A Measurement Study on the Advertisements Displayed to Web Users Coming from the Regular Web and from Tor

A Measurement Study on the Advertisements Displayed to Web Users Coming from the Regular Web and from Tor

Dario Adriano Bermudez Villalva†, Mirco Musolesi†, Gianluca Stringhini‡
†University College London, ‡Boston University
{*dario.bermudez.15,m.musolesi*}*@ucl.ac.uk*, *gian@bu.edu*

***Abstract*—Online advertising is an effective way for businesses to find new customers and expand their reach to a great variety of audiences. Due to the large number of participants interacting in the process, advertising networks act as brokers between website owners and businesses facilitating the display of advertisements. Unfortunately, this system is abused by cybercriminals to perform illegal activities such as malvertising. In this paper, we perform a measurement of malvertising from the user point of view. Our goal is to collect advertisements from a regular Internet connection and using The Onion Router in an attempt to understand whether using different technologies to access the Web could influence the probability of infection. We compare the data from our experiments to find differences in the malvertising activity observed. We show that the level of maliciousness is similar between the two types of accesses. Nevertheless, there are significant differences related to the malicious landing pages delivered in each type of access. Our results provide the research community with insights into how ad traffic is treated depending on the way users access web content.**

***Index Terms*—malvertising, cybercrime, measurement**

1. Introduction

People exposure to media and advertising has changed drastically since the advent of the Internet. As more online content becomes available, many companies seek to make their brands visible to millions of users around the world through online advertising. The inclusion of advertisements (ads) on websites generates a source of income for website administrators. Users interacting with websites are exposed to ads every day which drives an economy of billions of dollars [1]. As any other type of advertising, web-based advertising is a relationship between advertisers and publishers. While advertisers buy space on Web pages to show their ads, publishers get paid to display ads for others on the websites they own [2]. Due to the number of players involved, entities known as advervisting networks (ad networks) manage the buyer/seller process between advertisers and publishers.

Cybercriminals are abusing the online advertising ecosystem to conduct fraudulent activities such as malvertising [3], which is a more powerful mechanism of infection compared to other dissemination strategies because of the implied trust that exists between the parties involved in the ad delivery process. Publishers and advertisers trust the ad networks to deliver only genuine ads. At the same time, users believe that the ads displayed on Web pages are legitimate, therefore they are more likely to interact with them. As a result, miscreants infects a larger number of victims in a short amount of time. Although ad networks spend significant resources to effectively mitigate malicious ads by implementing inspection and monitoring techniques, malvertising is ubiquitous [3]–[7].

At the same time, an increasing number of users concerned about their privacy are migrating to anonymity networks such as The Onion Router (Tor) [8] to access online content. The main feature of these networks is that they provide user privacy by obfuscating the traffic between a client and a website or online service; therefore, the user can access the hosted content anonymously [9]. Previous work showed that a growing number of websites are limiting or rejecting access to users using Tor [10]. This leads us to believe that Tor users may be treated differently by ad networks as well, relegating them to the role of second-class citizens on the Internet. As a result, low quality ads from less popular ad networks might be delivered, putting Tor users at greater risk of being victims of malvertising. In addition, visiting websites from Tor might increase the risk of state sponsored malvertising attacks to users with high sensitivity to surveillance or countries with strong censorship policies in an attempt to deanonymize those users [11].

There is a considerable amount of literature suggesting that attackers utilize the Tor network to commit illegal activities [12], [13]; however, it is not yet known whether such network increases victimization risks to certain attacks. In this paper, we take the initial steps to explore user exposure to malvertising considering the type of network they use to access the Web. To this end, we collected online ads by crawling 20,000 websites from 6 different IPs using a regular connection to the Web (*regular network*) and the *Tor network*. Then, we compared the data to assess the level of maliciousness in the ads displayed. Our results show that ad networks deal with ad requests the same way regardless the type of access and do not discriminate users coming from Tor. Additionally, we observed that even though the level of maliciousness is similar in the regular network and the Tor network, ad servers when accessed from Tor are more likely to deliver ads from low ranked landing pages. This may suggest that ad networks are redirecting Tor users to less reputable ad networks associated with malicious websites.

2. Background and related work

Previous studies have focused on measuring the online ad ecosystem to understand its features. Liu et al. [14]

implemented a browser based tool that provides detailed measurements of the prevalence of different ad targeting strategies. Likewise, Barford et al. [15] developed a web crawler to collect display ads to determine whether the delivered ads depends on the user profile (cookies and browser profile). We have used some strategies presented in these papers to develop our tool to collect ads.

The complexity of the ad ecosystem allows cybercriminals to conduct malicious activities. In our work, we focus on malicious advertising known as malvertising and it is aimed to infect users with malware or redirect them to malicious websites under the control of criminals [4]. To this end, they inject ads into the ad network or set up a shady ad network to deliver malicious ads. Malicious ads take advantage of browser vulnerabilities to infect the victim's machine, lure users to download and install malicious software or redirects users to websites they have not planned to visit [3].

Previous research suggests that malvertising operations are low-cost and highly-effective compared to other malware dissemination techniques such as spamming and social networking [16]. Li et al. [5] investigated the topology of malvertising and proposed MadTracer, a machine learning detection tool that identity prominent features from malicious advertising nodes and their related content delivery paths. Different from our work, they relied on static analysis. Zarras et al. [3] studied maliciousness on ads to which users are exposed using different open tools to collect ads and detect malicious behaviour. Using a similar approach, we developed an ad collection tool from the scratch. Some detection techniques using statistical models have been proposed as well. For instance, Huang et al. [16] applied the Bayesian game model to inspect web-based malvertising. We differ from these studies as we focus on measuring the malvertising ecosystem from the end user perspective considering the type of network used for the access.

Anonymity networks serve an important purpose on the Internet ensuring privacy to users accessing to web resources. An example of such network is Tor. However, an increasing number of websites (publishers or advertisers) discriminate Tor users offering them a degraded service. [10] demonstrate the existence of differential treatment of Tor users by analysing website responses to Tor requests. In our work, we aim to observe whether ad networks are relegating Tor user requests to less reputable ad networks delivering low quality ads.

3. Methodology

Our research focuses on the ad delivery process from the end user's perspective. In this section, we present the methodology we use to generate and analyse a large corpus of advertisements. To this end, we crawl websites using the regular network connection and the Tor network to extract the displayed advertisements served by the ad networks. We then measure and analyse similarities and differences in both access environments.

3.1. Data collection

Collecting data about display ads is a task that requires the ability to identify and analyse the various elements of a webpage, identify ad elements and record the redirections triggered by the ad request [14]. When a user visits a publisher's website, third party tags embedded in the HTML code collect information from different ad networks to serve the ad. Typically, third party tags are enclosed within iframes and executes JavaScript at the time of page load to render the ad dynamically allowing the inclusion of external objects including images, videos, and other HTML documents.

When the third party tag is invoked, a request is made to an ad server which deals with the arbitration process. The ad request is redirected through the different ad networks to find the most suitable ad to be displayed in the publisher website. Ad request redirections are normally implemented through JavaScript and we are unable to record the HTTP chain of redirections taking place internally by the ad servers. However, most of the time, the code retrieved by the execution of the third party tag contains the last chain of the HTTP redirections which consist of an URL that includes the ad server serving the ad and the landing page of the advertiser.

Our approach consists in extracting all the advertising URL chains retrieved by the third party tags when the ad is rendered in the browser. To this end, we developed a Web crawler based on Selenium (a browser automation framework used to test Web applications). The main feature of Selenium is the use of a WebDriver which has the ability to drive a real Web browser natively as a user would. We use Mozilla Firefox for our experiments. This approach allows us to visits websites automatically and retrieve all the content including HTML and JavaScript code rendered, especially for advertisements.

We visit the top 20K websites from the Alexa top 1M list to observe patterns on the most popular Internet websites. Our crawler extracts the source code of the site including the code dynamically generated by JavaScript in the main HTML document and within the iFrames. Depending on the DOM structure of the page, ads are often embedded in nested iFrames spanning multiple levels [14]. We recursively extract the code included on them up to six levels and store it to disk. After storing the code generated for each website, we identify all the advertising URL chains using EasyList, a database of regular expressions that can be used to detect ads [15]. Although all the URLs obtained are ad related, we discard *Single URLs* (e.g. http://domain.com) as they do not represent ad-delivery paths. We focus our analysis on *Compound URLs* which are two or more single URLs linked by specific attributes such as a query string. For example: http://domain.com/path/?query1=test&query2=http://domain2.com. We test the single URLs of the compound URL against the EasyList again to infer ad membership. If is positive, the URL belongs to an ad server (ad URLs). Otherwise, the URL refers to the landing page of the advertiser (landing URLs). In our manual observations, we have noticed that most landing pages come after specific patterns in the URL chain such as `adurl=`, `redirecturl=`, etc.

It is important to note that not all the advertising URL chains included in the advertisements contain the landing page. Sometimes the URL chain consists only of URLs related to the ad servers. We still consider these URLs for our analysis as they represent the ad servers involved in the ad delivery process. As we observed manually in

these cases, when the ad is clicked, further redirections are triggered within the ad network to open the landing page in a different tab or a new window in the browser. Since we focus our study on display ads, we only collect data from automatic redirections without clicking on the ads.

The aim of our work is to measure malvertising from the user's perspective accessing Internet using Tor and a regular network to compare the data we collect. Thus, we deploy our crawler in both access environments using 3 physical machines on 3 different IP addresses. Each physical machine hosts 2 virtual machines (VM). The purpose of the first VM is to crawl the Web using a regular connection and has direct access to the Internet using the IP of the physical machine. The second VM, intended to crawl the Web using Tor, is configured to use a transparent proxy which establishes a connection to the Internet using the Tor network. In total, we have 3 crawlers for the regular network and 3 for Tor, and each one of them visiting the top 20K websites of the Alexa Top 1M list.

To add consistency to our experiment i.e. to ensure that the differences we find in our analysis are due to the access environment and not because other factors, we consider the following controls:

- We visit the same set of websites.
- We synchronise our crawlers to visit each website at the same time and in the same order for each of the 6 VMs.
- We use the same location: United Kingdom. Our physical machines are based in the UK for crawling the regular network. For Tor VMs, we configure them to use a single a UK exit node..
- We remove websites that are not present in all 6 IPs to build an homogeneous dataset for our analysis.

3.2. Detection systems

We are interested in detecting malicious and fraudulent activities that exploit online ads and how they differ between users accessing websites using their regular network connection and the Tor network. Specifically, if any component of the advertising URL chains (ad URL or landing URL) performs malicious activities (e.g., delivering malicious content, illicitly redirecting a user to malicious websites, etc.), we flag the ad as malicious. Correspondingly, we call any chain containing a malicious component a malvertising chain. Note that not all the ad URLs or landing URLs of the chain are always malicious. For example, a malicious component may redirect a user to a legitimate website. There are several blacklists to which URLs or files can be submitted to identify known malicious activity. We use antivirus scans provided by Virus Total and URL/domain blacklist services provided by Google Safe Browsing to classify ads as malicious.

3.2.1. VirusTotal. VirusTotal is an online tool used to analyse files, hashes or URLs to detect malware. It inspects items with over 70 antivirus scanners, URL/domain blacklisting services and other tools to extract different traits from the submitted content [17]. VirusTotal aggregates data from different antivirus engines, website scanners, file and URL analysis tools, and user contributions. VirusTotal also includes several characterization tools used for different purposes such as heuristic engines, known-bad signatures, metadata extraction, identification of malicious signals, etc. We submit our URLs to VirusTotal to identify malicious URLs, extract metadata and obtain domain categories.

3.2.2. Google Safe Browsing. Google Safe Browsing is a blacklist service provided by Google that allows clients applications to check URLs against updated lists of web resources related to malware and phishing and are constantly updated by Google. [18] This detection system aggregates information about maliciousness from various sources including data crawled by Google's search engine robots and client-side checks.Similarly to VirusTotal, we submit our URLs and receive as a response the type of threat related to the resource if the detection result is positive.

4. Results

In this section, we analyse several aspects related to malvertising in the regular network and the Tor network. We deployed our crawler on 20,000 websites from 3 different IPs for each type of access (60,000 in total) and we were able to extract data from 53,946 (90%) sites accessed from the regular network and 45,470 (76%) from Tor. When using the regular network, 9% websites gave 404 not found errors, 0.06% required a captcha, and 0.04% showed an access denied message. From Tor, 22% gave 404 not found errors, 1.7% required a captcha and 0.3% showed an access denied message. The difference on accessibility has to do with the fact a great deal of websites block Tor users [10]. After cleaning our dataset, 13036 websites were available for our analysis for each IP representing 65% of the 20K list. We aggregated the data from each type of access to perform the analysis.

4.1. Measurement dataset

We extracted 11,060 unique advertising URL chains from advertisements in the regular network and 10,041 unique chains from advertisements in Tor. As mentioned previously, a chain may contain 2 or more URLs including ad URLs and a landing URL.

Table 1 shows the distribution of URLs found for both access networks. We include the domains to which those URLs belong to and we call them ad domain and landing domain respectively. It is important to note that several URLs may belong to the same domain but the behaviour for each one is different depending on the redirections performed. As it can be seen, we obtained more advertising URL chains in the regular network compared to Tor. However, the number of URLs and especially domains are similar which may be an indication that the same ad servers and advertisers participate in the arbitration process in both environments.

Table 2 reports the top ten ad servers for each network. As expected, Doubleclick from Google leads the list with about 35% of all the ad requests in both cases. While the top ten ad servers cover approximately 65% of all the ad servers, the remaining servers (440 in the regular network

	Regular	Tor
Advertising chains	11060	10041
Ad URLs	2191	2016
Landing URLs	5639	5447
Ad domains	450	430
Landing domains	4877	4710

TABLE 1: Distribution of the advertising URL chains in the regular network and the Tor network. Chains may contain 2 or more URLs (ad URLs or landing URLs) and each URL belongs to a domain.

and 420 in the Tor network) represent approximately 35% of all the ad traffic generated in our experiments.

Regular		Tor	
Ad network	**Percentage**	**Ad network**	**Percentage**
doubleclick.net	35.60%	doubleclick.net	34.27%
pubmatic.com	6.77%	pubmatic.com	8.99%
adnxs.com	6.10%	adnxs.com	4.64%
mathtag.com	4.21%	tumblr.com	3.62%
openx.net	3.42%	openx.net	3.28%
tumblr.com	3.17%	mathtag.com	3.02%
smartadserver.com	2.09%	smartadserver.com	2.29%
casalemedia.com	1.79%	casalemedia.com	2.07%
weborama.fr	1.38%	de17a.com	1.76%
aolcdn.com	1.32%	weborama.fr	1.66%

TABLE 2: Top 10 ad servers for each web network. Doubleclick is the most common ad server representing one third of all them. 65% of all the ad traffic comes from the top 10 ad servers and the remaining, which are more than 400 ad servers account for about 35% of the total.

4.2. Maliciousness

To identify malicious ads shown to users visiting the websites from the regular network and the Tor network, we submitted our collected URLs using the VirusTotal API and Google Safe Browsing API. If any URL is flagged by either of the two scanners, we assume that it is a malvertising URL since it comes from an ad request and we consider its ad server or landing page as malicious. It is common for VirusTotal vendors to disagree with each other [19] causing false positives. In order to minimise this risk, we sample some URLs to empirically define a threshold. Therefore, we label URLs as malicious only if at least 3 vendors flag them as positive. In addition, we understand that the results of engines may update over time or they need some time to include new information about a URL that has not been previously scanned. Therefore, we performed subsequent scans for our URLs to obtain the most updated results.

As shown in the table 3, for each network, less than 0.5% of the ad URLs involved in the ad delivery process are malicious. Likewise, about 0.3% of the landing URLs where the ads are pointing are infected. In terms of domains, malicious ad domains represents about 1% of all the ad entities observed and approximately 0.3% of the landing domains are flagged as malicious. In general, the level of maliciousness is similar when accessed from the regular network compared to the Tor network. It is important to note that only one ad URL and only one landing URL are detected as malicious for a regular access by Google Safe Browsing. For Tor, a single ad URL was flagged as malicious.

	Regular	% of Total	Tor	% of Total
Ad URLs	10	0.46%	8	0.40%
Landing URLs	18	0.32%	19	0.35%
Ad domains	5	1.11%	5	1,16%
Landing domains	15	0.31%	17	0.36%

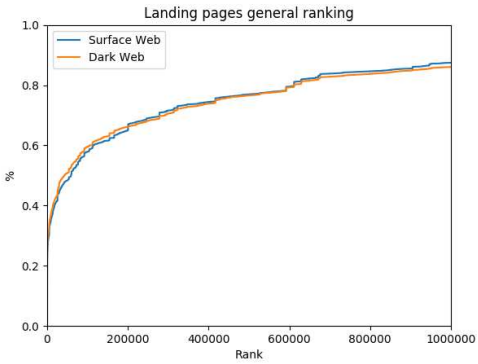
TABLE 3: Level of maliciousness in the regular network and the Tor network. In general, the level of maliciousness is similar in the regular network compared to the Tor.

We further analyse the landing pages resulting from the ad-delivery process to understand the nature of the advertisers involved in the ad ecosystem in the regular network and Tor. Firstly, we focus on the website rank on the aggregated data for each type of access to understand the level of popularity of the landing pages. If a domain is above the one million rank, we extract the rank from the Alexa web server directly. Usually, malicious pages are less popular because they are not indexed to popular search engines. The landing domains observed are in the range between 1 and 2,000,000. Approximately, 7% of them were 'unranked' in both networks i.e. Alexa did not have enough traffic data to assign a rank. Presumably, because these domains were used for a short period of time as part of a malicious campaign or they are very unpopular. Among the unranked domains, about 0.1% are malicious.

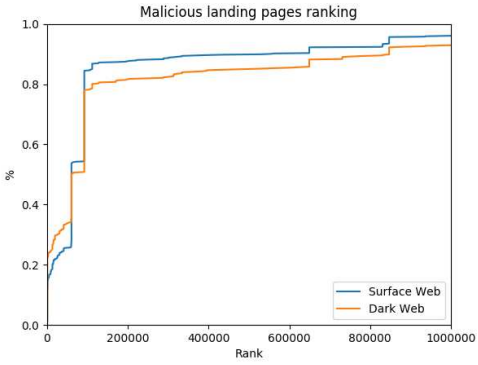
Figure 1a shows the CDF of the ranking of the landing domains observed for each type of access. It shows that the ranks of all the domains involved in the ad ecosystem are similar in the regular network and Tor. About 60% of the landing domains are in the top 100K list and 80% in the top 600K. The similarity in both networks may be an indication that the ad networks deliver ads from high-profile advertisers and do not discriminate traffic coming from the Tor network.

In terms of maliciousness, as can be seen in Figure 1b the rank of malicious landing domains is lower in Tor than the regular network. About 5% of the infected domains in Tor are in the top 10K compared to 50% in the regular network. In other words, most malicious landing pages delivered from ads accessed from Tor are low ranked websites. We used a Kolmogorov-Smirnov Test [20] to determine whether a significant difference exists between the ranking distributions for malicious domains. The results showed that there is a significant difference ($p<.001$) between the two type of access.This suggests that ad traffic from Tor is more likely to be redirected to shady networks which deal with less popular landing pages of dubious reputation.

Furthermore, we identified the ad domains which are unique for each network to verify if certain ad networks only deliver ad traffic in one network or another. We found that there are 35 unique ad domains in the regular network compared to 15 in the Tor network. Each one of them accounts for 0.01% or less of all our observations with the exception of one in the regular network that represents 0.06%. From these unique ad domains in the regular network, only 3 are malicious. No malicious unique ad domains were found in Tor.



(a) General ranking



(b) Malicious ranking

Figure 1: CDFs of the ranking of landing domains. For all the domains found, more that the half (60%) of the landing domains are in the top 100K list and 80% in the top 600K. For malicious domains, the ranking in the regular network is higher than the ranking in the Tor network.

4.3. Domain categorization

We analysed the categories of the landing domains taken from our advertisements to understand the type of websites targeted in the ad-delivery process. We used the VirusTotal API to extract the categories for each landing domain. We obtained 527 unique categories for websites in the regular network and 511 for Tor. Table 4 shows the top 10 categories of the websites pointed by the collected ads, the category of the landing pages is similar for both networks suggesting that ad networks do not target different websites if the ad traffic originates from Tor.

5. Discussion

Our study shows that the ad-delivery process is similar when users visit websites from the regular network and the Tor network in terms of ad traffic. There are small differences in the number of advertising URL chains, ad servers and landing pages in both networks but these are not significant. This may suggest that the volume of traffic related to the ad arbitration process in the regular network resembles the traffic observed in Tor. While it is known that Web traffic coming from Tor is blocked by some publishers [10], ad traffic flows in both networks without restrictions with the same ad networks and landing pages participating in the process most of the time.

Regular		Tor	
Category	**%**	**Category**	**%**
business	21.99%	business	21.78%
shopping	7.80%	shopping	7.99%
uncategorized	7.18%	uncategorized	7.21%
information technology	6.18%	information technology	5.92%
news and media	4.65%	news and media	5.49%
education	4.41%	financial services	4.43%
financial services	4.39%	education	4.14%
travel	3.43%	travel	3.19%
marketing	2.89%	marketing	3.04%
advertisements	2.07%	advertisements	1.99%

TABLE 4: Top 10 categories of the landing pages observed. The landing page categories are similar in the regular network and the Tor network suggesting that ad networks do not target different websites if the ad traffic originates from Tor.

Doubleclick is serving one third of the advertisements in both networks and the top 10 ad networks serve two thirds of them. The list is similar from the two networks with only one different ad server on each side. Thus, the same ad networks and advertisers are present across the different publishers regardless of the network. This support what we mentioned previously that the ad traffic is homogeneous in both networks and originates from the same ad networks. Considering that our advertising URL chains are the result of all the redirections after the ad exchange process, we may argue therefore that ad requests coming from Tor clients are treated in the same way as clients requested from normal Internet connections.

The level of malicious activity is also alike when browsing websites from the regular network and the Tor network, as the percentage of infected domains is similar for ad servers and landing pages. None of the top 10 ad servers, which represents two thirds of all the ad networks, are flagged as malicious. We find that approximately 0.4% of the ad domains and 0.3% of the landing domain are malicious in both Web environments which suggests that cybercriminals perform malicious advertising activities at the same level in the Surface Web and the Dark Web. Presumably because their target criteria is not based in the web environment used to access a website but in other factors.

In terms of the domains of landing pages, the general ranking of the landing pages served through the advertisements in the regular network corresponds to the general ranking in the Tor network. Most landing pages are in the top 100K. Therefore, advertisers involved in the ad-delivery process have a high level of popularity and their ads are served based on the arbitration process performed by the ad network without discrimination traffic from the Tor network. However, focusing our analysis on the ranking based on malicious landing pages, there is a significant difference between both networks. Ranking for malicious landing pages in the Tor network is lower that the ranking on the regular network. This may suggest that ad traffic from Tor users is redirected to less reputable ad networks associated with lower ranking advertisers and based in our previous evidence, it is likely that these advertisers host malicious content.

Categorisation of the landing domains shows some similarities between the regular and the Tor network.

The categories of the websites are the same in both networks with a small variation in the ranking. Therefore, ad networks are not serving ads with certain landing page categories depending on the network. Moreover, there are unique categories in each network but this does not depend on the network since these categories are a small portion of all the observations. The same pattern holds for unique ad servers suggesting that ad networks participate in the ad-delivery process regardless the network.

The comparison shows us that there are not substantial differences in the ad-delivery infrastructure between the regular network and Tor with almost the same ad networks participating in the process with few a exceptions. Similarly, categorization of landing pages follow the same trend in both web environments. Although malicious activity is similar in both environments, malicious landing pages belong to less popular domains probably associated to shady ad networks. Data collected from our observations may be useful for ad network administrators to understand malicious advertising activity in different layers of the web. Therefore, if malicious campaigns are set in any of the networks, they will be differentiated and the respective countermeasures applied.

There are some important limitations in this work. Our crawlers are not able to finish loading the content of the websites at the same time even though we send synchronized requests to load websites. Usually, due to the high latency produced by Tor while bouncing the traffic through several nodes, not all the websites load content at the same time between in both networks. Therefore, our results may be influenced by the time difference in each crawling as advertisements may change in time. Furthermore, we have sampled 20K popular websites but malvertising might be more prevalent in low-ranked websites. Therefore, as part of future work, we plan to crawl different subsets from the top 1M Alexa list to obtain more representative results. Then, we will crawls the same subsets multiple times and average the results to remove some biases such as the latency. At the same time, multiple crawls will allow us to build user profiles and explore whether users are targeted by malvertising campaigns due to their browsing habits depending of the type of access.

6. Conclusion

In this paper, we have explored the advertising ecosystem in the regular network and the Tor network to understand the ad-delivery infrastructure and specifically to compare malicious activity on display advertisements. We crawled websites in both networks to extract advertisements and studied various aspects related to malvertising. We found that there are not big differences in the ad-delivery process and that the level of maliciousness is similar in the regular and the Tor network. However, some ad networks deliver malicious advertisements related to low ranked landing pages. We believe that this work can help to the development of behavioural security systems aimed to detect and prevent malicious advertising campaigns in different layers of the Internet.

References

[1] P. IAB, "Iab internet advertising revenue report 2018 full year results," tech. rep., Market research report, 2019.

[2] B. Stone-Gross, R. Stevens, A. Zarras, R. Kemmerer, C. Kruegel, and G. Vigna, "Understanding fraudulent activities in online ad exchanges," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2011.

[3] A. Zarras, A. Kapravelos, G. Stringhini, T. Holz, C. Kruegel, and G. Vigna, "The dark alleys of madison avenue: Understanding malicious advertisements," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2014.

[4] J. DeBlasio, S. Guha, G. M. Voelker, and A. C. Snoeren, "Exploring the dynamics of search advertiser fraud," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2017.

[5] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang, "Knowing your enemy: Understanding and detecting malicious web advertising," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, (New York, NY, USA), pp. 674–686, ACM, 2012.

[6] X. Xing, W. Meng, B. Lee, U. Weinsberg, A. Sheth, R. Perdisci, and W. Lee, "Understanding malvertising through ad-injecting browser extensions," in *International Conference on World Wide Web (WWW)*, 2015.

[7] S. Ford, M. Cova, C. Kruegel, and G. Vigna, "Analyzing and detecting malicious flash advertisements," in *Annual Computer Security Applications Conference*, 2009.

[8] M. G. Reed, P. F. Syverson, and D. M. Goldschlag, "Anonymous connections and onion routing," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 482–494, May 1998.

[9] E. Marin, A. Diab, and P. Shakarian, "Product offerings in malicious hacker markets," in *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pp. 187–189, Sept 2016.

[10] S. Khattak, D. Fifield, S. Afroz, M. Javed, S. Sundaresan, D. McCoy, V. Paxson, and S. J. Murdoch, "Do you see what I see? differential treatment of anonymous users," in *23rd Annual Network and Distributed System Security Symposium, NDSS 2016, San Diego, California, USA, February 21-24, 2016*, 2016.

[11] S. Khattak, "Characterization of Internet censorship from multiple perspectives," Tech. Rep. UCAM-CL-TR-897, University of Cambridge, Computer Laboratory, Jan. 2017.

[12] D. S. Dolliver and J. L. Kenney, "Characteristics of drug vendors on the tor network: A cryptomarket comparison," *Victims & Offenders*, vol. 11, no. 4, pp. 600–620, 2016.

[13] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, (New York, NY, USA), pp. 213–224, ACM, 2013.

[14] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "Adreveal: Improving transparency into online targeted advertising," in *Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks*, HotNets-XII, (New York, NY, USA), pp. 12:1–12:7, ACM, 2013.

[15] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, (New York, NY, USA), pp. 597–608, ACM, 2014.

[16] C. Huang, M. N. Sakib, C. Kamhoua, K. A. Kwiat, and L. Njilla, "A bayesian game theoretic approach for inspecting web-based malvertising," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–1, 2018.

[17] VirusTotal, "How it works - VirusTotal." https://support.virustotal.com/hc/en-us/articles/115002126889-How-it-works.

[18] Google, "Google Safe Browsing." https://developers.google.com/safe-browsing.

[19] P. Peng, L. Yang, L. Song, and G. Wang, "Opening the blackbox of virustotal: Analyzing online phishing scan engines," in *Proceedings of the Internet Measurement Conference*, IMC 19, (New York, NY, USA), p. 478485, Association for Computing Machinery, 2019.

[20] F. Massey, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.