Quantifying Online Advertising Fraud: Ad-Click Bots vs Humans

Adrian Neal, Sander Kouwenhoven
firstname.lastname@oxford-biochron.com

Oxford BioChronometrics SA

January 2015

Abstract

We present the results of research to determine the ratio of Ad-Clicks that are human initiated against those that are initiated by automated computer programmes, commonly known as ad-bots. The research was conducted over a 7 days period in early January 2015, using the advertising platforms of Google, Yahoo, LinkedIn and Facebook. The results showed that between 88 and 98 percent of all ad-clicks were by a bot of some kind, with over 10 per cent of these bots being of a highly advanced type, able to mimic human behaviour to an advanced extent, thus requiring highly advanced behavioural modelling to detect them.

1 Introduction

In May 2014, according to the Financial Times[1] newspaper, part of a Mercedes-Benz on-line advertising campaign was viewed more often by automated computer programmes than by human beings. It was estimated that only 43 per cent of the ad impressions were viewed by humans. Later, in December, Google made a similar announcement[3] when it stated that its research has showed that 56.1 per cent of ads served on the Internet are never "in view". From our own informal research using existing data from detecting spam-bots, it was thought that the level of bots involved in ad fraud might be considerably higher than was being generally reported. Consequently, we set out to conduct a controlled experiment to answer the following questions:-

1. What is the ratio between ad-clicks charged for, ad-clicks from bots and ad-clicks from humans, and
2. How many different types of ad-click bots can we observe.

2 Internet Bots - what we know

According to Wikipedia[4], an Internet bot, also known as web robot, WWW robot or simply bot, is a software application that runs automated tasks over the Internet. Typically, bots perform tasks that are both simple and structurally repetitive, at a much higher rate than would be possible for a human alone. The largest use of bots is in web spidering, in which an automated script fetches, analyses and files information from web servers at many times the speed of a human. Each server can have a file called robots.txt, containing rules for the spidering of that server that the bot is supposed to obey or be removed.

In addition to these uses, bots may also be implemented where a response speed faster than that of humans is required (e.g., gaming bots and auction-site robots) or less commonly in situations where the emulation of human activity is required, for example chat bots.

There has been a great deal of controversy about the use of bots in an automated trading function. Auction website eBay has been to court in an attempt to suppress a third-party company from using bots to traverse their site looking for bargains; this approach backfired on eBay and attracted the attention of further bots. The United Kingdom-based bet exchange Betfair saw such a large amount of traffic coming from bots they launched a WebService API aimed at bot programmers through which Betfair can actively manage bot interactions.

Bot farms are known to be used in online app stores, like the Apple App Store and Google Play, to manipulate positions or to increase positive ratings/reviews while another, more malicious use of bots is the coordination and operation of an automated attack on networked computers, such as a denial-of-service attack by a botnet.

Internet bots can also be used to commit click fraud and more recently have seen usage around Massively Multiplayer Online Roleplaying Games (MMORPG) as computer game bots. A spambot is an internet bot that attempts to spam large amounts of content on the Internet, usually adding advertising links.

Bots are also used to buy up good seats for concerts, particularly by ticket brokers who resell the tickets. Bots are employed against entertainment event-ticketing sites, like TicketMaster.com. The bots are used by ticket brokers to unfairly obtain the best seats for themselves while depriving the general public from also having a chance to obtain the good seats. The bot runs through the purchase process and obtains better seats by pulling as many seats back as it can.

Bots are often used in MMORPG to farm for resources that would otherwise take significant time or effort to obtain; this is a concern for most online in-game economies. Bots are also used to artificially increase views for YouTube videos. Bots are used to increase traffic counts on analytics reporting to extract money

from advertisers. A study by comScore found that 54 percent of display ads shown in thousands of campaigns between May 2012 and February 2013 never appeared in front of a human being.

In 2012 reporter Percy Lipinski reported that he discovered millions of bot or botted or pinged views at CNN iReport. CNN iReport quietly removed millions of views from the account of so-called superstar iReporter Chris Morrow. A followup investigation lead to a story published on the citizen journalist platform, Allvoices[2]. It is not known if the ad revenue received by CNN from the fake views was ever returned to the advertisers.

3 Generally observed behaviour

All bots have a common set of properties. It can be said that a bot:-

- primarily exists, directly or indirectly, for economic gain,
- mimics, to any extent, the actions of a human using a computer,
- repeats such actions multiple times,
- initiates activity,
- executes only the minimum necessary actions to complete its task.

Bot behaviour, at the atomic level, falls into any one the following general classifications (*with examples of type*):-

1. Sends a single message (*Denial of Service Bots, Distributed Denial of Service Bots, Ad Click Bots, Ad Impression Bots*),
2. Sends a single message and waits for response (*Email Spam Bots, Ad Click Bots, Ad Impression Bots, Online Banking Bots*),
3. Sends multiple messages asynchronously (*Denial of Service Bots, Distributed Denial of Service Bots*),
4. Sends multiple messages asynchronously and waits for one or more responses (*Online Spam Bots*).

In behaviours 2 and 4, the sender address (i.e. the IP Address) must be valid for the response to be received (although not necessarily the point of origin), while behaviours 1 and 3 can accomplish their task without this prerequisite condition, making them considerably harder to detect their true point of origin.

4 How the research was conducted

In order to limit the level of non ad-platform bot activity being recorded, individual web pages were created specifically as the click target for the ad, one per ad platform. HTTP GET logging software was enabled for each of these web pages, recording each HTTP GET request that was made to the web server. Embedded on each of the target web pages was a JavaScript library, providing data collection functions to the web page. These functions were designed to record:-

1. Device-specific data, such as the type of web browser being used by the device, predetermined calculations to estimate CPU capabilities, hashing of HTML CANVAS elements to determine screen resolution, etc.

2. Network-specific data, such as the geo-location of the ip address, determining if the ip address was a proxy server, details of the DNS used, fixed-size data packet transmission latency tests, etc.

3. Behaviour-specific data, such as when and how the mouse and keyboard were used for devices that raise mouse and keyboard events, while for mobile devices, recording the data from the gyro, accelerometer and touch screen events.

Each of the three data sets that were being collected from the web page, were sent to their own separate web server using a variety of transmission methods. These were:-

1. Creating an empty SCRIPT Document Object Model Tag element, setting the SRC attribute to the URL of a collection script and parsing the collected data as a HTTP GET parameter.

2. Creating an new IMG Document Object Model Tag element, and again setting the SRC attribute to the URL of a collection script and parsing the collected data as a HTTP GET parameter.

3. Creating a Document Object Model HTTPRequest instance (also known as an AJAX request) to post the data to a collection script on the same server from where the web page was loaded.

Including the server HTTP GET request logs, this gave us in total four streams of data, which were relatively independent of each other, providing us with the ability to create much richer models of ad-bot behaviour and enabling us to create thoroughly-researched ad-bot classifications.

The advertising platforms used were Google, Yahoo, LinkedIn and Facebook. The ad-click budget allocated was around £100 (GBP) per platform, which was the maximum lifetime budget for the ad campaign and was used as fast as possible on each platform.

5 Types of ad-fraud bot detected

While observing the behaviour of bots, we were able to create six classifications of bot types, that we propose as a class of the Kouwenhoven-Neal Automated-Computer-Response Classification System and are described thus:-

Basic - (Ad-Clicks Only) Identified through the difference between the number of Ad-Clicks charged by a specific ad platform, and the number of consolidated HTTP GET requests received for the unique URL that was designated as the ad-click target for the ad campaign running on the ad platform.

Enhanced - Detected through the correlation of a HTTP GET request received by an ad-server for a specific ad, with the AJAX-transmitted record of the web-browser load event. If the recorded load event is inconsistent with the standard load event model, the HTTP GET was made by a bot.

Highly Enhanced - Detected through the use of advanced JavaScript processor metrics. A bot is evident if the client-side code execution is inconsistent with known code execution models.

Advanced - In an elementary attempt to impersonate human behaviour, the page is loaded into the web-browser, but the combination of the length of time that the page is supposedly viewed and the subsequent number and type of supposed user activities show very high levels of inconsistency with our models of normal human behaviour.

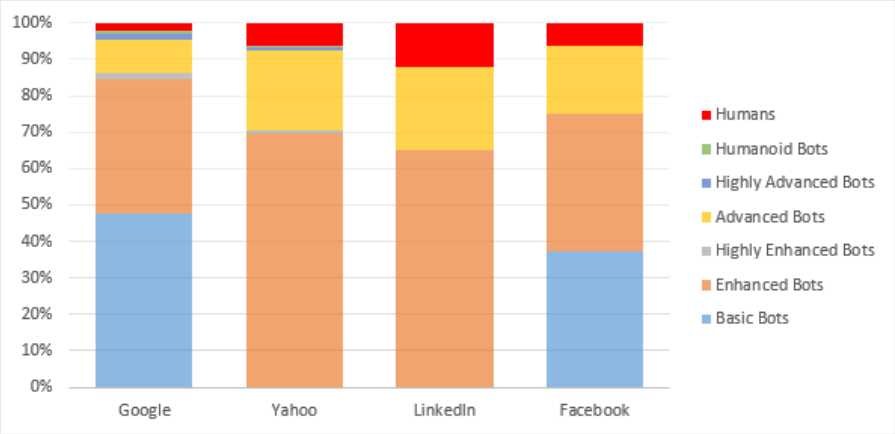
Highly Advanced - A significant attempt at impersonating human behaviour, the bot views the page for an amount of time that would seem reasonable. Both mouse and keyboard events are triggered and the page might be scrolled up or down. However, using cluster analysis, the pseudo randomness is highly detectable.

Humanoid - Detected only through deep behavioural analysis with particular emphasis on, for example, recorded mouse/touch movements, which may have been artificially created using algorithms such as Bezier curves, B-splines, etc., with attempts to subsequently introduce measures of random behaviour, mimicking natural variance.

6 Results

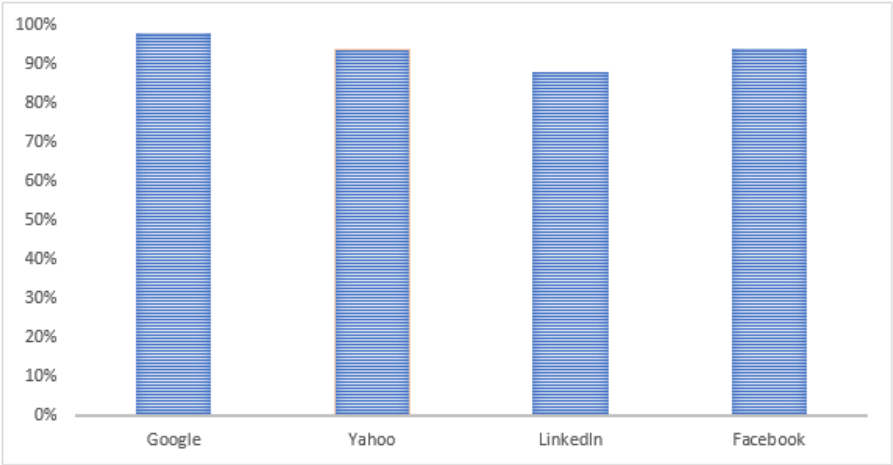
Our research found that at best, 88 percent of the ad-clicks were made by bots on the LinkedIn ad platform, while at worst, 98 percent were from bots on the Google ad platform.

Figure 1: Ratio of Ad-Bot Clicks to Human Clicks



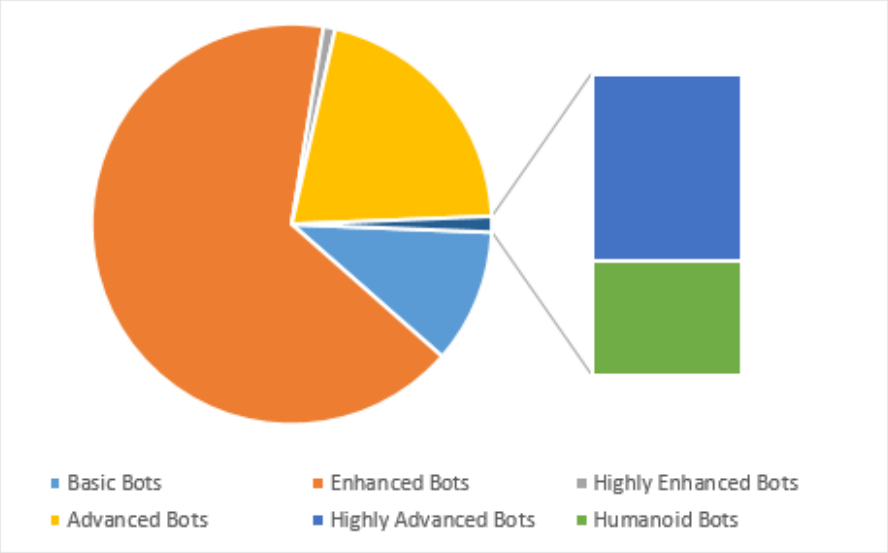
There were no instances where we were not charged for an ad-click that was made by any type of bot.

Figure 2: Prevalence of Overcharging of Ad-Clicks



The prevalence of the different types of ad-bot was not entirely as expected. We expected that the majority of bots would be of the basic type and that they would diminish in a linear fashion as they became more advanced. This was not the case, as the Enhanced bot was by far the most widely observed, with the second being the Advanced bot.

Figure 3: Types and Prevalence of Ad-Bots



The limited sample size and duration of this test notwithstanding, these findings are in keeping with our general observations of bot activity through conventional bot detection software, which analyses Internet traffic as a whole on a post real-time basis.

7 Conclusion

There are perhaps few industries where overcharging on such a scale as demonstrated here would be tolerated, but until very recently, the ability to model both human and bot behaviour at the necessary level of complexity (and thus hold advertising platforms to account) was not commercially feasible.

However, with the rise of what is commonly referred to as Big Data, the ability to collect, store and process vast amounts of data in real-time at reasonable cost, while modeling complex human (and human-like) behaviour, has fundamentally changed the balance of power in the relationship between advertisers and the advertising platforms.

References

[1] R. Cookson. *Mercedes online ads viewed more by fraudster robots than humans.* Financial Times, 2014.

[2] Percy Lipinski. *CNN's iReport hit hard by pay-per-view scandal.* Allvoices, 2013.

[3] Z Wener-Fligner. *Google admits that advertisers wasted their money on more than half of internet ads.* Quartz, 2014.

[4] Wikipedia. *Internet Bots.* Wikimedia Foundation, Inc., 2015.