Dear Editor,

As you kindly suggested us, we have made some changes in the paper to address the reviewer's comments. Next, we detail the specific changes we have made to the paper in order to address every issue.

We hope that the new version of the paper meets the quality requirements for being published in the Computer Networks journal.

Please, transmit our sincere thanks to the anonymous reviewers for their help in improving this paper.

Sincerely yours,

Gabriel Maciá-Fernández
Corresponding author

Reviewer #1:

The authors investigate the problem of identifying web browsing patterns without inspecting packet payloads. The problem is important and challenging; and solving this problem allows service providers to facilitate targeted advertisements.
The paper has the following strengths:
(1) The paper is well-written,
(2) The proposed algorithm is clear
(3) The authors thoroughly investigate many details and challenges in recovering web browsing patterns
(4) The performance evaluation section, while relying on relatively limited data sets, highlights the accuracy of the proposed algorithm

The following point needs further elaboration:
As admitted by the authors, the proposed algorithm requires extensive web crawling. It is worth adding a section or at least a paragraph quantifying the involved overhead and how it scales with the number of investigated web pages. The diversity in the popularity of web pages may be included in this discussion to bound the overhead.

> *The method suggested in this paper focus on the detection process once the crawling has been carried out. It is true that assuming a comprehensive and up-to-date crawling is not realistic. For this reason, the strategy in this paper is to consider that a limited crawling will be done, and see how it affects to the suggested detection process. Specifically, Section 4.4 explores the functioning of this process when either traces or crawling information are not up-to-date, and Section 4.6 explores the situation in which only a partial profile is obtained from a website.*
>
> *We agree with reviewer 1 that quantifying the overhead of the crawling process and evaluating how it scales is a really interesting issue in this context. Many papers have focused on efficient crawling indeed. Yet, we still think that the scope of this paper is the evaluation of the detection method and not the crawling process. However, we*

agree with reviewer 1 that there are some hints that should be given for the crawling process to be efficiently adapted to our detection method. For this reason, we have extended Section 7 (Discussion – Too much crawling?) to include these issues. Moreover, we consider that it is a very interesting suggestion for building an efficient crawler specially designed for our methodology. We have also included this point in the conclusions section as future work.

Reviewer #2:

The paper addresses the relevant problem of non-integrity violating
tracing of user web behaviour.

While the topic of the paper is interesting, it is severely hampered
by the presentation. The paper is technically acceptably sound, but
the writing is very opaque, making the paper very heavy
reading. Unnessecarily so. Furthermore, the nomenclature is
unconventional and confusing, in particular in regards to the notions
of root files, objects, slices and elements. An example of
this issue is the definition of "root file", which is inconclusive as
well as circular.

We have followed the suggestions of reviewer 2 and, thus, we have adopted the terminology for a web page structure as given in this seminal paper [21]. Coherently, the notion "root file" has been replaced by "page file", and the term "objects" has been maintained. Regarding page files and objects located in external servers, we adopt the terminology "third-party" taken from [23]. We explicitly cite these references in the text to clarify the readers about this point. In addition, the text has been modified accordingly (appearances of "root file" → "page file", and R set (root files set) in 3.2.3 has been renamed to PF set (page files set)).

Regarding the terms "slices" and "elements", we have not found any other paper referring to the same concepts, so we have chosen to clarify them by inserting Figure 2 and modifying text in 3.2.2 ("web page-based trace slicing" and "extracting web page features") in order to better clarify this terminology.

The methodological descriptions as well as experimental descriptions
would benefit from diagrams to lighten the density of the text.

The densest parts of the paper have been extended by adding diagrams and figures. Specifically, (new) Figures 2(a) and 2(b) explain the model of the trace and the way in which useful information is extracted from it. In addition, (new) Figure 3 gives an overview of the proposed detection algorithm.

Reviewer #3:

The paper proposes a scheme to build web access history from TCP headers, instead of examining the contents of the packets. The motivation is to allow ISPs to tap into advertising markets with violating current laws. This is a very interesting issue, especially because such header information is easy and allowed to share among different parties. Although such a scheme is somehow violating users' privacy, and may be soon targeted by legal actions.

The authors are well aware of related work and present a clear design how to take advantage of statistical learning to reconstruct access patterns. They have shown good understanding of current web trace analysis and developed fairly reliable solutions.

However, although this paper has good practical values, it does not provide too much contribution in the theoretical aspect. How to model the profile of a web site and how to control the false rate are more interesting theoretical questions. This fails to discuss such issues with broad impacts.

Again, it is a solid work. I support to accept it. However, I wish it could lead to a more theoretical study which would be more interesting.

> *We agree with this reviewer that a theoretical study about the model profile and about false detection rate yielded by our methodology is an interesting contribution.*
>
> *In our paper, we present a model for the profile of a web site, which is composed of different pages, each one represented by several attributes (size, location, links, etc.) within which the pages contents are not considered as in other approaches (e.g. for content-based advertising it is important the contents of visited web pages, so this information is retrieved in the crawling process).*
>
> *Regarding the development of a theoretical framework which allows the estimation of false rates (positives and negatives), there are many issues involved. First, as pointed out in our paper, the depth of the crawling in a certain web site affects the detection rate. Second, sources of noise (caching, compression, cookies, embedded dynamic code, variable headers sizes, etc.) should also be modeled in this framework, which is not straightforward. While the first point could be modeled by taking into account a probabilistic approach considering the probability for a page to be visited within a web-site, the modeling of sources of noise is not straightforward at all. For this reason, we consider that this problem deserves itself a deep study and decided not to afford it in this paper.*
>
> *As we agree with the reviewer at this point, we have considered it useful to include these ideas as future work in Section 8.*