the page, generally concentrated over the navigation bar and the text.

Quantitatively, Figure 4(a) shows the average number of mouse moves per visit across all visits to our servers. The results show a clear separation in behavior between the bulk vendors and the higher tier vendors. Traffic from bulk vendors have negligible or no mouse movement, while traffic from the other vendors exhibit some kind of user UI interaction. Figure 4(b) shows CDFs of the number of mouse moves per visit for all visits to the servers. The distributions show the differences between tiers in more detail. For traffic from Google Adwords, for instance, over 75% of the visits had at least 10 mouse moves per visit. Traffic from the bulk vendors are nearly vertical lines — nearly all of the visits had no interaction. From the perspective of mouse activity, traffic from bulk vendors looks suspiciously inorganic.

4.3 Link accesses

Similar to mouse activity, we expect some fraction of real users to click on embedded links in the pages on the sites. Some of these links refer to subpages (Section 4.1.1), but since PeachySkin is essentially a link aggregator to other cosmetics-related sites, the majority of the embedded links across the pages are to external sites. As a result, with organic traffic we expect a subset of visitors will be interested in the site content and access additional links on the pages. In contrast, we would expect that automatically generated traffic will typically have few to no link accesses. Of course, automation can also extract embedded links and visit them as well, but doing so again adds complexity (and effectively results in additional "free" traffic to other sites).

To capture accesses to external links, we encode the external link as a URL that first accesses our site, which then proxies the access to the external site. (Results from search engines and other sites use an equivalent technique to also capture user link access behavior.) To reduce confusion from other types of "clicks", we will refer to these as "link accesses". These link accesses then appear in our Web server logs and are easily identified.

[3]Modulo users with JavaScript disabled, or using ASCII browsers like Lynx, which from our results do not appear prevalent.

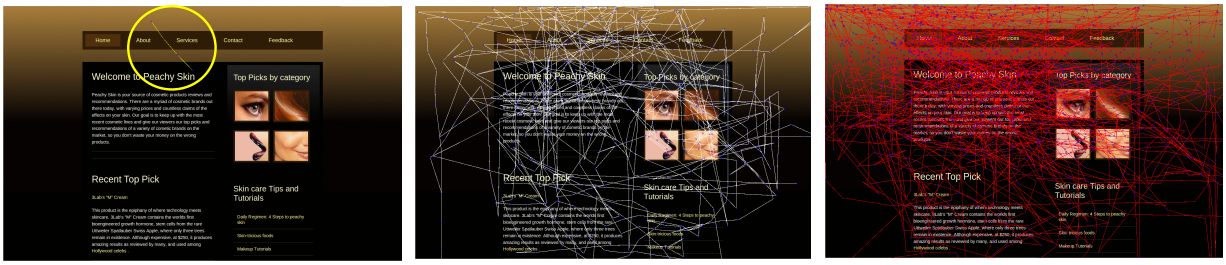
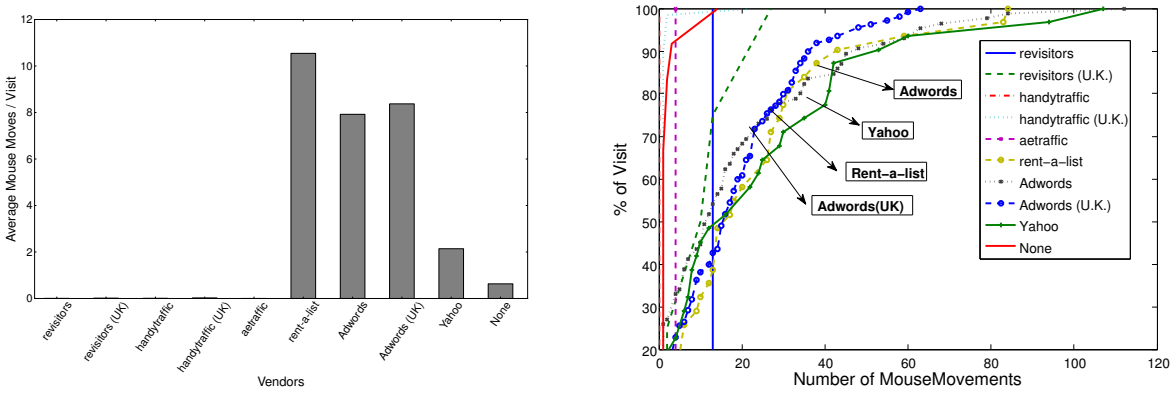


Figure 3: User mouse activity overlayed on the main page visited by traffic from three vendors, one representative from each tier.



(a) Average # of mouse moves per visit

(b) CDF of # of mouse moves per visit across all visits

Figure 4: User mouse activity recorded on each site.

Figure 5(a) shows the average number of link accesses per visit across all visits to our servers. Traffic from bulk vendors result in negligible accesses to links on our pages, while visitors via the higher tier vendors access two or more links on average. Figure 5(b) shows the CDFs of the number of link accesses per visit for all visits. The distributions show the behavior in more detail. As might be expected, a small percentage of visits had many link access via Google and Yahoo. Traffic via the bulk vendors resulted in negligible, if any, link accesses. As with mouse activity, from the perspective of accesses to embedded links traffic from bulk vendors looks suspiciously inorganic.

4.4 User-Agent

Next we examine the distributions of User-Agent strings of the visitors to our sites as another possible signature. The User-Agent field in HTTP requests identifies the client software used to make the request. Web browsers set the field to identify the browser software and the operating system on which the browser is running (Web servers can use this information to tailor content accordingly). Crawlers and other automated clients set the User-Agent field using a unique, often self-identifying string. Automated clients, such as crawlers looking for malware and cloaking, can also use a popular browser+OS User-Agent combination to superficially hide their nature.

In general, we expect the users visiting our sites to reflect the popularity distribution of browsers and operating systems. We used the `user-agent-string.info` tool [2] to extract OS and browser information from the User-Agent strings from the requests to our servers. Figure 6 shows the distribution of browser and operating systems combinations for three representative traffic vendors, one from each tier. Reflecting browser and OS popularities, Windows and IE dominate traffic from the middle of top-tier vendors. In contrast, Linux and Firefox dominate traffic for the bulk vendor Aetraffic. Google Adwords and Rent-a-list have a relatively rich variety of browsers and OSes, including smartphones, whereas the low-tier Aetraffic has two dominant OS/browser combinations.

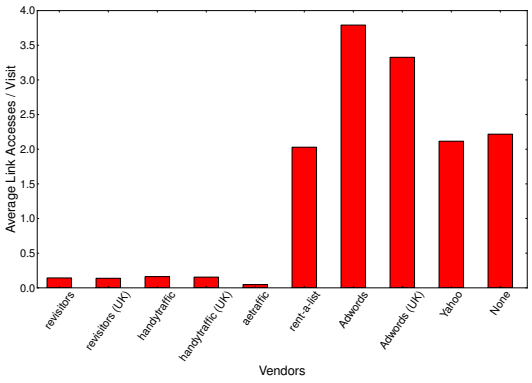
4.5 Referrers

We use the `Referer` field from HTTP requests, when present, to locate the page which led users to visit our sites. We then visit the referrer site and take a snapshot of the page, capturing the context in which our site was advertised in real time.

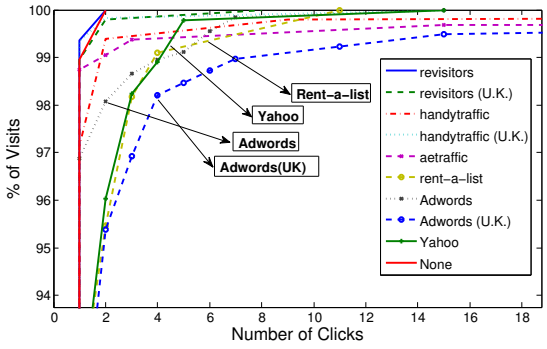
Automatically snapshotting the referring page does not work in every instance for a variety of reasons. Vendors like Google Adwords prevent disclosure of such information by proxying/inserting a referrer's field, resulting in an empty page.[4] On the other hand, other vendors like Revisitors have a referrer field that is always a subpage that is part of their domain.[5] Although we do not observe precisely how our site is advertised, we do however learn the general mechanism they use to advertise it. Early in our experiments with one (now-defunct) bulk traffic vendor QualityTrafficSupply, using our snapshotting tool to visit the referrers induced a HTTP denial-of-service attack on our server. Although annoying to deal with, such behavior serves as a heavy-handed signature that the vendor employs dubious means for delivering traffic to their clients. To avoid this and conserve space, we enabled our snapshotting tool for a limited amount of time. Overall, we were able to obtain snapshots for 10–40% of the total visits to all of our sites.

[4]http://googleads.g.doubleclick.net/pagead/ads?client=ca-pub-0121688737141704

[5]http://www.revisitors.com/admin/?VFJDSz0xMzcx

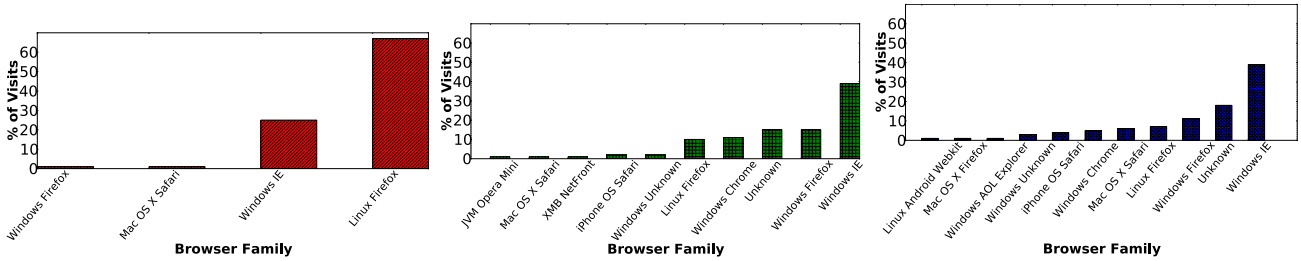


(a) Average # of link accesses per visit



(b) CDF of # of link accesses per visit across all visits

Figure 5: User activity accessing embedded links on each site.



(a) Aetraffic (b) Rent-a-list (c) Adwords

Figure 6: Web browser distribution of traffic from three vendors, one representative from each tier.

Of the successful snapshots, we manually analyzed a random 10% subset of them visually to infer how traffic was directed to our sites. Figure 7 displays some example snapshots. Starting with bulk traffic vendors, all referrers from Revisitors and HandyTraffic for instance came from subpages of the their Web site. These pages rotated different advertisements every time they were accessed (typical advertisements varied from Casino to Adult ads) These referrer snapshots of the bulk traffic sellers suggest that the means of delivering traffic from these vendors is questionable, and consistent with earlier indications of it being automatically generated.

From the middle tier, Rent-a-list, we captured snapshots of a site called `hits4pay.com`. The snapshots showed an error for user name and password. When we visited the page manually, we discovered it was a site that paid people to view advertisements. From previous characteristics such as mouse activity and link accesses, Rent-a-list appeared to deliver organic traffic much like the pay-for-click sites Google and Yahoo. Visiting the referrer URL confirms that the traffic is indeed organic, but nevertheless is fraudulent.

The snapshots from Yahoo displayed various pages, some cosmetic related while others were ad aggregation sites. Some even had our PeachySkin site still advertised on them, as shown in the snapshot in Figure 7.

4.6 Timing metrics

We examined a variety of timing characteristics of the traffic we purchased, including the distribution of durations that users spent on the site, the distribution of the time-of-day of arrival, the time series of daily traffic volumes, and the time in between site visits.

For the first two characteristics, we found little to distinguish the traffic from any of the vendors. We used Google Analytics to estimate the durations of visitors on a site, and it reported little variation across sites. (It also reported anomalous results for one site, making us hesitant to interpret the duration estimates from Google's heuristics too closely.) When looking at the distribution

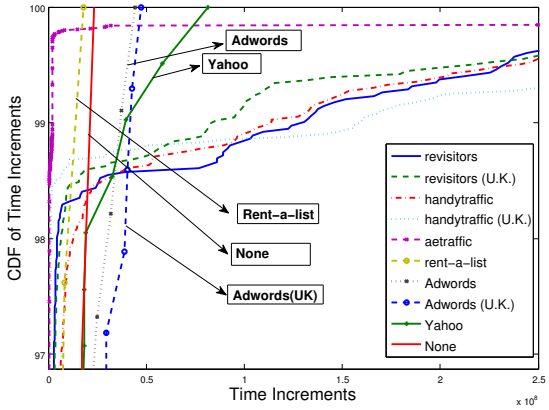


Figure 8: CDF of time between any two clicks normalized.

of time-of-day arrival of traffic to the sites, we also did not see much differentiation. Normalizing traffic to the local time zone of the clients, traffic from all vendors generally fell within the expected diurnal range without any outlying modes.

Finally, we also computed the distribution of time in between visits to each site. Because different vendors have different arrival rates, we normalize the inter-arrival time for each vendor by the amount of traffic received from the vendor to make them comparable. Figure 8 shows the CDFs of the normalized inter-arrival times for all visits for each vendor. Again, we see a clear separation between the low-tier vendors and the others. Google, Yahoo, and Rent-a-list have distributions that fall within narrow time bands without long tails, while the inter-arrivals for bulk vendors for the most part are much smaller, yet have long tails with much larger inter-arrival times.

4.7 Blacklists

If a vendor generates automated traffic to a site, one platform

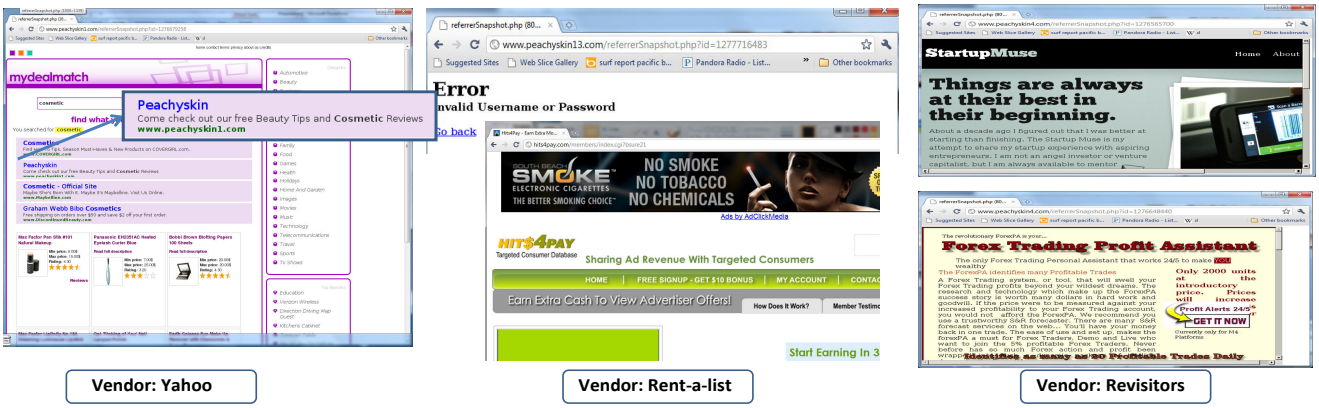


Figure 7: Referrer snapshots, one example from each tier.

for generating such traffic could be click bots on compromised hosts [5]. Due to various kinds of undesirable activity, such as sending spam, compromised hosts are often blacklisted. A potentially interesting question, then, is to what degree traffic to our servers from the various vendors come from hosts that are also blacklisted.

We cross-referenced the IP addresses of hosts visiting our sites with the database of blacklisted hosts maintained by the Composite Block List (CBL). We checked whether the host was ever blacklisted, whether it was on the blacklist at the time of the visit, and also calculated the duration between a host visit and how long before it propagated to the CBL.

We found that only a small fraction of hosts visiting our sites were blacklisted on the CBL across all the vendors. (Note, though, that for the bulk vendors this does translate to hundreds of blacklisted hosts in absolute numbers.) Revisitors (UK) had the highest percentage, with 13% of its click traffic from blacklisted hosts. Otherwise, only 4–10% of traffic from bulk traffic vendors came from blacklisted hosts, and high-tier vendors had an even smaller percentage (0–4%) of traffic from blacklisted hosts with the exception of Adwords (UK) with 12%. Although it might seem surprising that high-tier vendors have a non-zero percentage, our servers received more click traffic than what the high-tier vendors charged for. For example Adwords (UK) charged for 75% of the clicks. In this case, either the high-tier vendors saw the traffic and decided not to charge according to their quality heuristics [8], or this traffic did not originate from the vendors.

5. CONCLUSION

The multi-billion dollar online advertising market has given rise to an interesting secondary ecosystem of bulk traffic vendors who contract to deliver clicks to a site in exchange for fixed up-front payments. In this paper we empirically evaluated the traffic delivered from a sample of these secondary traffic vendors, comparing them to traffic purchased from classic pay-per-click advertising networks embodied by Google and Yahoo. Our primary goal was to understand the extent to which the quality of purchased traffic differs between traditional ad networks and these bulk traffic vendors.

On the one hand, we see strong evidence that the bulk traffic vendors we purchased from are not directing organic traffic to our sites. A range of traffic characteristics either directly do not match what would be expected from real users, or have profiles that differ substantially from PPC traffic.

At the same time, if these vendors do automate traffic to the sites of their customers, they perform some effort to mask such behavior. For the most part, region-specific traffic appears to indeed originate from that region, and coarse-grained timing characteristics are not out of line with those from PPC sites. However, when evaluating traffic from these bulk vendors in isolation, the masking reveals itself as incomplete.

6. REFERENCES

[1] V. Anupam, A. Mayer, K. Nissim, B. Pinkas, and M. K. Reiter. On the security of pay-per-click and other Web advertising schemes. In *Proc. of the 8th WWW*, 1999.

[2] ASAP Consulting. user-agent-string.info API. `http://user-agent-string.info/`.

[3] B. Brow, B. Elgin, and M. Herbst. Click fraud: The dark side of online advertising. Cover Story BusinessWeek, Oct 2006.

[4] Click Quality Team, Google, Inc. How Fictitious Clicks Occur in Third-Party Click Fraud Audit Reports. `http://www.google.com/adwords/ReportonThird-PartyClickFraudAuditing.pdf`.

[5] N. Daswani and M. Stoppelman. The Anatomy of Clickbot.A. In *Proc. of 1st HotBots*, 2007.

[6] Digital Element. NetAcuity IP Intelligence. `http://www.digitalelement.com/our_technology/our_technology.html`.

[7] M. Gandhi, M. Jakobsson, and J. Ratkiewicz. Badvertisements: Stealthy Click-Fraud with Unwitting Accessories. *J. Digital Forensic Practice*, 1(2):131–142, 2006.

[8] Google Team. How does Google detect invalid clicks? `http://adwords.google.com/support/aw/bin/answer.py?hl=en&answer=6114`.

[9] Z. Gyongyi and H. Garcia-Molina. Web Spam Taxonomy. In *Proc. of 1st AIRWeb*, April 2005.

[10] B. J. Jansen and A. Spink. Sponsored Search: Is Money a Motivator for Providing Relevant Results? *Computer*, 40(8):52–57, 2007.

[11] B. J. Jansen and A. Spink. Are Pop-Ups Always Annoying? The Moderating Effect of Ad Relevance on Consumers' Attitude Toward Ads and Websites. 2008.

[12] M. Kane. Pop-ups: Unpopular, but effective. http://www.emarketer.com/Article.aspx?R=1003861, Jan 15 2003.

[13] B. Krishnamurthy and C. E. Wills. Privacy diffusion on the web: A longitudinal perspective. In *Proc. of 18th WWW*, April 2009.

[14] F. Qiu, Z. Liu, and J. Cho. Analysis of User Web Traffic with a Focus on Search Activities. In *Proc. of 8th WebDB*, 2005.

[15] A. Tuzhilin. The Lane's Gifts Vs. Google Report. http://googleblog.blogspot.com/pdf/Tuzhilin_Report.pdf.