# An Exploration of Key Traits of Click Fraud

**Andrew Swindlehurst – Data Analyst at PPC Protect**

**Abstract:** Click fraud costs advertisers billions of dollars every year in lost advertising budget. Yet despite efforts to reduce this budget waste, click fraud is still set to rise over the upcoming years. Understanding click fraud is paramount to preventing it, so in this paper, key traits of click fraud are identified, and rates of click fraud are examined against various factors. No correlation between cost per click or keyword search volume against rates of click fraud has been found, although a weak correlation between keyword competition and click fraud is clear.

## Introduction

Click fraud can be defined as the practice of deceptively clicking on adverts with the intention of either increasing third-party website revenues or exhausting an advertiser's budget (Wilbur, K. C. & Zhu, Y., 2009). The malicious practice of click fraud is nothing new and has plagued online advertisers for well over almost two decades. It has been reported in publications for well over a decade (Mann, C. C., 2006) after it gained prominence in the mid-2000's. Despite this prominence and the demonstrable harm (Li, X. et al. 2011), there seems to be little progress being made to remove click fraud from advertising platforms. The lack of action on the part of advertising platforms may stem from the fact that some search engines may have an interest in allowing some click fraud to occur under certain conditions (Wilbur, K. C. & Zhu, Y., 2009). Even though this practice may run contrary to the interests of the businesses advertising on those platforms.

Since the first pay-per-click (PPC) platforms were launched, various efforts have been made to reduce the damage done by the effects of click fraud. These efforts to identify and reduce click fraud have come in many novel forms, notably "bluff ads" (Haddadi, H. 2010) and probability-based blocking using the Dempster-Shafer Evidence Theory (Walgampaya, C. et al., 2010). Only recently has there been significant progress in building an effective way of combating click fraud, which is in part due to machine learning and adaptive algorithms (Saad, S., 2011).

Despite all the innovations made to prevent click fraud, the impact it has on the advertising industry is not diminishing. Click fraud still cost advertisers an estimated $7.2 billion in 2016 (ANA, 2016) and brands $6.5 billion in the US alone. New platforms for click fraud have arisen over the last few years, most notably smartphone click fraud which grew 102% from January to April 2017 (Pixalate, 2017). Botnets are another fraud platform that has been responsible for a significant amount of fraud over the past few years, the click fraud operation "Methbot" generated $3-5 million in fraudulent revenue every day at its peak (WhiteOps, 2016) and there is no doubt more waiting to be discovered.

Little is known about the critical traits of fraudulent clicks in general, for example, when fraudulent clicks are most likely to occur, or if keyword factors (i.e., cost per click) have any effect on the rate of fraud. In this paper, we demystify lesser known traits of click fraud. The aim is to identify which ads may be most at risk of click fraud, thus allowing advertisers to be more vigilant. Using data from clicks protected by the PPC Protect algorithm, we can gain an insight into real click fraud traits.

## Results

Understanding how the PPC Protect platform operates is key to understanding the data presented throughout this report. As a click passes through the algorithm and is identified to be fraudulent, the IP address is blacklisted, preventing ads from being displayed to that IP address again. Blocking the IP address reduces the amount of repeat click fraud the ad experiences.

A hypothetical example could be a fraudster committing programmatic click fraud. The fraudster aims to click on an ad 20 times, and has access to two IP addresses and a single computer. The fraudster clicks the ad three times before the system identifies them as fraudulent. They then swap IP addresses to their second, click for the fourth time and are automatically blocked again as the system identifies them as the fraudster through other variables (i.e., device fingerprinting).

The fraudster can no longer see ads from either of his two IP addresses, preventing the remaining 16 fraudulent clicks. This is excellent for the advertiser as they are saving more of their ad budget, but it does hamper the range of data collected for this study. This hampering is due to the database only registering two of the four clicks committed as fraudulent (the two clicks where the system identified the fraudster). The limitations of the data

do not reduce the serviceability of the data collected by the system, as out of the 20 possible fraudulent clicks, only two will be recorded as fraudulent. This "non-repeat" click fraud is what will be referred to as fraud from now on in this paper.

### Click Fraud Over Time

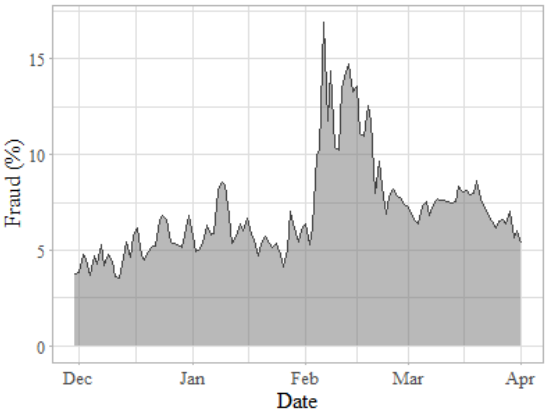


*Fig.1 – Fraudulent clicks as a percentage of all clicks over the period of 2017-11-30 to 2018-04-01.*

The effect of this "non-repeat" click fraud can be seen in Fig. 1 above. The mean percentage of fraudulent clicks over the 123-day period was 7.01%, meaning the system was blocking 7.01% of new IP addresses each day. For 13 days in February, the system experienced a significant jump in fraudulent clicks, this increase of fraudulent clicks is due to a new, somewhat large and heavily frauded ad campaign loading into the protection system. On the few days the system found a significant amount of fraud and started blocking the fraudulent users, the amount of fraud was tending back down to the average rate of fraudulent clicks. This was due to fraudulent users having been added to the blacklist by the system.

### Click Fraud Per Weekday

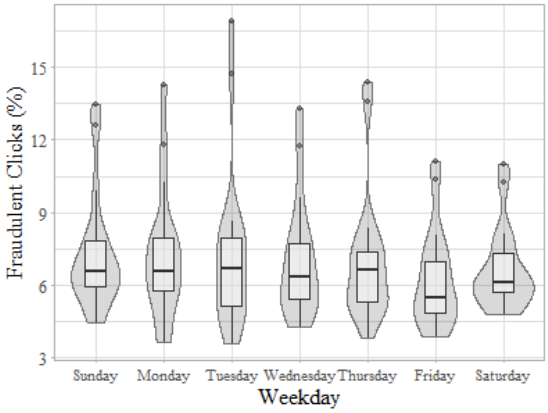


*Fig.2 – Violin plot displaying fraudulent clicks as a percentage of all clicks over the days of the week.*

Friday is the day with the lowest rate of fraudulent clicks, shown in Fig. 2. Friday is the only day with a significant amount of fraud below the average rate. Saturday has little variation and is consistently around average, whereas the midweek days such as Tuesday show high variation and the highest rates of fraud. If we remove the outliers which correspond to the spikes of fraud in February, there is no significant variation of fraud over the week.

### Click Fraud and Cost Per Click

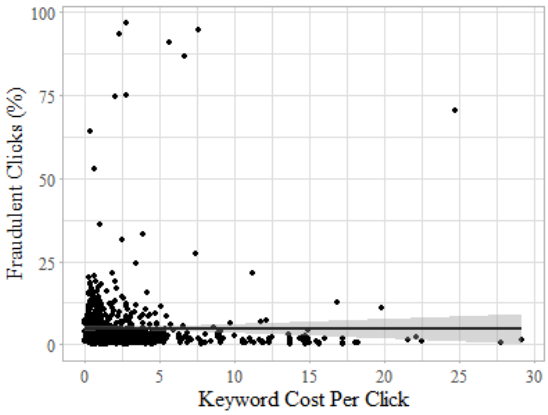


*Fig.3 - Fraudulent clicks as a percentage of all clicks against keyword Cost per Click (points) with a linear regression line.*

Using the cost per click (CPC) values from Google AdWords Keyword Planner, it is shown that there is no correlation between CPC and the rates of click fraud (Fig. 3). Using linear regression analysis to find the correlation coefficient between CPC and the rate of fraudulent clicks we find a value of negative 0.003. A correlation coefficient value this small shows there is no correlation between CPC and the rate of fraudulent clicks. Therefore adverts at any CPC are as likely as any other ad to experience fraudulent clicks.

### Click Fraud and Keyword Search Volume

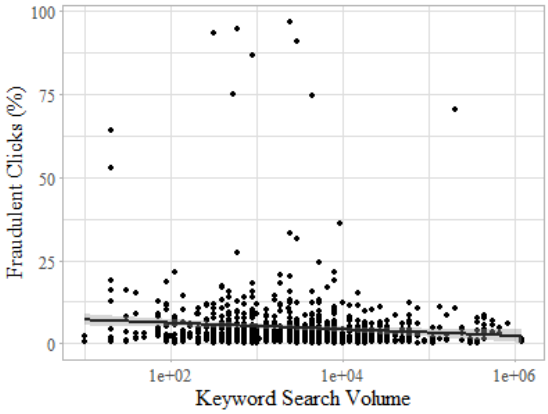


*Fig.4 - Fraudulent clicks as a percentage of all clicks against keyword search volume (points) with a linear regression line.*

Performing the same steps using search volume in place of cost per click produce similar results. Linear regression analysis returns a correlation coefficient of negative 0.022 (Fig. 4). Though stronger than the correlation between CPC and rates of fraudulent clicks, the correlation between volume and rates of fraudulent clicks is still very weak, if non-existent.

### Click Fraud and Keyword Competition

The correlation between keyword competition and rates of fraudulent clicks are much stronger than that of search volume and CPC. Performing linear regression analysis on keyword competition and rates of fraudulent clicks results in a correlation coefficient of positive 0.104, showing a positive relationship between keyword competition and rates of fraudulent clicks (Fig. 5). Therefore, if advertisers are bidding on highly competitive keywords, they are more likely to experience fraudulent clicks.

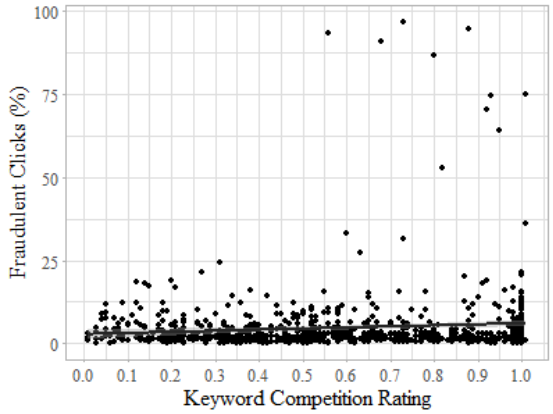


*Fig.5 - Fraudulent clicks as a percentage of all clicks against keyword search competition (points) with a linear regression line.*

### Click Fraud Per Industry

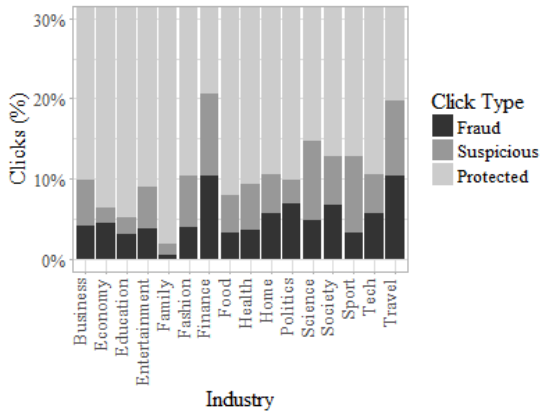


*Fig.6 - Fraudulent clicks as a percentage of all clicks over various industries.*

Examining rates of fraud in various industries, the data shows Finance and Travel are the most frauded industries (Fig. 6). Bloomberg obtained similar findings where Finance was the most frauded industry (Eglin et al., 2015). However, Travel in Bloomberg's findings ranked fourth, and Family ranked second. The drastic difference in findings between Bloomberg and the data in this study may be due to the small sample size of 'Family' data available to us. Out of the 16 industries listed, Family accounted for less than 0.01% of all clicks. Therefore it should be expected to see differences due to this small sample size.

### Click Fraud Per Top Level Domain

Large variations in the rate of fraud are experienced by different Top Level Domains (TLDs). Generic TLDs (i.e. *.com, .gov, .edu)* experience fraud at well over twice the average, with *.net* experiencing a massive 17% fraudulent clicks (Fig. 7), and the 'new' generic TLD *.xyz* facing 17.5% fraudulent clicks. Though it must be noted that these account for a small number of clicks in the database, roughly 5%. 'Localised' TLDs (domains targeting specific areas of the world, ie. *.ca*) experience fraud dependant on the targeted area. The Australian TLD *.au* experienced 12.3% fraud, 175% above the average rate. Many experience average, if not slightly above average fraud, notably *.co.uk* with 9.5% and *.com* with 8%.

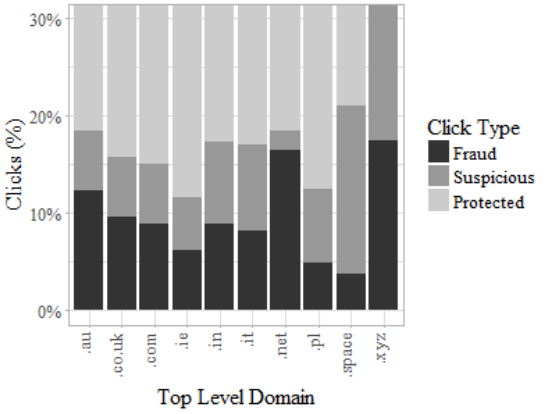


*Fig.7 - Fraudulent clicks as a percentage of all clicks over various top level domains.*

### Click Fraud and Advert Type

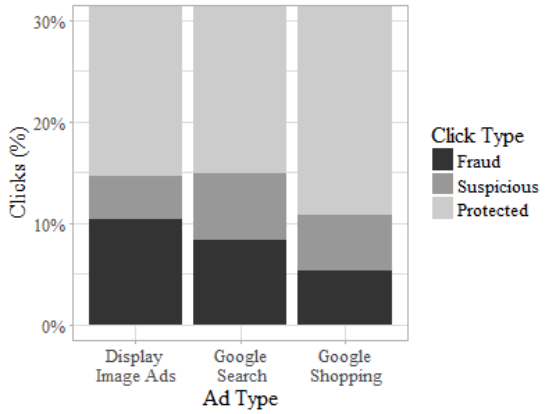


*Fig.8 - Fraudulent clicks as a percentage of all clicks over various advert types.*

Display adverts face the highest rates of fraud of precisely 10%, 142% more of the average rate of fraud (Fig. 8). Google Search comes in second with 7.0%, slightly below average. While Google Shopping has a fraudulent rate of 6%. The finding that Google Shopping is experiencing fraud on a similar level of Google AdWords is surprising since Google Shopping has had very little coverage regarding click fraud, with the majority of coverage given to Google Search and display advertisements.

### Click Fraud and Device Type

The majority of click fraud comes from desktops with a rate of 12% of all clicks being fraudulent (Fig. 9). Clicks from tablets follow closely behind with 10.2%. Mobile devices are found to have a click fraud rate of 7%, which is a much higher rate than expected given click fraud from mobiles is only responsible for 2% of all click fraud (ANA, 2016). From this, it is possible to deduce that the mobile platform is the least likely for 'repeat' fraud. Though this may change in the future with the rising prominence of programmatic mobile click fraud.

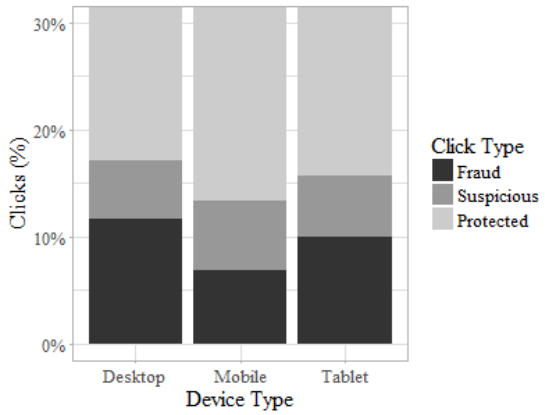


*Fig.9 - Fraudulent clicks as a percentage of all clicks over various device types.*

### Click Fraud and Match Type

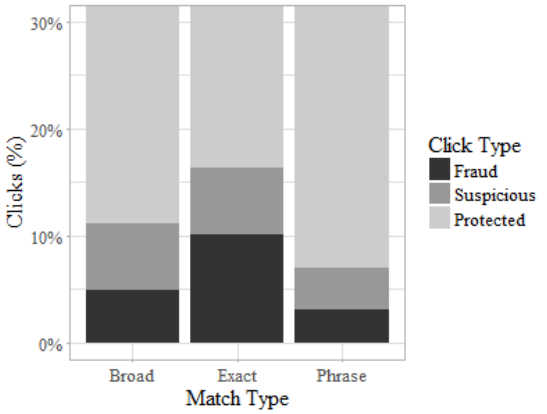


*Fig. 10 - Fraudulent clicks as a percentage of all clicks over search match types.*

Finally, examining search match types, it can be shown that the majority of fraud is committed on exact search terms (Fig. 10). The exact match type has a fraudulent click rate of 10.2%, much higher than average and much higher than broad or phrase matches, with 5% and 3% respectively.

## Conclusion

Click fraud costs advertisers billions of dollars each year, a number that is not expected to fall in the coming years. Understanding how to identify possible fraudulent clicks is paramount to being able to prevent it. Highlighting the key traits of click fraud is an essential step in understanding and detecting click fraud.

In this paper, it is shown through linear regression analysis, that there is no correlation between a keyword's search volume or a keyword's cost per click with the rate of fraudulent clicks. However, we have shown a small correlation between a keyword's competition factor and the rate of fraudulent clicks. High rates of click fraud in the travel and finance industries have been confirmed. The high rates of click fraud have been shown to effect generic TLDs and location-specific TLDs alike. Click fraud on mobile is the lowest rate of the three device types seen, confirming findings by the Association of National Advertisers (Association of National Advertisers, 2016).

Although advancements made in the identification and prevention of click fraud is still set to rise in the future. Hopefully, the information in this paper may aid advertisers in identifying and reducing click fraud.

## References


Wilbur, K. C. & Zhu, Y. (2009). Click Fraud. Marketing Science, 28(2), p.293–308.

Mann, C. C. (2006). How Click Fraud Could Swallow the Internet. [online] Wired. Available at: https://www.wired.com/2006/01/fraud/ [Accessed Apr. 2018]

Li, X., Zeng, D., Liu, Y. and Yang, Y. (2011). Click Fraud and the Adverse Effects of Competition. IEEE Intelligent Systems, 26(6), p.31-39.

Haddadi, H. (2010). Fighting online click-fraud using bluff ads. SIGCOMM Comput. Commun. Rev., 40(2), p. 21-25.

Walgampaya, C., Kantardzic, M. & Yampolskiy, R. (2010). Real Time Click Fraud Prevention using multi-level Data Fusion, World Congress on Engineering and Computer Science, 1.

Saad, S., Traore, I., Ghorbani, A., Sayed, B., Zhao, D., Wei Lu, Felix, J. and Hakimian, P. (2011). Detecting P2P botnets through network behavior analysis and machine learning. Ninth Annual International Conference on Privacy, Security and Trust.

Association of National Advertisers & White Ops. (2016). Bot Baseline 2016-2017. [online] Available at: https://ppcprotect.com/resources/FraudInDigitalAdvertising2016.pdf/ [Accessed Apr. 2018]

White Ops. (2016). The Methbot Operation. [online] White Ops.. Available at: https://ppcprotect.com/resources/WO_Methbot_Operation_WP.pdf/ [Accessed Apr. 2018]

Pixalate. (2017). Programmatic Click Fraud Benchmarks: January – April 2017. [online]. Available at: https://ppcprotect.com/resources/WO_Methbot_Operation_WP.pdf/ [Accessed Apr. 2018]

Eglin, B., Riley, M., Kocieniewski, D. & Brustein, J. (2016). How Much of Your Audience Is Fake? [online] Bloomberg. Available at: https://www.bloomberg.com/features/2015-click-fraud/ [Accessed Apr. 2018]