

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 ad-blocker. 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 as 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 criteria 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 four criteria ?
- 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 experiments on various websites.

1.1 Proposed Approach

In this project we propose to analyze, what constitutes an acceptable advertisement. We plan to crawl web pages from different sources and analyze them. Additionally, we plan to investigate machine learning as a tool that would aid us in the analysis of the web pages and aid in discriminating between acceptable and unacceptable ads. In particular, we attempt to see if we can discover any patterns from unacceptable ads and see if those patterns could be used to refine the criteria for acceptable ads.

2. RELATED WORK

General aspects of adblocking have been discussed in various studies. These studies have focussed on economic impact that adblockers have on advertising. [3] 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 understanding 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, i.e; 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.

3. METHODOLOGY

We give an outline of the questions that we attempt to answer, and then elaborate:

- Q1. How to detect if an ad violates the current characteristics of an acceptable ad?
- Q2. How do we classify if an ad is intrusive or not?
- Q3. Can an ad be annoying or intrusive even after satisfying all these four criteria?
- Q4. Also, based on classification of ads as annoying or intrusive, can we then detect some common patterns in the web pages that are deemed non-acceptable?
- Q5. If so, can this be used to refine the notion of an acceptable ad?

Note that we term 'unacceptable' ads as ads that don't follow the criteria for acceptable ads as well as ads that could be considered annoying or intrusive even if they satisfy the 4 characteristics.

1. One way to answer Q1 is to look at the working of existing ad blockers and see if they are able to perfectly predict whether an ad violates the current notions of acceptability.

We start by looking at Adblock Plus - how does it deal with acceptable ads. All the current criteria for acceptable ads involve the page structure, and this implies that Adblock Plus must have some way of analyzing the DOM structure of the web page content. Thus, a critical step is the analysis of the Adblock Plus code. We then check if:

- (a) Does adblock Plus block ads even if they satisfy the four criteria?
- (b) Does Adblock Plus allow ads even if they do not satisfy the four criteria ?

While ideally (a) or (b) should not happen, there might still be some exceptions. Such exceptions provide scope for improvement or refinement of the criteria used to identify acceptable ads.

2. The next step is to obtain the data - this can be done by crawling the pages that contain annoying or intrusive ads and then storing the web page contents, and the same for the pages without these ads. For obtaining the data, we crawl through the ad servers domain list that is obtainable from lists [1] [2] [4] A question arises here: How to obtain pages with only acceptable ads? For now, a simple way to look at acceptable ads is to consider that all ads blocked by the Adblock-Plus plugin as unacceptable, and the ones permitted through as acceptable. It needs to be thought of if there is a better approach.
3. We now have a set of pages, some with acceptable ads and some without. We attempt to perform the following:
 - (a) We then identify characteristics of acceptable ads or unacceptable ones - there are multiple ways to evaluate this: one to Looking at the DOM structure and identify if there are common patterns between ads that are acceptable, one to crowdsourcing and obtain users' subjective analysis of what constitutes an acceptable ad, the other is to try to have an image of pages with acceptable and unacceptable ones and see if they have some visual similarities, or match some aesthetic criteria. we go with analysis of the dom structure, as it is an objective measure.
 - (b) Given a new ad, can we then classify it as an acceptable or not? One way to do that is if we have a set of annoying and acceptable ads, given a new ad - can we classify it as acceptable or not. This looks to be a typical machine learning classification problem provided we identify and specify the relevant features in the web page structure. This can be used to answer Q2.
4. If possible, we check if we can refine the notion of an acceptable ad or not by the characteristics or modes of classification used in the previous step. That is, if we identify structural similarity with respect to acceptability, can this be exhibited as a further constraint to

redefine what an acceptable ad is? This would help us answer Q3 - Q5.

One example of a criteria to determine if an ad is acceptable or not: frequency of occurrence of the ad. If the same ad is loaded every time a set of pages is viewed, will it correlate with the ad being unacceptable?

This whole process would have implications in ad fraud detection as well, as analysis of web pages would be helpful in determining ad fraud, assuming the correlation between unacceptable ads and fraudulent ads is not very low.

4. 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.

5. REFERENCES

- [1] C. Crawl. Common Crawl, 2011.
<https://commoncrawl.org/>.
- [2] J. Crussell, R. Stevens, and H. Chen. Madfraud: Investigating ad fraud in android applications. In *Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services*, pages 123–134. ACM, 2014.
- [3] DigitalTrends. Adblock plus accused of shaking down websites., 2014. <http://www.digitaltrends.com/web/adblock-plus-accused-of-shaking-down-websites/>.
- [4] P. Lowe. Peter Lowe. Ad blocking with ad server hostnames and IP addresses., 2013.
<https://blog.pagefair.com/2015/google-losing-billions-adblock-devils-deal/>.
- [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.