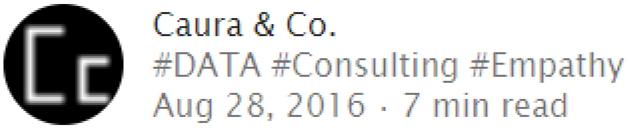
# Investigating Google Adwords Click Fraud



Caura & Co.
#DATA #Consulting #Empathy
Aug 28, 2016 · 7 min read

***Update 1:***

*After an internal investigation, Google called and admitted the discovery of 25 previously not reported fraudulent clicks and promised a refund.*

*I, of course, pushed for confirmation of that in writing. In writing they have not admitted any guilt—just restated that there are 25 "invalid" clicks, which I now see on my Adwords account.*

> *When first you came, I fancied you might be a Thief: now that you try to bribe me from my duty, I am sure you are one;*

Disclaimer: I am not singling out Google. All platforms, including Facebook, are guilty of fraudulent clicks. Google just happens to be the most transparent.

My wife recently started a preschool in Bellevue while I took ownership of her online advertising and analytics. I'm no stranger to online advertising. I know well-dialed campaigns are a crapshoot in comparison to promotions offline.

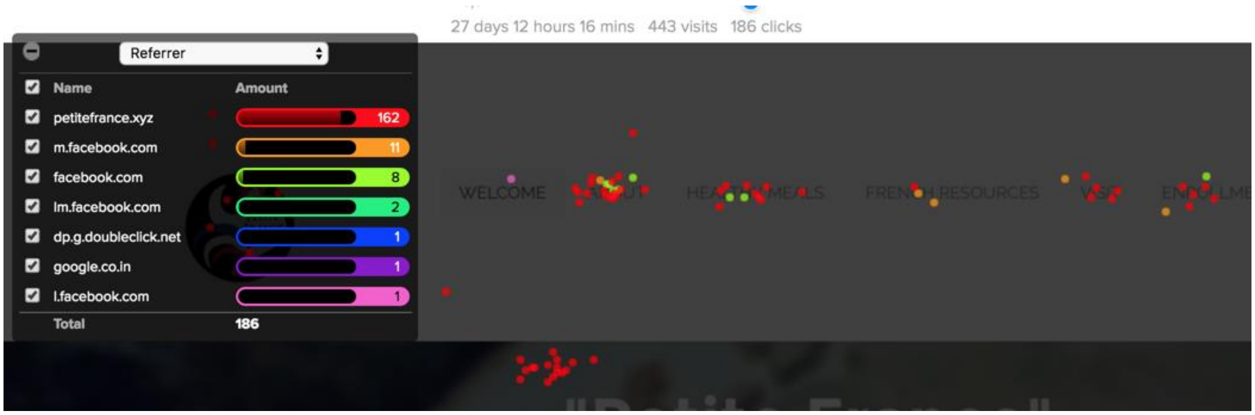
I assumed I could outsmart the system by focusing on traffic from Google searches alone and ignoring display.

Shockingly, I learned how naive I was about the scale and source of the fraud. Below are my findings.

## Where Has All the Money Gone?

It has been 3 weeks and I have spent close to $400. More than $350 were spent on search. Google Adwords showed 132 clicks, but my wife's business did not see any of it.

Digging deeper, I discovered her site's heat map had no engagement from Google referrers:



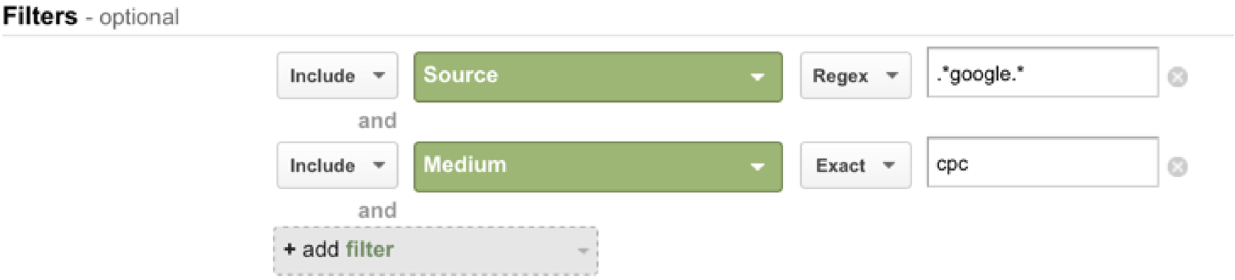
27 days 12 hours 16 mins 443 visits 186 clicks

Referrer

| ☑ Name | Amount |
|---|---|
| ☑ petitefrance.xyz | 162 |
| ☑ m.facebook.com | 11 |
| ☑ facebook.com | 8 |
| ☑ lm.facebook.com | 2 |
| ☑ dp.g.doubleclick.net | 1 |
| ☑ google.co.in | 1 |
| ☑ l.facebook.com | 1 |
| Total | 186 |

Clicks were mostly from direct emails my wife sent out earlier.

How is that possible? It was time to look at the Google sessions.



**Filters** - optional

Include Source Regex .*google.*

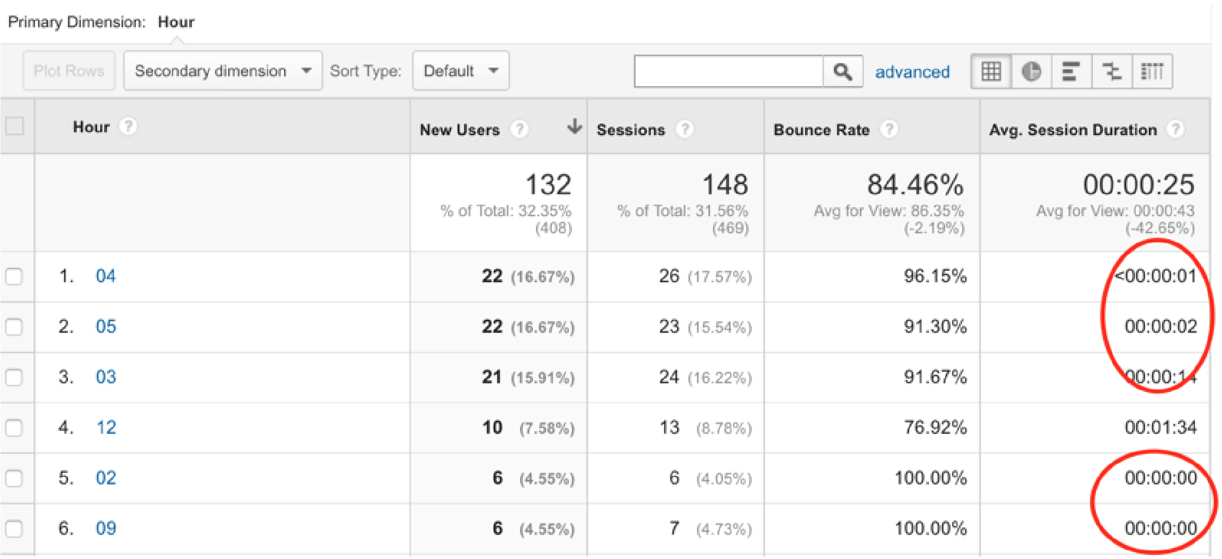
and

Include Medium Exact cpc

and

Selecting CPC medium for paid traffic only.

What are all these near-zero time sessions?



Primary Dimension: Hour

| Hour | New Users | Sessions | Bounce Rate | Avg. Session Duration |
|---|---|---|---|---|
| | 132<br>% of Total: 32.35% (408) | 148<br>% of Total: 31.56% (469) | 84.46%<br>Avg for View: 86.35% (-2.19%) | 00:00:25<br>Avg for View: 00:00:43 (-42.65%) |
| 1. 04 | 22 (16.67%) | 26 (17.57%) | 96.15% | <00:00:01 |
| 2. 05 | 22 (16.67%) | 23 (15.54%) | 91.30% | 00:00:02 |
| 3. 03 | 21 (15.91%) | 24 (16.22%) | 91.67% | 00:00:14 |
| 4. 12 | 10 (7.58%) | 13 (8.78%) | 76.92% | 00:01:34 |
| 5. 02 | 6 (4.55%) | 6 (4.05%) | 100.00% | 00:00:00 |
| 6. 09 | 6 (4.55%) | 7 (4.73%) | 100.00% | 00:00:00 |

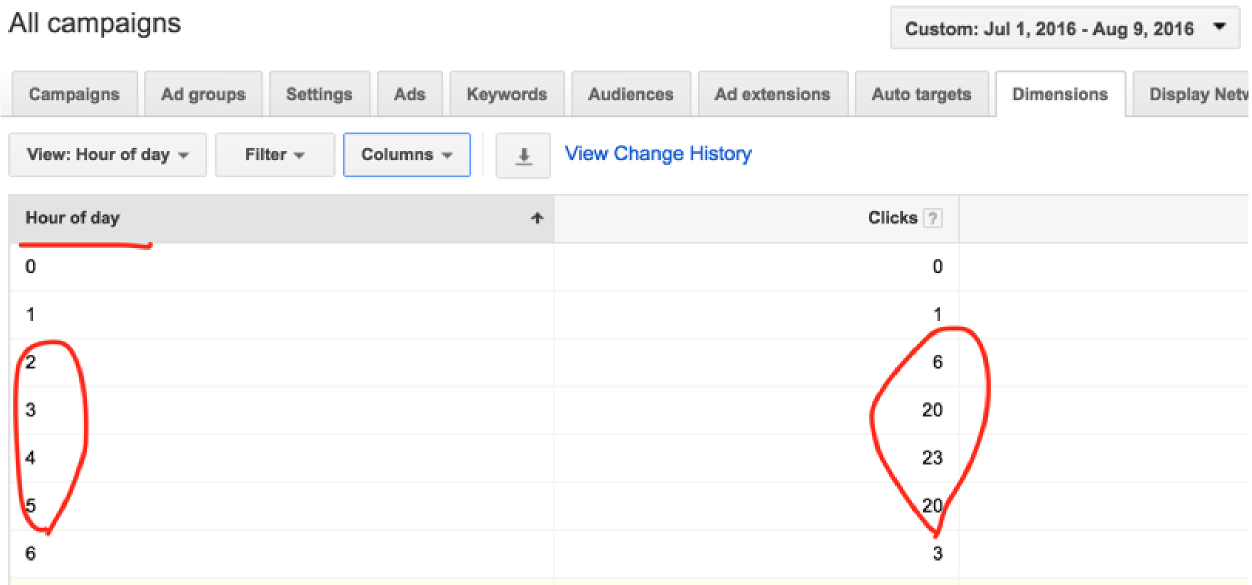
In Google's ideal world, high bounce rate reflects a disconnect between content and what users expect.

At this point I know that a big portion of clicks did not even engage with the site. The billion dollar question: can I prove it?

# Adwords shooting blanks

So far I found that there were 71 clicks (53% of total clicks) between 2:00 am and 5:00 am.

Adwords confirms the clicks and the hours:



Custom: Jul 1, 2016 - Aug 9, 2016

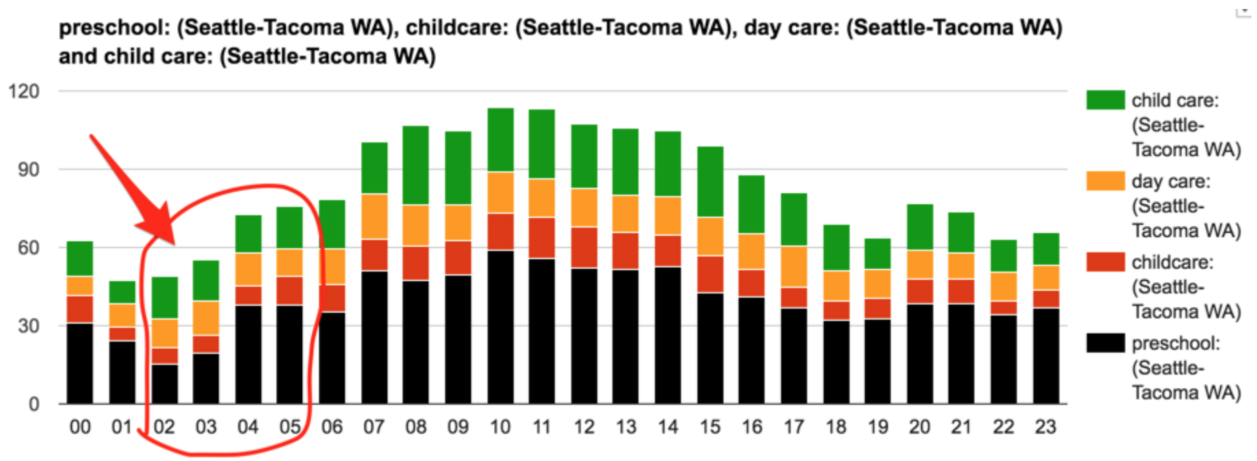
View: Hour of day

| Hour of day | Clicks |
|---|---|
| 0 | 0 |
| 1 | 1 |
| 2 | 6 |
| 3 | 20 |
| 4 | 23 |
| 5 | 20 |
| 6 | 3 |

Are parents with young children searching for a preschool in the middle of the night? Time to ask my friends who have young kids. Maybe I'm missing something.

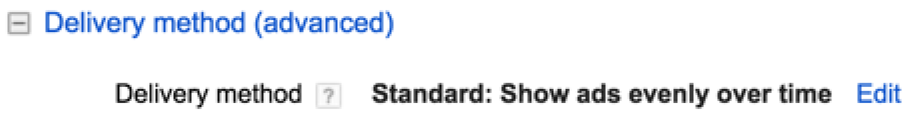
Three calls later, I am convinced my results are suspicious. But as, at one time, a student of math, I am well aware that I need a bigger sample size.

Google Search Trends to the rescue:



Only 13% of all searches for my clicked keywords happen between 2 and 5 am. The ads shown for my wife's business were triggered by variations of these keywords.

This is unusual because I selected standard ad delivery across my campaigns:



⊟ Delivery method (advanced)

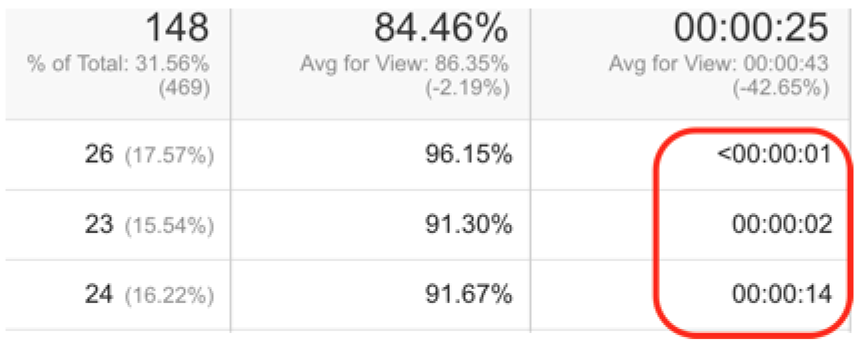
Delivery method ? Standard: Show ads evenly over time Edit

Change History confirmed the same Delivery method

So, no, parents are not searching for a preschool in the middle of the night. Moreover, the exposure of my wife's site to search during off-peak times is abnormal.

An additional issue affecting our exposure is how competitors schedule their ads. But, it plays less of a role because of the following observations.

Lets look now at these sessions closer:



| | | |
|---|---|---|
| 148<br>% of Total: 31.56%<br>(469) | 84.46%<br>Avg for View: 86.35%<br>(-2.19%) | 00:00:25<br>Avg for View: 00:00:43<br>(-42.65%) |
| 26 (17.57%) | 96.15% | <00:00:01 |
| 23 (15.54%) | 91.30% | 00:00:02 |
| 24 (16.22%) | 91.67% | 00:00:14 |

This is short for a preschool website. In life there are a few activities that we devote a lot of time to: finding the right preschool must be at the top of the list. I bet that even with crappy sites, parents spend more than a few seconds.

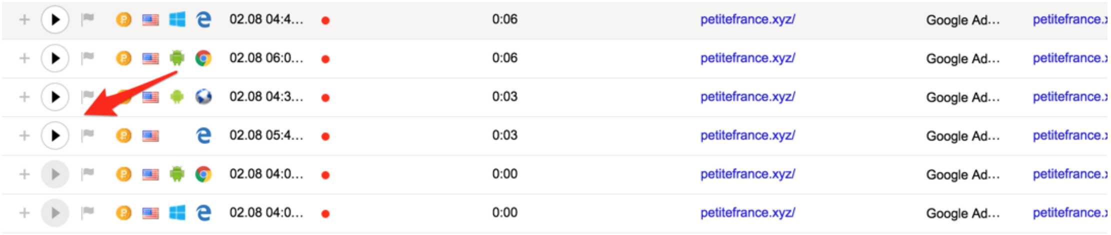
But I still wonder if prospective customers who clicked the ads were immediately overwhelmed by the website.

The bounce rate shows me there was no interest beyond the landing page. But, did prospective customers even see the website? Let's look at the page load time:

1. average page load time (which is 2.33 last I checked) > session duration time
2. avg. page load time is within or better median internet page load times

Unfortunately, we can't conduct this analysis using just Google Analytics. With Google Analytics, the session duration is expected to be zero for the 100% bounce pages, so even if sessions lasted 10 seconds, GA would still capture them at zero seconds.

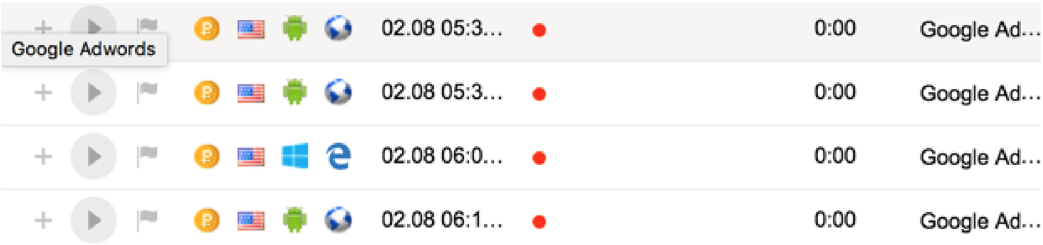
Luckily, from the get-go I had more tools in my arsenal. Come GA's competing product, Yandex Metrica, and my favorite Webvisor. Since I had Webvisor installed, Yandex has been tracking full duration of sessions.



| | | | | | |
|---|---|---|---|---|---|
| 02.08 04:4... | ● | 0:06 | petitefrance.xyz/ | Google Ad... | petitefrance.x |
| 02.08 06:0... | ● | 0:06 | petitefrance.xyz/ | Google Ad... | petitefrance.x |
| 02.08 04:3... | ● | 0:03 | petitefrance.xyz/ | Google Ad... | petitefrance.x |
| 02.08 05:4... | ● | 0:03 | petitefrance.xyz/ | Google Ad... | petitefrance.x |
| 02.08 04:0... | ● | 0:00 | petitefrance.xyz/ | Google Ad... | petitefrance.x |
| 02.08 04:0... | ● | 0:00 | petitefrance.xyz/ | Google Ad... | petitefrance.x |

Webvisor starts recording as soon as 1 second after the html header loads

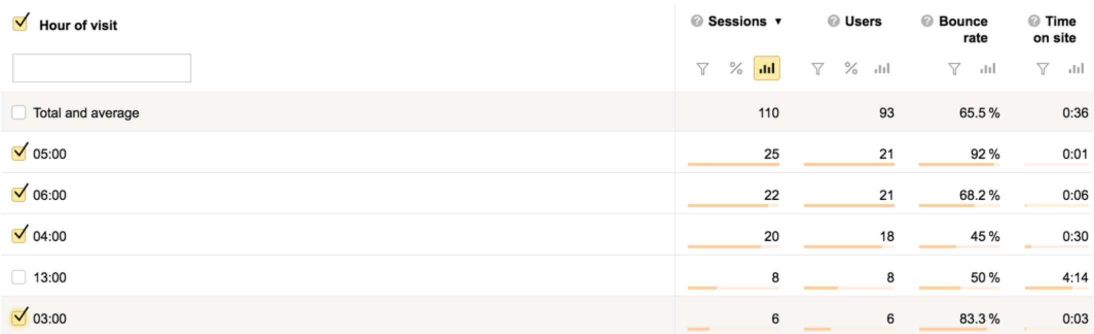
The majority of Google Adwords clicks are zero durations:



Google Adwords

| | | | |
|---|---|---|---|
| 02.08 05:3... | ● | 0:00 | Google Ad... |
| 02.08 05:3... | ● | 0:00 | Google Ad... |
| 02.08 06:0... | ● | 0:00 | Google Ad... |
| 02.08 06:1... | ● | 0:00 | Google Ad... |

Here are visits from Google Adwords. Remember those 2:00 am—5:00 am visits?



| ☑ Hour of visit | Sessions ▾ | Users | Bounce rate | Time on site |
|---|---|---|---|---|
| □ Total and average | 110 | 93 | 65.5 % | 0:36 |
| ☑ 05:00 | 25 | 21 | 92 % | 0:01 |
| ☑ 06:00 | 22 | 21 | 68.2 % | 0:06 |
| ☑ 04:00 | 20 | 18 | 45 % | 0:30 |
| □ 13:00 | 8 | 8 | 50 % | 4:14 |
| ☑ 03:00 | 6 | 6 | 83.3 % | 0:03 |

Measuring session duration with the time counter is an approximation. The server is not pinged every millisecond. What was actually 3–4 seconds, may have shown up as 1 second. And looking at a page for 3–4 seconds is probably enough time to decide if it's crap.

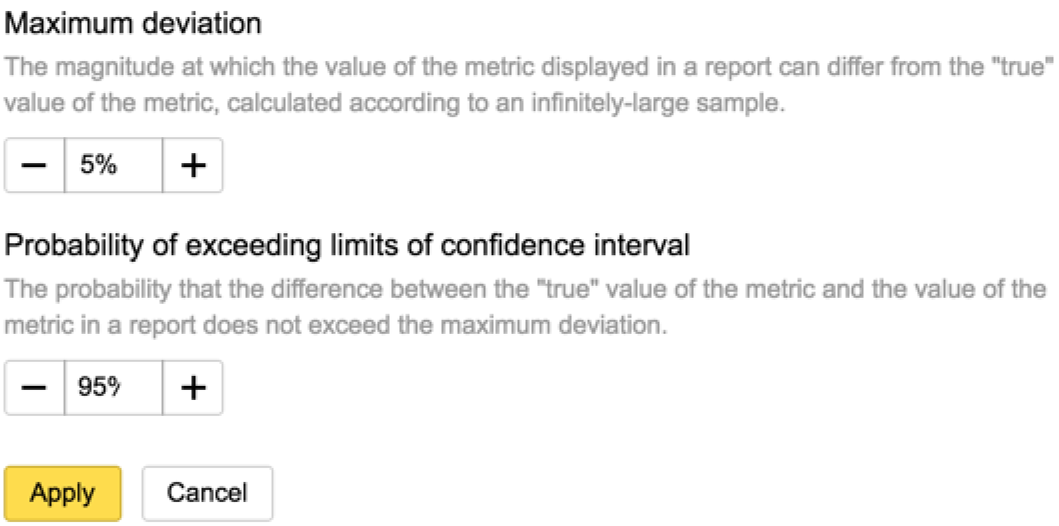
I am not convinced by this argument. Yet, if I play skeptic, I have to entertain this possibility too:

$$H_0 : \ Page\ Load\ time \ - \ Time\ on\ Site \ \le \ 0$$

$$H_a : Page\ Load\ time \ - \ Time\ on\ Site \ > \ 0$$

Need to demonstrate the alternative Hypothesis to prove the above 3-4 second argument is bogus.

Conveniently, Yandex Metrica has a tool to build confidence intervals around “Time on Site” (for simplicity, assume Page Load Time is 2.5–3 seconds).



**Maximum deviation**

The magnitude at which the value of the metric displayed in a report can differ from the "true" value of the metric, calculated according to an infinitely-large sample.

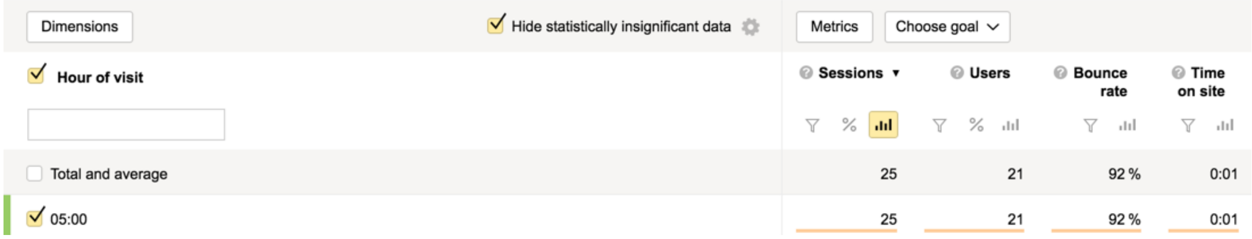
− 5% +

**Probability of exceeding limits of confidence interval**

The probability that the difference between the "true" value of the metric and the value of the metric in a report does not exceed the maximum deviation.

− 95% +

Even at this confidence interval, about a third of the midnight clicks did not see the website because it did not have time to load.



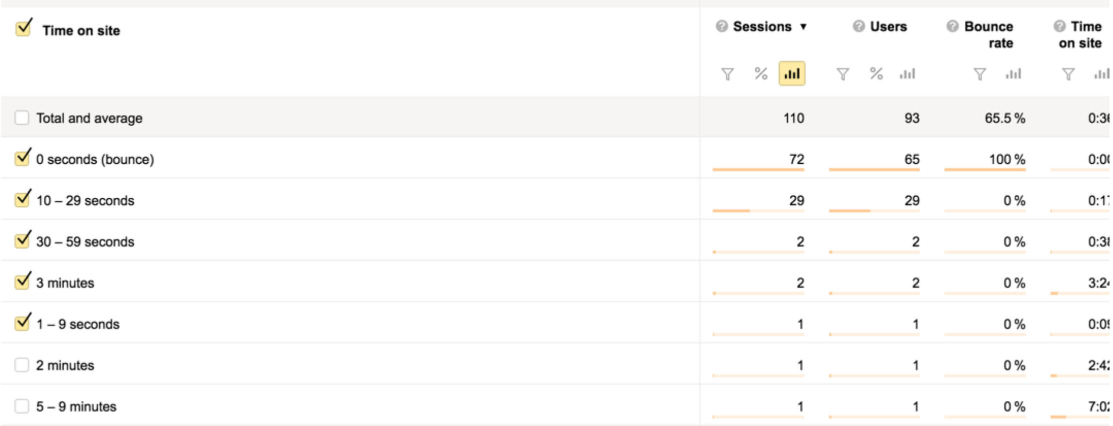
☑ Hide statistically insignificant data

| ☑ Hour of visit | Sessions ▾ | Users | Bounce rate | Time on site |
|---|---|---|---|---|
| □ Total and average | 25 | 21 | 92 % | 0:01 |
| ☑ 05:00 | 25 | 21 | 92 % | 0:01 |

Rejecting the sceptic 3–4second argument in favor of the alternative (my argument)

# What is going on here?

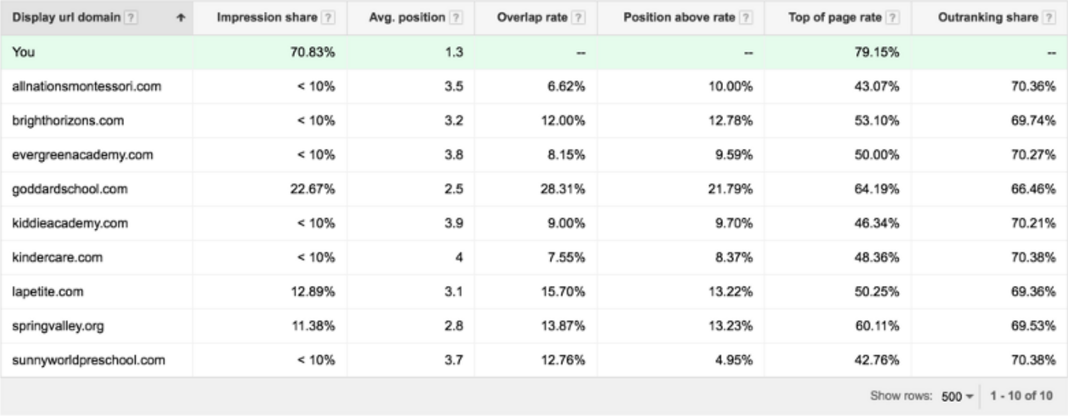
To summarize, about 65% of clicks are bounces. 53% are abnormal and occur in the middle of the night. And half of these abnormal clicks (22% of total) are fraudulent.



| ☑ Time on site | Sessions ▼ | Users | Bounce rate | Time on site |
|---|---|---|---|---|
| □ Total and average | 110 | 93 | 65.5 % | 0:3 |
| ☑ 0 seconds (bounce) | 72 | 65 | 100 % | 0:0 |
| ☑ 10 – 29 seconds | 29 | 29 | 0 % | 0:1 |
| ☑ 30 – 59 seconds | 2 | 2 | 0 % | 0:3 |
| ☑ 3 minutes | 2 | 2 | 0 % | 3:2 |
| ☑ 1 – 9 seconds | 1 | 1 | 0 % | 0:0 |
| □ 2 minutes | 1 | 1 | 0 % | 2:4 |
| □ 5 – 9 minutes | 1 | 1 | 0 % | 7:0 |

So far, 22% is my lowest fraud estimate. However, 53% is my best guess. But, who cares? Who can benefit from this information?

My wife's business is competing with many (hundreds of) daycare centers and preschools. Just 9 of these are advertising on Google Adwords for the same keywords:



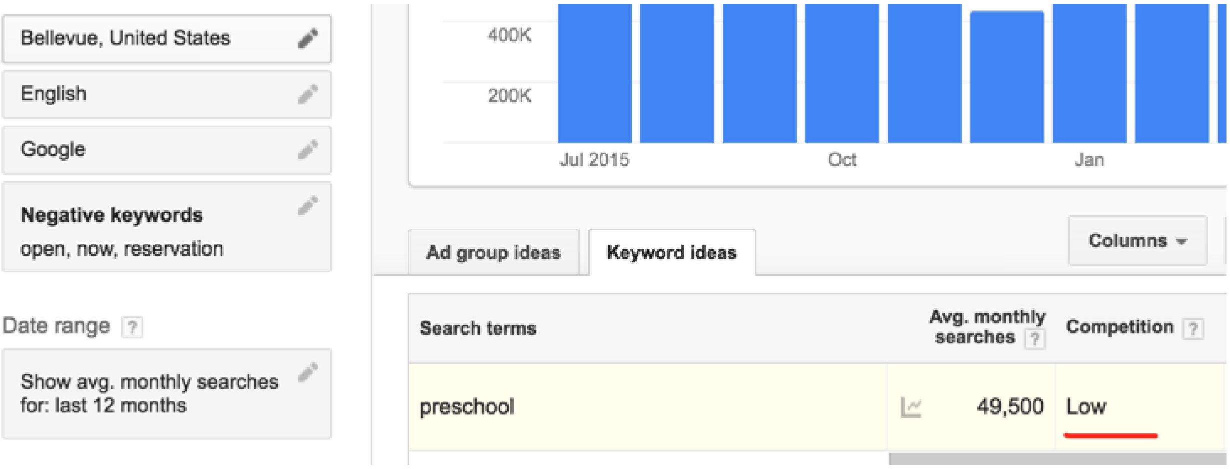
| Display url domain | Impression share | Avg. position | Overlap rate | Position above rate | Top of page rate | Outranking share |
|---|---|---|---|---|---|---|
| You | 70.83% | 1.3 | -- | -- | 79.15% | -- |
| allnationsmontessori.com | < 10% | 3.5 | 6.62% | 10.00% | 43.07% | 70.36% |
| brighthorizons.com | < 10% | 3.2 | 12.00% | 12.78% | 53.10% | 69.74% |
| evergreenacademy.com | < 10% | 3.8 | 8.15% | 9.59% | 50.00% | 70.27% |
| goddardschool.com | 22.67% | 2.5 | 28.31% | 21.79% | 64.19% | 66.46% |
| kiddieacademy.com | < 10% | 3.9 | 9.00% | 9.70% | 46.34% | 70.21% |
| kindercare.com | < 10% | 4 | 7.55% | 8.37% | 48.36% | 70.38% |
| lapetite.com | 12.89% | 3.1 | 15.70% | 13.22% | 50.25% | 69.36% |
| springvalley.org | 11.38% | 2.8 | 13.87% | 13.23% | 60.11% | 69.53% |
| sunnyworldpreschool.com | < 10% | 3.7 | 12.76% | 4.95% | 42.76% | 70.38% |

Show rows: 500 ▾ 1 - 10 of 10

Most of these schools have hundreds or even thousands of locations across the country:

*goddardschool.com, lapetite.com, kindercare.com, brighthorizons.com, kiddieacademy.com, evergreenacademy.com*

It is amazing how few preschools advertise on Adwords for our target location. According to Yelp, there are hundreds of them in our location.



Bellevue, United States

English

Google

**Negative keywords**
open, now, reservation

Date range ?

Show avg. monthly searches for: last 12 months

| Search terms | Avg. monthly searches ? | Competition ? |
|---|---|---|
| preschool | 49,500 | Low |

Preschools don’t advertise?

Have they been pushed out due to low conversion?

# 3rd party js scripts—what do they do exactly?

After skimming through the above list of large franchises, I discovered the following 3rd party scripts on their websites:

> simpli.f*i (Local Programmatic Advertising & DSP Platform),* silverpop.com *(Marketing Automation), Clicktale (Monetizing & Conversion), and Omniture (does not require introduction) among them*

Following Google's own suggestion, I searched for "*simpli.fi virus*". It returned a lot of web results - not that I am trying to accuse simpli.fi or any of the above 3rd party services by association. There will always be plenty of bad agents out there. The question is whether Google is genuinely doing absolutely everything to fight them, or is it the case of one hand feeding the other. After all, the interests align closely.

# Bots—they surf, browse, click

So far I have several theories for what is going on, the central premise being that one of the 3rd party ad services used on these corporate sites, delivers low CPC (good) and high conversion (also good) to its customer by directing spam traffic towards competing higher bidders (not good), such as my wife's business.

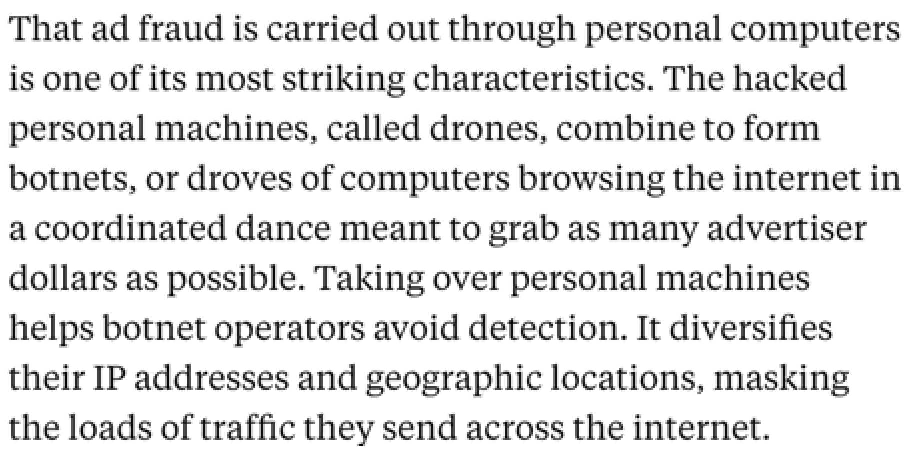
In fact, Criteo, Google's largest ad competitor, even conducted its own technology study in a suit against SteelHouse, one of its smaller competitors, when the latter managed to steal away a number of Criteo's clients (Lara O'Reilly's original story). The gist of the study: clicks to your site can be generated on behalf of a user without that user's knowledge of your website.

Another alternative explanation: **Botnets.**

When I brought up clicks during odd hours to my dad, he recognized the pattern from his work as a Network Security Expert (he works at Checkpoint). He explained it like this:

> *A big share of our phones and computers are infected with malware that can send traffic anywhere at someone's mere will. To make this traffic less noticeable to the owner of the device, much of this traffic is send when the device is not used by the owner (aka in the middle of the night).*

I then did my own search and found this backstage interview by Alex Kantrowitz with a Google security expert, incidentally with the same first name as that of my father:



That ad fraud is carried out through personal computers is one of its most striking characteristics. The hacked personal machines, called drones, combine to form botnets, or droves of computers browsing the internet in a coordinated dance meant to grab as many advertiser dollars as possible. Taking over personal machines helps botnet operators avoid detection. It diversifies their IP addresses and geographic locations, masking the loads of traffic they send across the internet.

from Sasha, one of Google's fraud fighters

# This Warrants a Study

What I am really saying is that someone should put together a test. To put together such a test would only require:

- event tracking/collection,
- factor identification,
- correct sampling

So far such research happened exclusively in start-ups that were quickly bought out by Google. @veritasium conducted his study of Facebook fraud, but the closed nature of the experiment leaves room for Google and Facebook officials to undermine it.

Why not perform an open study?

*Segah Meer is a Data Technologist building a Data Consulting business of the future out of Seattle*