Revisited: An Exploration of Key Traits of Click Fraud

Andrew Swindlehurst – Data Analyst at PPC Protect

Abstract: Click fraud is an ever-present thorn in the side of advertisers, so understanding click fraud is essential for advertisers to minimise losses. In March 2018, PPC Protect released 'An Exploration of Key Traits of Click Fraud' in the hope of bringing light to key traits of click fraud. The study was limited in some areas of data collection, restricting the information to only clicks protected by the PPC Protect algorithm. In this study, the same key points are covered using both an unprotected data set and a new protected data set. It is found that fraud rates in the unprotected data set are over 400% higher than the protected data set, with rates of fraudulent clicks being 29.75% and 6.97% of all clicks respectively.

Introduction

Click fraud has been a problem since the launch of online advertising (Mann, C. C., 2006), and despite some predictions that the growth rate of online advertising fraud is declining (ANA, 2017), advertisement fraud is still predicted to reach $50 billion by 2025 (WFA, 2017). The presence of advertising fraud has been shown repeatedly and can be substantiated by the recent discovery of '3ve' (security.google.com, 2018).

3ve was a major fraud operation taken down through industry collaboration. The collaboration, led by Google and WhiteOps, resulted in the indictment and arrest of its perpetrators (Justice.gov, 2018). 3ve used malware, spread by drive-by downloads and email attachments, along with Border Gateway Protocol-hijacked IP addresses to control 1.7 million IP addresses and caused $36 million in lost ad spend (CISA, 2018).

In March 2018, PPC Protect released the paper 'An Exploration of Key Traits of Click Fraud' (to be referred to as the 2018 study), intending to increase understanding of click fraud (PPC Protect, 2018). The 2018 study examined the relationship of click fraud with several key factors, including:

- Cost per Click (CPC)
- Keyword Search Volume
- Keyword Competition
- Industry
- Time
- Top Level Domain
- Advertising Network
- Device Type
- Search Match Types

The paper succeeded in highlighting important relationships and has allowed advertisers to better understand which of their Ads may be targeted by fraudsters. However, the paper has a flaw in its data collection. All data used in the 2018 study had the PPC Protect system protecting their advertisements, leading to an obfuscation view of rates of fraud on unprotected advertisements.

The obfuscation actual fraud rates in the 2018 study can be avoided by including recent data collected by PPC Protect. A large enough sample size of 'unprotected' data has been collected allowing direct analysis of fraudulent activity as detected by the PPC Protect algorithm. The objective of this paper is to demystify traits of click fraud using the new 'unprotected' data and recent protected data. The new data allows for a novel comparison between the unprotected data, new protected data and the previously collected protected data used in the 2018 study.

Data Collection

The unprotected data presented in this study is not affected by the PPC Protect algorithm. As it unaffected, this data can be taken at face value. The protected data, however, cannot be taken at face value and requires further clarification.

The protected data presented within this report has the same limitations that were present

within the data in the 2018 study. So it is essential to understand the issue within these sets of data. The PPC Protect system is designed to prevent fraud at the earliest possible time, thus limiting the number of fraudulent clicks that can be collected and used in the data sets. An example scenario is as follows.

A fraudster is planning to click on an advert ten times. After three clicks, this particular fraudster is detected as fraudulent and is blacklisted. Their IP address is added to Google's IP exclusion list automatically and can no longer click on any advertisements for that campaign. In this scenario, three fraudulent clicks would be included in the data set, despite the fraudster intending to fraudulently click ten times.

The unprotected data, when put into the scenario described above, would collect all ten fraudulent clicks, as there is no automated blocking occurring on the unprotected data set. Use of data not protected by the PPC Protect algorithm is the crucial difference between the 2018 study and this paper; it is possible to see the rates of fraud without the PPC Protect algorithm obscuring the actual amount of fraudulent activity.

Results.

The new data set consists of 9.7 million clicks, 8 million protected clicks and 1.7 million unprotected clicks. These clicks can be broken down into three categories, Legitimate, Suspicious and Fraudulent clicks. Legitimate clicks are clicks that have passed through the PPC Protect system and have been determined to be legitimate users, and no action has been taken against these clicks. Suspicious clicks are clicks that exhibit some signs of fraudulent activity, but not enough to confidently identify them as fraudulent clicks. An IP address that commits a suspicious click will be monitored much more closely than a legitimate click. Fraudulent clicks are clicks that have been identified as fraudulent and blocked within the protected data set.

The unprotected data set had a fraudulent click rate of 29.74%, meaning almost 30% of clicks showed a high level of fraudulent activity. This fraudulent activity rate is over 426% higher than the fraudulent activity rate of the protected data set, which sits at only 6.97%. The 2018 study has a comparable average fraud rate to the 2019 protected data set at 7.01% fraudulent clicks.

	Data Set	
Click Type	Protected	Unprotected
Suspicious	5.62%	3.02%
Fraud	6.97%	29.74%

Table 1: Percentage breakdown of fraudulent and suspicious clicks in the protected and unprotected data sets.

Click Fraud Per Weekday

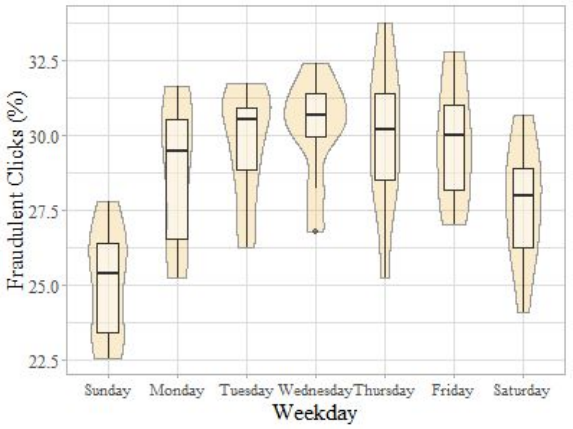

Figure 1: Violin plot displaying unprotected fraudulent clicks as a percentage of all clicks over the days of the week.

Fraud drastically decreases during the weekend, as can be seen in figure 1. This decrease in fraud rates is not mirrored in the protected data set, shown in Figure 2, where the average rate of fraud per weekday is within a standard deviation. The fact that the decrease

in the rate of fraud over the weekend is not matched in both protected and unprotected data would indicate that most programmatic click fraud occurs during the week.

In the 2018 study mid-week days, specifically Tuesday, are mentioned due to their high variation in rates of fraud. This idea is corroborated by the unprotected findings with Thursday having the highest variation with a maximum of 35% and a minimum of 25.7%, an overall difference of 9.3%.

The protected data shows no significant variation between weekdays along. The 2018 study also showed little change in rates of fraud over the week. Compared to the 2018 protected data set, there are considerably more outliers in the 2019 protected data set; this can be credited to the much larger data set.

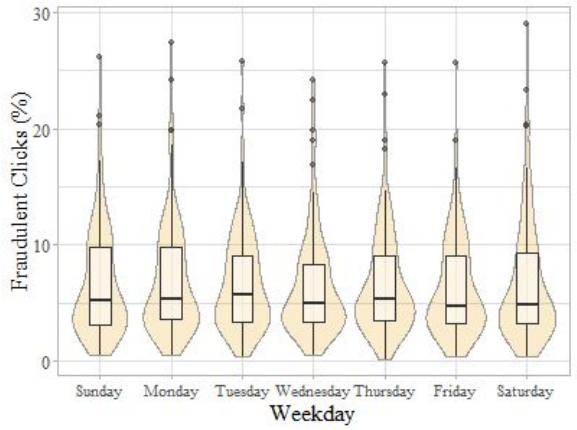

Figure 2: Violin plot displaying protected fraudulent clicks as a percentage of all clicks over the days of the week.

Click Fraud And Cost Per Click

Using an in house tool to harvest cost per click (CPC) values from Google for over 10,000 keywords. Fraud rates were calculated for each keyword and plotted for the unprotected data set (Figure. 3) and the protected data set (Figure. 4).

As with the 2018 study, linear regression analysis was used to find the correlation coefficient, if any, between CPC and the rate of fraudulent clicks. The 2018 study found an insignificant (positive) correlation between CPC. Both the protected and unprotected data sets reflect this, showing no significant correlation between CPC and rates of fraudulent clicks.

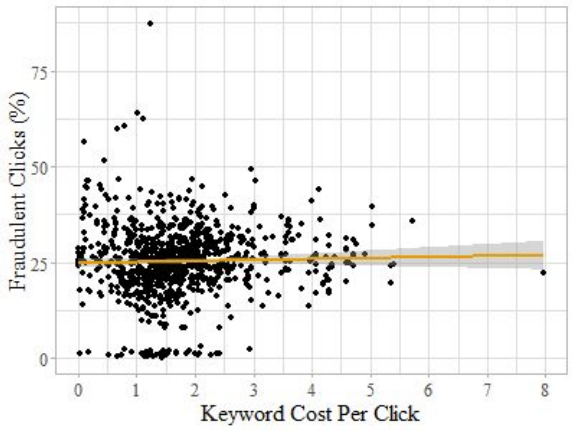

Figure 3: Scatter plot displaying fraudulent clicks as a percentage of all clicks against Cost Per Click with a linear regression line (Unprotected clicks).

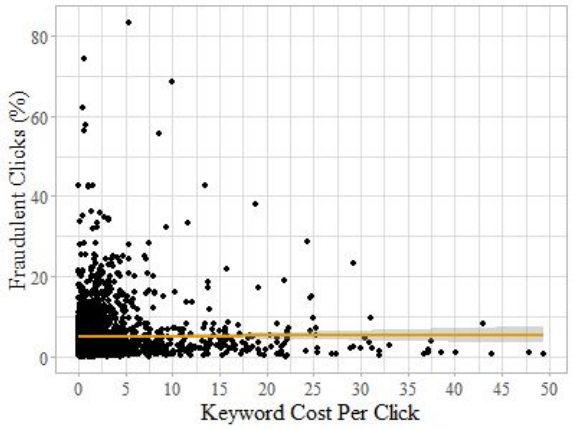

Figure 4: Scatter plot displaying fraudulent clicks as a percentage of all clicks against Cost Per Click with a linear regression line (Protected clicks).

Click Fraud And Search Volume

Using the same tool and steps as the cost per click, search volume data for individual keywords has also been collected. The unprotected data set has an insignificant (positive) correlation between the rate of

fraudulent clicks and keyword search volume (Figure 5). This relationship is unlike the 2018 study and the protected data (Figure 6), where both have an insignificant (negative) correlation.

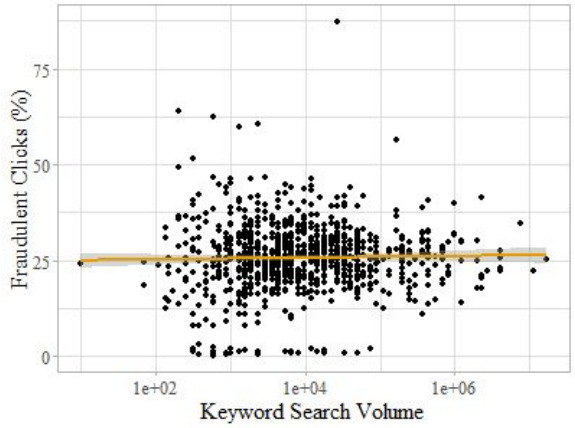

Figure 5: Scatter plot displaying fraudulent clicks as a percentage of all clicks against Search Volume with a linear regression line (Unprotected clicks).

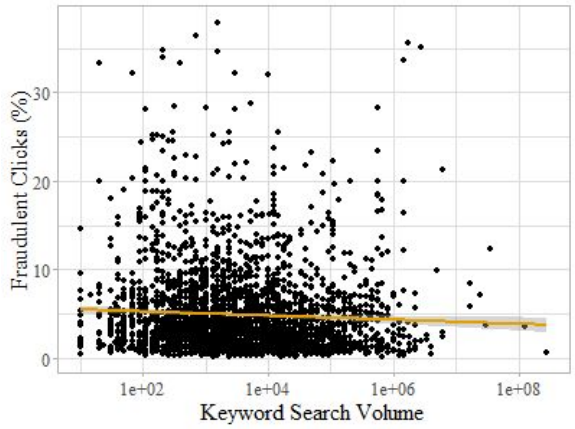

Figure 6: Scatter plot displaying fraudulent clicks as a percentage of all clicks against Search Volume with a linear regression line (Protected clicks).

Click Fraud And Keyword Competition

Data on individual keyword competition was collected with the same tool as search volume and cost per click. Similar to search volume and cost per click data, linear regression analysis shows an insignificant correlation between search volume and rates of fraudulent clicks in the unprotected data (Figure 7) and no relationship in the protected data (Figure 8). The findings are contrary to the results of the 2018 study, as that study found a higher rate of fraud on more competitive keywords. However, the correlation was only a minor correlation, suggesting that there is only an insignificant (or no) correlation between keyword competition and rates of click fraud.

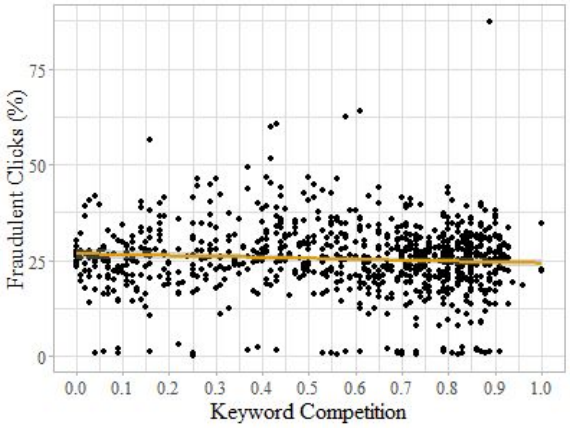

Figure 7: Scatter plot displaying fraudulent clicks as a percentage of all clicks against Search Volume with a linear regression line (Unprotected clicks).

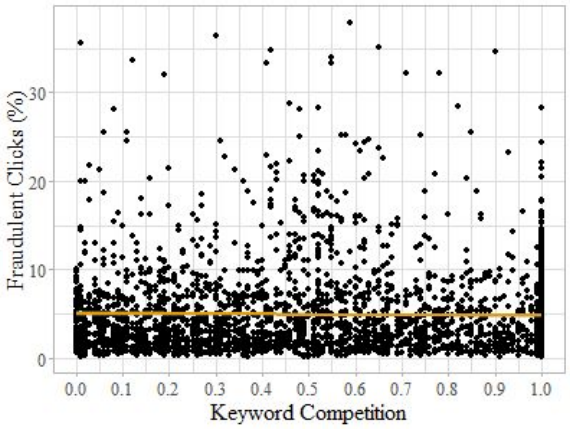

Figure 8: Scatter plot displaying fraudulent clicks as a percentage of all clicks against Search Volume with a linear regression line (Protected clicks).

Click Fraud Per Top Level Domain

Only a limited number of top level domains (TLDs) were available within the unprotected data set (Figure 9). In the unprotected data set, .co.uk experiences slightly above average fraud rates, whereas the .com TLD experiences slightly below average fraud. In the protected data set, .com also experiences slightly below average fraud rates; however, unlike the unprotected data set, .co.uk experiences a rate of fraud considerably below average (Figure 10).

Mirroring the 2018 study, .xyz experiences the highest amount of fraud, closely followed by .net domains. Both of these TLDs experience fraudulent clicks at twice the rate of the norm. Unlike the 2018 study, data for .eu websites were available, which also exhibited much higher than average fraudulent click rate.

Figure 9: Fraudulent clicks as a percentage of all clicks over various top level domains. (Unprotected clicks).

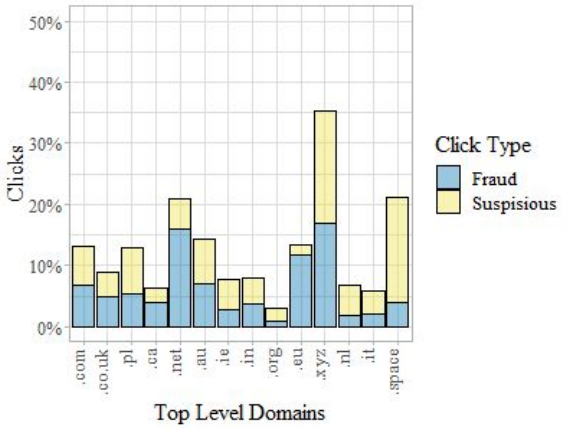

Figure 10: Fraudulent clicks as a percentage of all clicks over various top level domains. (Protected clicks).

Click Fraud And Advert Type

No display adverts were collected in the unprotected data set. However, it can be seen the Google shopping adverts receive much lower rates of fraud than the average (Figure 11). This is consistent in both the 2018 study and the protected data (Figure 12). Google search ads experience slightly under average fraud rates in the protected data set, remaining consistent with the 2018 study.

Display ads in the protected data set show much higher rates of fraud than either search or shopping ads. Display ads are a prime target for fraudsters. Some estimates suggest roughly 60% of all display ad budgets are wasted due to fraudulent clicks. Assuming the difference between the rates of click fraud on the unprotected data set and protected data set are consistent between networks, the 60% estimate would be consistent with our findings.

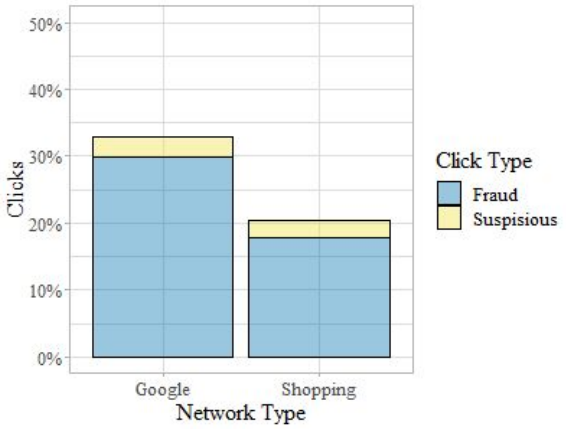

Figure 11: Fraudulent clicks as a percentage of all clicks over Advertising Networks. (Unprotected clicks).

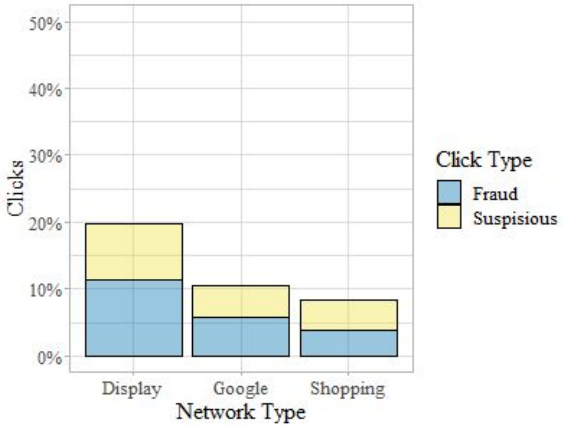

Figure 12: Fraudulent clicks as a percentage of all clicks over Advertising Networks. (Protected clicks).

Click Fraud And Device Type

In the unprotected data set, mobile fraud is considerably above the average rate of fraud (Figure 13). Such rates of mobile fraud are much higher than observed in either the protected data set or the 2018 study. The rate of mobile fraud observed is higher than the 2% of all clicks suggested by ANA (ANA, 2016).

Compared the 2018 study, there has been a considerable reduction of desktop fraud in the protected data set (Figure 14). In the 2018 study, desktop fraud had the highest rates of fraudulent clicks, whereas, in the protected dataset, desktop fraud is slightly below average.

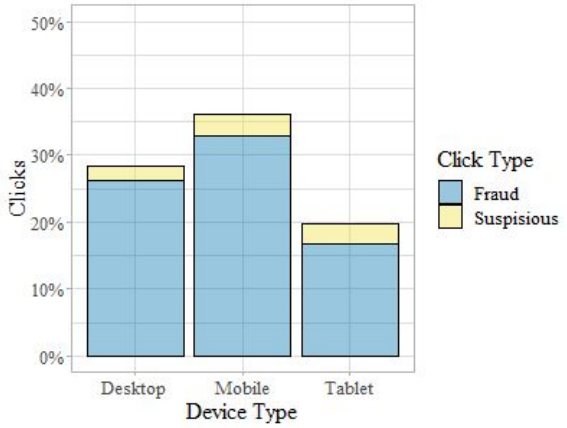

Figure 13: Fraudulent clicks as a percentage of all clicks over device type. (Unprotected clicks).

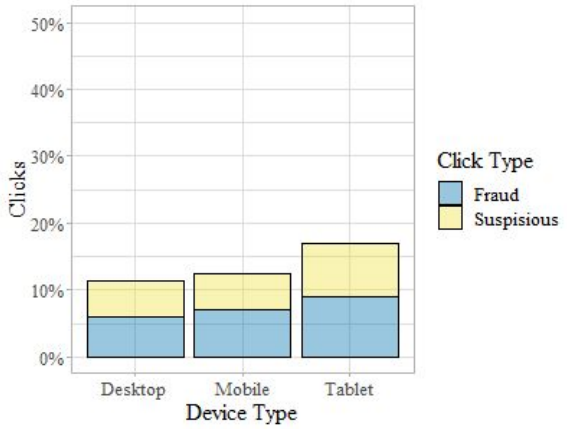

Figure 14: Fraudulent clicks as a percentage of all clicks over device type. (Protected clicks).

Click Fraud And Phrase Match Type

The broad match type experienced a less than average rate of fraud in both the unprotected (Figure 15) and the protected (Figure 16) data sets. The rate of fraud on the broad match type was far closer to the average rate of fraud than in the 2018 study, however. Phrase match, though unavailable in the unprotected data, received the lowest rate of click fraud similar to the 2018 study.

The exact match type experienced the highest rate of fraud in the unprotected data set and marginally the highest rate of fraud in the

unprotected data set. Compared to the 2018 data set, exact received a much closer to average result in the protected data. In the 2018 study, the rate of fraud on the exact match type was much higher than average, whereas in the protected data only minutely above average.

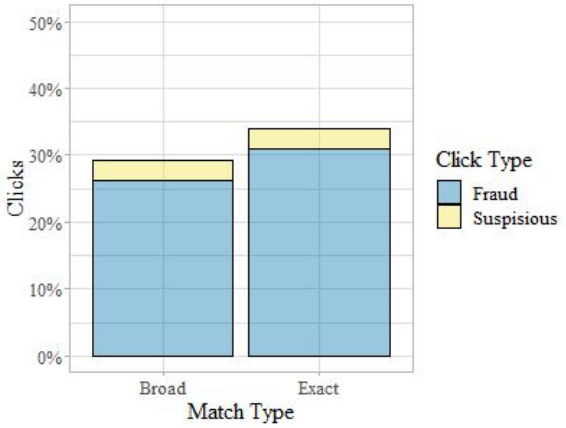

Figure 15: Fraudulent clicks as a percentage of all clicks over phrase match type. (Unprotected clicks).

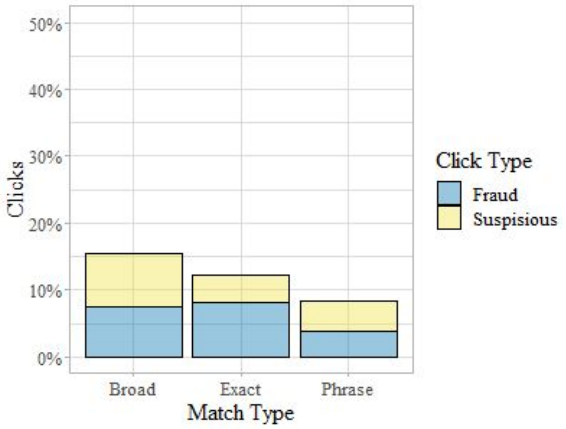

Figure 16: Fraudulent clicks as a percentage of all clicks over phrase match type. (Protected clicks).

Conclusion

With the recent discovery of 3ve, it is clear that programmatic click fraud is a perpetual thorn in the side of advertisers. Understanding identifying features of click fraud is essential to preventing it. The 2018 study made headway to highlighting some of the key traits of click fraud to help advertisers better understand and detect click fraud. However, the 2018 study wasn't completely clear due to limited data collection.

In this paper, two data sets were examined, a protected data set similar to that which was examined in the 2018 study, and an unprotected data set. This unprotected data set is free from the interference of the PPC Protect algorithm and can be interpreted directly. It was found that the rates of click fraud detected in the unprotected data set were significantly higher than the protected data set, an increase of over 400%.

Using linear regression analysis, no correlation between cost per click, keyword competition and search volumes were found in either the protected and unprotected data sets. Well above average rates of mobile fraud were also discovered using the unprotected data, which was missed in the 2018 study. Many patterns observed in this study also confirmed the findings of the 2018 study, such as .net and .xyz being prime targets for click fraud.

This paper has made significant improvements over the previous 2018 study. Hopefully, in clarifying the data, the information contained within this paper will be considerably more useful when trying to aid advertisers in what fraudsters are targeting and how they are doing it.

References

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]

Association of National Advertisers & White Ops. (2016). Bot Baseline 2016-2017. [online] Available at: https://ppcprotect.com/resources/FraudInDigitalAdvertisi ng2016.pdf/ [Accessed Apr. 2018]

World Federation of Advertisers & The Advertising Fraud Council. (2016). Compendium of ad fraud knowledge for media investors. [online] Available at: https://ppcprotect.com/resources/FraudInDigitalAdvertising2016.pdf [Accessed Apr. 2018]

Google (2018). Industry collaboration leads to takedown of the “3ve” ad fraud operation. [online] Available at: https://security.googleblog.com/2018/11/industry-collaboration-leads-to.html [Accessed Apr. 2018]

Google & White Ops (2018).The Hunt for 3ve. [online] Available at: https://services.google.com/fh/files/blogs/3ve_google_whiteops_whitepaper_final_nov_2018.pdf [Accessed Apr. 2018]

United States Attorney’s Office (2018). Two International Cybercriminal Rings Dismantled and Eight Defendants Indicted for Causing Tens of Millions of Dollars in Losses in Digital Advertising Fraud. [online] Available at: https://www.justice.gov/usao-edny/pr/two-international-cybercriminal-rings-dismantled-and-eight-defendants-indicted-causing [Accessed Apr. 2018]

Cybersecurity and Infrastructure Security Agency (2018). 3ve – Major Online Ad Fraud Operation. [online] Available at: https://www.us-cert.gov/ncas/alerts/TA18-331A [Accessed Apr. 2018]

PPC Protect (2018). An Exploration of Key Traits of Click Fraud. [online] Available at: https://ppcprotect.com/resources/key-traits-of-click-fraud.pdf [Accessed Apr. 2018]