

Comments About Click Investigation of A Client's Content Network Ads in Google Adwords

August 9, 2007

Stuart Jenner, Marketek Consulting Group

Contact info: stuartj@marketek-consulting.com

Introduction

For the past few months, I have been investigating “mystery clicks” or “click fraud”, depending on one's perspective, on behalf of a client who had bought Content Network ads in the Google Adwords system.

This particular technology client spent in excess of six figures per year for three years (Feb 2004 to Feb 2007, when I had paused the campaign and started my investigation) on ads in Google. As I looked at the allocation of spending, I was struck by the very high – over 80% - level going to Content Network ads. I was suspicious because the client has a specialized product and it seemed unlikely there was that much available relevant ad inventory on relevant publications. I wondered if perhaps there was some click fraud, for which the client should receive a refund.

In spring 2007, Google announced it was going to make a report available about Content Network ad locations. When this report came out, data was available from December 2006 to present. (Other clients had data from the report available starting later, not December). Once the report was available:

- I looked very closely at data for a three month period of December 2006 to Feb 2007, comparing the Google Performance Report and the client's log file analytics package. I found:
 - The client only seemed to be receiving 22% of the traffic that Google's reports said Google was sending to the client's site. (There were 11,486 clicks on the Placement Report, and 2,191 visits reported in the client's analytics tool as having “googlesyndication” in the referring URL string. There were 2113 unique URLs).
 - Of the 11,486 reported in Google, I estimate 70 to 90% was on sites that were completely irrelevant to the client's target audience. Any clicks coming from these sites were likely fraudulent or mistakes.
 - Google had made a credit to the client, however, it was only about 5% of what the client had spent. I do not take any credit for this refund, it appeared on the statement before I contacted Google.
- I contacted Google in June. Google support had some interesting and helpful insights. However, there are still a number of open issues.
- My conclusion is the client did significantly overpay. The clicks may not have been technically “fraudulent”, however, they are extremely suspicious. There are issues with both the clients and Google's reports during the time the ads were running, making it impossible to fully reconcile the data and specify just how much the client was overcharged for clicks it did not receive, and how much it was overpaying for sites where their ads should not have been running in the first place.

I am writing up some of my experiences in the hope they can be helpful to Google, and possibly also to others. I have assisted clients with web advertising, search engine marketing and optimization and related since 1997.

My Two Wishes

1. I wish I could have access to the Google Performance Report of Content Network ads for a few days in Jan 2007. At one time the data was available from Dec 2006, but recently the start date for pulling data was moved to June 2007. I would like to compare a few days against the client's log file reporting tool reports to evaluate more closely how visits are recorded.
2. Google had a quality control process where clients were not charged for ads on clearly irrelevant sites.

Key Findings

1. I can not definitely say there was click fraud. The lack of data from Google and issues with the client's reports make it impossible to definitely say whether there was fraud.
 - However, definitely there are some very big differences of what Google reported it drove to the clients site and what the client reports receiving, and I don't think all of these differences are due to quirks of the reporting tools.
 - i. Google also suggested Javascript differences on a browser could account for differences: some tools break down if Javascript is not working. I don't think this explains the huge differences I observed though.
 - ii. Google also suggested I read this report:

www.google.com/adwords/ReportonThird-PartyClickFraudAuditing.pdf

However, this does not address the situation of a site reporting far less traffic in its own reporting tools than Google reports in the Placement Performance report. This article is over a year old and needs to be updated.

- There are legitimate differences of what gets reported as a "click" in one tool compared to another. Just because one source differs from another, for example a log file analytics report and a Google Analytics report, doesn't mean there's necessarily fraud.
 - Also, a lot can go wrong in setting up log file reporting tools
2. The new Google Performance Report of Content Network ad spending is helpful. However, it falls short in the following areas:
 - If we look at the total for "Content Network" in an Ad Group report, and the total in a Performance Report, we get a different number. The problem is Google does not break out ads shown on Gmail and elsewhere in the Google network in the Performance Report.
 - The data availability changed during the time I was working on my investigation. Originally, I was working with data from December 1, 2006 to present. Later, data was only available from June 1, 2007 to present. I asked a Google rep why, and whether the data in the original time period was accurate or not. I did not get a complete response to this question. Instead, the Google rep sent me an Acrobat file with all the IP addresses for a certain time period. This was just about useless. I am guessing that the data was not fully accurate but that he didn't want to say this. I am not sure at all that the reported data from June 1 to present is accurate either, I do not feel I have a good way to assess it.

3. Google Content Network has some serious problems at matching sites to ads.

- The particular client I was assisting was a major technology company.
- I checked several hundred of the URLs that the client's ads ran on, and the reality is: only a few were what I would consider to be "legitimate" sites. Many were "Made for AdSense" sites, with content that appeared written in order to trigger ads.
- This particular client had the single work "proxy" in its ad buy. There were many other terms too. However, I noted many sites showed up that were "proxy" surfing sites, aimed at people who wanted to hide their tracks. People go to these sites to then visit other sites, such as Ebay or social networking, or job hunting sites, during the workday when they should be working.
- In setup, the ads were specified as English only and were targeted only to the US.
 - i. I was thus very frustrated to observe sites with .cn (China) suffixes and language, sites in Spanish and other non-English languages, sites in English but from outside the US (www.newsday.co.tt/classifieds/0,90.html or www.geekzone.co.nz/content.asp or china.alvin.hk), and worst of all, sites that were clearly NOT relevant for a technology audience. Some were Adult sites others were misspellings like "24hourfotness.com" (not 24hourfitness). Others were link spam sites, like americanexpresscredit.com and some were combo adult and link spam (sexalg.com, sexbnat.com). This page www.myspaceproxy1.com had 23 clicks at a cost of \$40.48 – note this is NOT a Myspace page, it is a clever proxy hiding page.
 - ii. I put some of these sites in email to the Google rep and attached a spreadsheet of their Performance Report. He didn't seem to read what I wrote, instead he wrote back saying "Despite the locations of the websites, the clicks your ads received should have come from English-language sites. After going through a number of the sites on which your ad was featured, we did not find any content on the same page as your ad that wasn't in your target language. If you have any evidence of sites in other languages on which your ads were showing, please let us know so we can take appropriate action."
 - iii. In a follow up email, which I never heard back on, I did send specific non-English URLs, for example,
liny1.022idc.cn/jiadi.html
www.100programas.com
anoniem-surfen.eigenstart.nl
- More frustrations: some of the sites did not load when I tried to check them. It had only been a few months since the ads had been running. The sites were not present in the cache of Google, and were not included in archive.org. My guess is the sites were "junk sites" that had been closed down. They often had alphanumeric combo URLs. An example: 10589-https.com, which had 63 clicks at a cost of over \$170.

- Another frustration: where did the ad actually run? Some of the pages I would visit did not have any ads on them, yet the report listed them as the specific page where the ad ran. It is always possible an ad wasn't running that particular time I checked the page, but still, the number of times this happened was quite a surprise. See this Dutch adult site: zulu.messageboard.nl/12866/viewtopic.php
- Google's site targeting, which prices ads per thousand impressions (in contrast to Content Networking price per click), does allow one to specify sites of interest. However, Content Network does not allow one to proactively specify sites. Instead, one can only specify which sites one does NOT want an ad running on after the ad has already run. There are problems with this:
 - i. There can be thousands of sites. Many of them may only have 1 or 2 impressions. Checking every one of these sites is very time consuming.
 - ii. There's no way to exclude a class of sites. One can't say "exclude all domains that have an alphanumeric combo" for example.
 - iii. One apparently can't exclude sites with other characteristics, such as sites with a country name suffix or content for a specific language. This did not happen much, but it should not happen at all.
- I want to close with by mentioning this site where ads ran: <http://www.adtest.info/> It combines link spam, German content, photo and MP3 ads and hints at adult content, but has nothing related to enterprise content to trigger an enterprise software company's ad.

4. There appear to be major problems with "Click-fraud" or "inappropriate clicks"

- As mentioned above, the quality of sites is often dubious. I have no idea how people would find them, they certainly would not rank well on any terms in Google's organic listings. I also have no idea why someone interested in enterprise software would be at them.
 - What are the odds that a major technology vendor would have two clicks, out of five total impressions, on this fan page about the soap opera General Hospital: www.angelfire.com/ma/LazyLobsterPub/fanpage.html
- I can not prove this was fraud. However, it certainly unlikely qualified clicks were coming from these pages. Google's making money, and the sites are, but is my client? Not likely!
- The Google support rep, and Google's Adwords Help, discuss the matching algorithms they use. The rep's email said "we found that their traffic fit a pattern of normal user behavior." The bottom line is: no algorithm can replace human judgment. Just as human judgment is sometimes trumped by greed, I think Google's algorithm could be trumped by greed as well. The reality is: this particular client had not set separate bids for Content Network ads. Instead they'd left in place a very high bid that applied to their Content ads as well as search ads. I think this high bid trumped other factors in the Google algorithm.

Suggestions for Google

1. Companies spending \$9K per month for 41 straight months that then significantly cut their spending are probably not satisfied customers. You might want to do a better job of staying in touch with your customers who are spending six figures per year on ads.
2. Don't withdraw report data from a certain date range, then not explain why it is no longer available.
3. Do allow people to specify in much greater detail where their ads should run in the Content Network. You might want to combine Site Targeting and Content Network and let advertisers select how they want to pay.
 - a. In particular, pay attention to the original language and geography specified for the Campaign.
 - b. Also allow people to specify characteristics of a site to include, or a type of site to exclude. I realize there's some subjectivity, so maybe a "don't run my ads on sites like this" feature would be good.
 - c. Some advertisers DO like having ads on parked domains, or typos. Sometimes these are a relevant advertising tool.
4. Consider caching pages to show advertisers that the ads actually ran
 - I realize this could be a huge amount of disk space. On the other hand, it is very odd that I could go to a specific URL and not see any ads on the pages where my ad supposedly ran. If a site isn't cached in the Google organic results, that does say something about the quality of the site.
5. The matching algorithms need a considerable amount of improvement, as mentioned in the previous section.
6. Link to articles about comparing Google Analytics to specific log file packages, IF these articles have helpful advice.
 - I was checking with a major brand name analytics package, and it had some major differences. It took me a while to learn these, but it turns out the people at the vendor, and presumably others, were aware of them for a long time. I would have saved a lot of time if I'd known these quirky differences up front.
7. Make a cache of the sites where ads run. Even if the specific ad is not shown in the cache, as suggested above, at least we'll be able to tell what the site is about and whether it was legitimate or not.
8. Provide an easy way to indicate which sites should be moved into "site targeting" out of "Content Network"
9. Allow people to specify Content Network sites, not just to exclude them.
10. Provide examples of reports and of how third party Analytics tools were set up to help with comparisons of Google Ad reports. Google Help has great text, but often no screen shots. Maybe allow vendors to include links to screenshots of their specific setups, so Google is not directly responsible for this content but still makes vendors aware of its importance.

11. Your support team supplied an article about Click Fraud, where clients are concerned about too many clicks showing up on their site. However, you don't seem to have a white paper about the opposite angle, of too few clicks on a site's log compared to the clicks showing up in the Google reports.
 - I realize there's a certain amount of drop-off. Someone may click on an ad then not wait for it to load. However, how often does this really happen? Probably not 80% of the time which is what seemed to be the case for this client.

Lessons for Advertisers

For Advertisers considering Content Network targeting, here are some words of advice:

1. Google's support team is helpful. However, don't expect a reply to every question.
2. Log file reporting tools take some work to set up. Do a lot of research about what you're trying to do, and make sure your tool is sufficiently accurate. For example, one report mentions how some tools differ significantly from others. Here's the report:

<http://www.stonetemple.com/articles/analytics-report-may-2007.shtml>

Also see this, comparing Adwords and Google Analytics.

<http://forums.seochat.com/website-analytics-76/big-discrepancy-between-clicks-reported-by-adwords-vs-analytics-clicktracks-135068.html>

3. Start with low bids, then gradually increase your bid.
 - Make sure these bids are separate from your regular search bids
 - High bids are more likely to attract fraudulent clicks and to attract more impressions
4. Do look at the Placement Performance Report. See what sites your ad is actually showing on.
5. Be willing to drop the Content Network. It may be more trouble than it is worth.
6. Compare what you see in the Google reports with your own log file reports.
 - You may want to activate the Google Analytics tracking to make this easier to see. However, be aware of a potential impact on slowing down the user experience – at least this was a concern of my client.
 - Specifically activate the tracking URLs in Analytics and Adwords.
 - Ads from Google's Content Network should show up with a URL of "googlesyndication" in them. However, there's a chance the URL might instead only show up in your log files as a referral directly from a site. So, be careful in checking.

- I searched on URLs in the third party reporting tool with “google” in them. It appears some of these may be from ads, and others from natural “organic” results. For example,

www.google.com/search?sourceid=navclient&ie=UTF-8&rls=GGLJ,GGLJ:2006-42,GGLJ:en&q=CLIENT

According to the Analytics reporting person, this could come from either organic or paid ads.

7. If you think there’s click-fraud going on, where illegitimate clicks are depleting your budget, please be aware of the following:

- Google does adjust charges if its automated tools detect something fraudulent
- They will also investigate if you file a complaint. However, it appears you may need to do this within 60 days of when the charge occurs.
- Unless you’re spending on a huge campaign, you may be money ahead just swallowing your losses and not investigating. The time to figure this out may be more costly than the return.
- You should definitely have some sort of tracking going on and use that data as well as the clicks themselves to assess what’s going on. The analytics tracking URLs will be helpful in this regard.

8. Remember Site Targeting, with ads priced per thousand, is always an option.

- You may want to run Content Network for a month or two, see which sites appear, then try to run those sites in Site Targeting and close down the Content Network ads. However, there are no guarantees the sites will be available in Site Targeting.

Extracted data

Here’s a comparison of data from the two sources.

Site	Google reported # of clicks from Placement Performance Report	Client’s analytics package reported # of clicks	Comment: how much do Google reported clicks differ from client’s analytics package?
Nighi	83	11	Google over by 72, or 650% of 11.
Openvpn	63	40	Google over by 23, or 58% of 40.
Nexopia	14	2	600% over
Mylot	10	7	42% over
myspaceproxy1.com	23	4	475% over
Metroflog	65	6	983% over
Listenernetwork	66	3	2,200% over
apache-ssl.org	4	6	An oddity: ANALYTICS TOOL reports more clicks than Google does. This is

			the only instance I could find of this happening.
Answers.com	414	19	More than double
Reference.com	14	7	Double
Format is a number followed by https, for example, 10559-https.com	321 (this one of 10166-https.com Had 49 and this one 10589-https.com Had 63)	The two specific URLs have 37 and 45.	Analytics tool doesn't make it easy to see all searches with -https.com in the domain
Total with Googlesyndication	11,486	2,191	9,295 clicks more per Google than per Analytics tool. Avg CPC was \$3.07 during this time, so \$28,535.

Specific URLs from the Client's Analytics report

In the Analytics Tool list of URLs, I found over 1000 URLs with this structure:

http://pagead-us.google syndication.com/pagead/ads?oe=utf-8&client=ca-opera_800x30&format=800x30&alternate_ad_url=http%3A//textads.opera.com/ads/%3Fformat%3D800x30&channel=754%20en&url=http%3A//www.opera.com/download/lng/854/ouw854_ru.lng

Note: Opera does show up in the Placement Performance Report. Analyzing this is hard in the Client's Analytics data because the URLs for searches conducted using the Opera browser are also captured. So, this specific example is not included in the table above, but is mentioned here because so many URLs had this structure.

Nighi (listed in the table above).

From the Analytics report:

http://pagead2.google syndication.com/pagead/ads?client=ca-pub-8824833016068065&dt=1166399060328&imt=1166399060&format=120x240_as&output=html&channel=8341500775&url=http%3A%2F%2Fnighi.com%2Fgateway%2Fleft.php%3Furl%3Dhttp%3A%2F%2Fwww.youshare.com%2Fview.php%3Ffile%3DBaabull.wmv&color_bg=000000&color_text=FFFFFF&color_link=FFFFFF&color_url=FFFFFF&color_border=000000&ad_type=text&ref=http%3A%2F%2Fnighi.com%2Fgateway%2Fgateway.php%3Furl%3Dhttp%253A%252F%252Fwww.youshare.com%252Fview.php%253Ffile%253DBaabull.wmv&u_h=768&u_w=1024&u_ah=738&u_aw=1024&u_cd=32&u_tz=-300&u_his=3&u_java=true

There are 11 URLs like this, each with one visit according to the Analytics tool.

Apache-SSL (listed in the table above)

I found about 80, each with two visits, with a structure similar to this:

http://pagead2.googlesyndication.com/pagead/ads?client=ca-pub-7420574511380438&dt=1165594733406&lmt=1161911173&format=120x600_as&output=html&url=http%3A%2F%2Fwww.apache-ssl.org%2F&ad_type=text_image&ref=http%3A%2F%2Fmembers.domainhost.com%2FwebControl%2Fsslsetup.bml&cc=1997&u_h=800&u_w=1280&u_ah=770&u_aw=1280&u_cd=32&u_tz=-300&u_his=3&u_java=true

Note: Apache-SSL was the only one I found where Google's report showed less traffic than the Analytics tool reported.

Answers.com (listed in the table above)

These URLs do not contain any indication of Google in them. Here's the URL from the Google Placement Performance report:

www.answers.com/topic/high-level-data-link-control

This specific one is not found in the ANALYTICS TOOL report.

Domain Parking (not listed in the table above)

I found 30 or so with this structure:

<http://domains.googlesyndication.com/apps/domainpark/domainpark.cgi?client=TUCO3114&s=adoolt.com>

Reference.com (listed in the table above)

I found some clearly from third party sites, but that had "google" in them. THERE WERE NONE OF THESE WITH MORE THAN 1 SEARCH in the Analytics report. An example is this one from reference.com, which drove 1 visit:

http://dictionary.reference.com/ads.html?q=Network&adq=Data%20Network%20Security&orig_channel=definition&google_page_url=http%3A//dictionary.reference.com/search%3Fq%3DNetwork