Can we refine the notion of what an acceptable ad is?

Arjun Gurumurthy
Dept Of Computer Sciences
UW-Madison
Madison,Wisconsin
arjun@cs.wisc.edu

Rogers Jeffrey
Dept Of Computer Sciences
UW-Madison
Madison,Wisconsin
rl@cs.wisc.edu

ABSTRACT

Ad blockers have caused an increasing amount of debate over the recent years, with the scope of arguments ranging from their economic impact to ethical considerations. A majority of ad blockers rely on whitelists and blacklists of websites to control the ads being displayed. Adblock Plus, one of the widely used ad blockers, has enumerated a set of rules that an ad displaying page should follow for an advertisement to be determined as acceptable or not. In this project we investigate the behavior of adblock plus to .Our study begins with the analysis of the ad-block plus filtering-mechanisms. Existing addons provide us with capability to understand the filtering process. We then analyze the behavior of adblock plus on 1500 webpages to see how adblock plus implements or enforces the acceptable ads criteria. We use our analysis to see if minor modifications to the page structure can cause the ads to pass through the filtering process of the adblocker. Our results show that adblock plus is too restrictive as it blocks non advertising related content. We also observe that adblock plus can be obviated by minor modifications to the web pages.

Keywords

Ad Fraud,Impression Fraud,Acceptable Ads

1. INTRODUCTION

Content and services which are offered for free on the Internet are sometimes monetized through online advertisements. This relies on the implicit understanding that the viewers who consume free content do so at the cost of viewing ads. However, recent times have seen the rise of ad blockers that control the display of advertisements on websites. These adblockers rely on a list of filters for blocking or allowing advertisements.The lists can be categorized into whitelists and blacklists. Blacklists contain rules for blocking advertisements whereas whitelist acts as exception for the rules that are in the blacklists. In 2011 adblock plus started monetization of the whitelist by launching what is known was an "Acceptable ads" program.As a part of the acceptable ads program adblock plus proposed that ads displayed on a page have to adhere to. A publisher adhering to this crtieria can pay a fee to adblock plus and get themselves whitelisted. It is known that the e-commerce and advertising giants like Amazon, Google and Microsoft paid an undisclosed sum to adblock plus to get themselves whitelisted.

While the number of users of adblock plus continues to grow, on the other hand the number of filters added to the whitelist has also grown over the years. <Insert fig>, indicates the growth of whitelist since 2011. Unsurprisingly, one can see an increasing trend in the number of ads allowed by adblock plus. This raises the following questions.

Q1. Does Adblock Plus allow ads even if they do not satisfy the criteria of an acceptable ad?

Q2. Does adblock Plus block ads even if they satisfy the four criteria?

Q3. As more sites are added to the whitelist when can adblock plus be deemed in-effective?

We attempt to answer Q1 and Q2 in our project by performing crawling web pages from different sources and analyzing the behavior of adblock plus on these webpages.

2. BACKGROUND

Adblock plus is one of the widely use adblockers. <statistics> Adblock plus relies on filter lists for its functioning. The filter lists can be categorized into whitelists and blacklists. Blacklists contain rules for blocking advertisements whereas whitelists acts as exception for the rules that are in the blacklists. The rules are CSS and Xpath selectors that are prefixed with a predefined set of symbols. These rules interpreted as regular expressions by the adblock plus engine. The notation is that filters prefixed with an "@@" is a whitelisting filter and filters prefixed with "##" or #@ is a blocking or blacklisting filter. The blocking filters can be broadly classified into:

1. Element Hiding filters
2. URL filters

Element hiding filters hide the elements present in an webpage from the user whereas the URL filters block the web page from requesting a resource , be it a webpage or javascript. The filters are published and maintained by easylist [3].

3. *RELATED WORK*

General aspects of adblocking have been discussed in various studies. These studies have focussed on economic impact that adblockers have on advertising. [2] suggests that adblock railroads publishers and content providers to pay for their sites to be whitelisted. <report> suggest that in their report estimate the number of adblockusers at 2014 to be 144 million. They also [5] describe a method of classifying ad traffic by leveraging adblock plus. They use libadblock plus to classify the web traffic based on hits from the easylist and the whitelist. They also note that it is not possible to associate HTTP traffic with ad objects and that information is needed about the structure of the web page to achieve higher accuracy in detection of ad traffic. Finally, they establish that it is not possible to identify hidden ads without knowledge of the html page content. By aiming to look at whether advertisements are acceptable or not, we are not just dealing with fraudulent ads that are propagated across the Internet, but also ads that a user would perceive as annoying or intrusive. In our process of understaning what an acceptable ad is, we look into existing definitions of what it means to be an acceptable advertisement. [7] describe criteria for an acceptable ad as defined by Adblock plus:

- Advertisements cannot contain animations, sounds, or "attention grabbing" images.
- Advertisements cannot obscure page content or obstruct reading flow, iė; the ad cannot be placed in the middle of a block of text.
- Advertisements must be clearly distinguished from the page content and must be labeled using the word "advertisement" or equivalent terms.
- Banner advertisements should not force the user to scroll down to view page content. [7] further present a comprehensive study of adblock plus and an analysis of how the users perceive acceptable ads. They report that 90% of the users viewing an all grid layout ads could not distinguish them from the content. Allowing this ads thus seems to be in conflict with adblocks acceptable ads policies.

Furthermore , in their analysis of PPV networks [6] describe a viewport size filter mechanism (a viewport is the user's visible area of a web page) for detecting ad views that are too small to be seen by the user.

It can be clearly seen that the html content of the web page has a ton of information that can be used to characterize the nature of the ad being displayed in the web page. We thus concentrate our discussion on analysis of web page structure to detect and redefine unacceptable ads.

4. METHODOLOGY

THe first step in our study was to gain insight into the working of adblock plus and the different filters it uses. [7] explain in detail about filter specification and the functionality of different kinds of filters. Additionally, we used a tool called the Element Hiding helper [4] to identify the correspondence between the filter and the actual web page elements displayed. Element Hiding Filter Helper a Firefox add-on that aids in analyzing the ads that were blocked. It identifies and highlights an element blocked by adblock plus, provides the content type of the element and the maching URL or element hiding filter that caused the element to be blocked.

We then proceeded to examine and measure the impact of AdBlock Plus, through three different experiments:

(a) The impact of whitelists on AdBlock Plus's blocking mechanisms

(b) If we do away with the notion of lists, can we see how AdBlock Plus performs with respect to element hiding filters alone.

(c) Fuzzing the page to prevent AdBlock Plus from recognizing the ads in the page.

4.1 Crawling

For our experiments we needed to examine a corpus of webpages with advertisements. Crawling of webpages is the standard way of obtaing the content of the webpages. However normal crawling of webpages would just fetch the source of the page offline and there by mask the additional communication made between the advertisement and the adserver. This might result in elements missing from the page source which would have otherwise been present if the page were online. This calls for a crawling mechanism which mimics the behavior of an actual browser. The crawler should also provide us with the capability of analyzing the DOM(Document Object Model) of the page.

AdBlock Plus has developed a tool called abpcrawler [1] that launches a webpage in realtime. This tool takes in as input a list of urls to crawl. The tool launches the URLs in the browser and logs the filters,both blacklist and whitelist that had matches in the pages.

This tool uses the Gecko layout engine of Firefox to launch the browser with a pre-specified url. This tool also provides the source of a webpage in xml and also returns the visual representation of the webpage as an image. Additionally the tool also logs the adblock plus filters applied to the page as a json. This allows us analyze the type of filters fired for a given webpage.

4.2 Combination of Filter Lists

We also needed capabilities to specify the filter lists used by adblock plus when it examins a webpage. AdBlock Plus provides two means of configuring the whitelist and blacklists through it's GUI. Users could add subscription to the filters they wanted by specifying an URL to the filter-file. Additionally, an option is provided to the users for configuring adblock plus to use the whitelist, i.e acceptable ads. This option is enabled by default.

The abp crawler,however, did not have means through which filter lists could be configured. We created Firefox profiles for this purpose, one for a particular combination of filter lists used in our experiments. Adblock plus was installed and the necessary filter lists were configured for each of the profiles. The abpcrawler was enhanced to use the profile information. This enhanced abpcrawler working is outlined in the following steps:

Thus for our experiments, the configuration in step 2 alone needs to be changed. For example, to see the impact of the change in whitelist (a), the corresponding whitelist URL is configured, and the crawler is run.

Measuring performance of ads: One key aspect to measure is to see the impact of URL filters on AdBlock Plus's

Algorithm 1 Crawler

```
1: procedure CRAWL WEB PAGES
2:     configurations ← various filter configurations
3:     urls ← urls to crawl
4:     for all urls do
5:         for all configurations do
6:             Launch firefox with configuration
7:             In JSON format, write to a file the following
8:             The content type of the element
9:             The filter regex applied (null if no filter is applied)
10:            The location of the element within the page.
       return
```

performance. We expose a new URL that has only the element hiding filters (a subset of easy_list.txt) and configure it to run the crawler. We can now test the performance of AdBlock Plus with three configurations - the âĂŸnormalâĂŹ AdBlock Plus (the default typical configuration, with Acceptable Ads enabled), EasyList alone (Acceptable Ads disabled) and Element Hiding (only element hiding filters), and measure the number of ads blocked. A naive measurement of the number of ads blocked is to count the number of non-null and non-allowed filters in the JSON file. This doesnâĂŹt always work - with whitelist and blacklist, the number of ads blocked shows a higher number than with only the blacklist. The reason for this is that when URLs are blocked, the call never gets through and hence the number of elements in the page is considerably less. This would mean that even though more elements are explicitly blocked, the actual number of blocked elements is higher. To get the actual measurement, we have a new measure, the total number of elements in the page.

The procedure is as follows Get the total number of elements in the page For each configuration, Blocked elements in page = Get the count of non-null and non-allowed filters in the page. Effectively blocked elements = Total number of elements (1) - Element count in page for current configuration Total blocked elements = Blocked elements in page+Effectively blocked elements.

This is then repeated for multiple configurations.

In addition, to check whether it is possible to circumvent the ads, we do the following: Circumventing element hiding filters: AdBlock Plus blocks ads by looking into the blacklist and whitelist and hiding elements that are prefixed with the element hiding filter tag. This means that the blocking mechanism can be considered as a form of string matching process. Any way to circumvent the string matching process would then allow the ad to be displayed. One method we followed was to use a URL redirection filter - this would result in the element to not be blocked as the string would not be matched. Blocking of Non-Ad elements One other experiment would be to check if elements that are not ads are blocked by AdBlock Plus

5. CONCLUSION AND FUTURE WORK

In our analysis, we observed that WideTable outperforms other views considered in the study. The better performance was unaffected by the multiple memory configurations we tested in our analysis.

A major takeaway in our analysis is that the cost of workload is affected significantly by the disk access patterns. Therefore the cost formulation should incorporate not just the buffer pool memory but also query characteristics such as types of joins, I/O access patterns, disk bandwidth etc. A sophisticated cost model involving the above parameters can help us in a more accurate modeling of the workload.

We also feel that SSB is a simple benchmark which doesn't offer variety in terms of relational operations performed in the query. A complex benchmark like TPC-H should offer more challenges in analysis.

Lastly, the LP formulation should allow selection of multiple views, instead of one view per workload, as it is assumed right now. This will allow us to choose the best view for a given query and potentially maximizing the performance gain for the workload. It will also make the LP not so obvious, unlike the current LP which can be solved using a simple greedy algorithm.

6. REFERENCES

[1] A. Crawler. ABP Crawler. https://github.com/adblockplus/abpcrawler.

[2] DigitalTrends. Adblock plus accused of shaking down websites., 2014. http://www.digitaltrends.com/web/adblock-plus-accused-of-shaking-down-websites/.

[3] EasyList. EasyList., 2016. https://easylist.github.io/.

[4] E. H. Helper. Element Hide Helper. https://adblockplus.org/elemhidehelper.

[5] E. Pujol, O. Hohlfeld, and A. Feldmann. Annoyed users: Ads and ad-block usage in the wild. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 93–106. ACM, 2015.

[6] K. Springborn and P. Barford. Impression fraud in on-line advertising via pay-per-view networks. In *USENIX Security*, pages 211–226, 2013.

[7] R. J. Walls, E. D. Kilmer, N. Lageman, and P. D. McDaniel. Measuring the impact and perception of acceptable advertisements. In *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, pages 107–120. ACM, 2015.