POLITECNICO DI TORINO
Repository ISTITUZIONALE

Data Analysis and Modelling of Users' Behavior on the Web

Original
Data Analysis and Modelling of Users' Behavior on the Web / Vassio, L., Mellia, M.. - STAMPA. - (2019), pp. 665-670. (2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM) Arlington, VA (USA) Aprile 2019).

Availability:
This version is available at: 11583/2800194 since: 2020-03-16T13:08:33Z

Publisher:
IEEE

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

19 June 2026

Data Analysis and Modelling of Users' Behavior on the Web

Luca Vassio
Politecnico di Torino, Italy
luca.vassio@polito.it

Marco Mellia
Politecnico di Torino, Italy
marco.mellia@polito.it

***Abstract*—The research developed during my PhD [1] was driven by the need to understand how people interact with the web. This information gives ISPs and network managers better visibility and understanding of how users and web services change over time. Thanks to traces and logs of users' traffic, my work focuses on two complementary aspects: (i) data analytics, and (ii) user modelling.**

In this work, I show how to reconstruct users' online activity from passive measurements and to model their behaviour. I introduce machine learning approaches to identify the intentionally visited web-pages and web-sites. I highlight device usage evolution, the structure of the navigation and the interactions with social networks and search engines. I build users' profiles and then I show how to re-identify users in a future time thanks to their behavioural fingerprints. This is also instrumental for security applications. I next study the interaction with online ads, capturing the impact of the temporal dynamics of shown advertisement and improving revenues.

I make available all the anonymized datasets and code for the community, to guarantee results reproducibility and foster further analyses.

***Index Terms*—data analytics, modelling, passive traces, network monitoring, machine learning, human behaviour, fingerprinting, recommendation systems.**

I. Introduction

The Internet and its pervasive use transformed our approach with the world. The research developed during my PhD [1] was driven by the need to understand how people interact with the web, capturing its characteristics and changes, and modeling people inner habits and interactions. Traces and logs of users' behaviors collected in the Internet (i.e., passive measurements) offer invaluable information to obtain this goal. Thanks to passive traces, I study the behavior of the users, with focus on two complementary aspects: (i) data analytics, and (ii) user modeling.

There are many key challenges to face: (big) data requires the use of scalable software and hardware. It demands also the introduction of innovative methodologies and meaningful metric to obtain trustable, filtered, clean and useful information. Data analytics is performed by means of a variety of statistical, machine learning and data mining approaches. Moreover, it is also a pre-requisite for creating analytic models of the studied phenomena, that should be as much as possible adherent to the reality. Lastly, understanding the applicability of derived models is a fundamental step for optimizing performances and understanding possible scenarios.

More in details, I analyze 3 years of data of about 25 000 households, reconstructing and analyzing users' online activity. In summary, the followings are the main contributions to the research community, obtained thanks to my thesis [1]. All these results are also instrumental for ISPs and network managers who can get a better understanding of what people do online.

- I propose a new machine learning approach for the identification of web-pages and web-sites explicitly visited by users in HTTP and TCP logs, collected by passive network monitors ([2], [3], presented in Section II).
- I present a characterization of clickstreams that differs from previous efforts (e.g., [4]) for (i) covering a large population during a long period of time, and (ii) accounting for different devices used to browse the web. Thanks to this, I am able to highlight device usage evolution, the intrinsic structure of the navigation and the interactions with social networks and search engines ([5], presented in Section III-A).
- I model paths of users on the web, representing them in a succinct and interpretable manner. I can easily and automatically inspect and cluster the interests of users and communities. I can automatically extract groups of similar or likely connected web-sites, and monitor the interests and browsing patterns of a single user or communities ([6], presented in Section III-B).
- I explore techniques for users' fingerprinting and identification, using only the domains of visited web-services. This has several implications to cybersecurity and privacy ([3], presented in Section IV).
- I model the user interaction with online ads, introducing a behavioral model validated and tuned on real traces. I improve the revenue of advertisements systems by optimizing the timings when ads are shown to users ([7], presented in Section V).

Following the scientific approach, I made available the anonymized datasets that I used for the community.

II. Dataset and methodologies

A. Dataset collection

The starting points are passive traces, often consisting in raw data, automatically saved in different kind of logs. I rely on Tstat [8] to collect passive data. Tstat is a deep packet

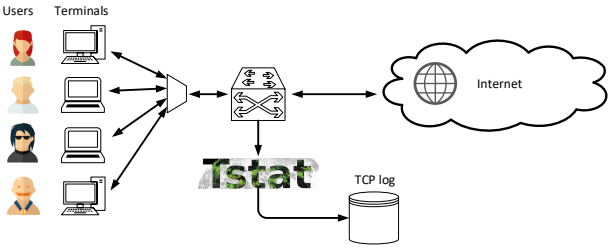

Fig. 1. Tstat is installed at a PoP and monitors the network, logging information from both TCP and HTTP connections.

inspection tool for network monitoring that logs information from both TCP and HTTP connections (see Fig. 1). Tstat monitors each TCP connection, exposing information of more than 100 metrics. It implements DPI mechanisms to identify application layer protocols, such as HTTP and HTTPS, and records the server fully qualified domain name the client resolved via DNS queries. To reduce the privacy risks, Tstat anonymizes IP addresses and removes parameters from URLs.

For the works presented here, I used Tstat in an European ISP network and in my university campus in Torino. For the ISP, three probes have been installed in Points of Presence (PoPs) of different cities, where they observed about 25 000 households overall. Each household is assigned, and uniquely identified by, a static IP address. Users connect to the Internet via ADSL or fiber, using a single access gateway offering Ethernet and WiFi home network. These passive traces can be classified as big data and therefore requires use of scalable software and hardware: data storage and manipulation has been done thanks to the use of Hadoop and Spark on a Big Data cluster.[1]

Part of my work required also active traces, obtained instrumenting browser applications and crawlers. In Section V I used passive traces from an advertisement platforms (i.e., Avazu), publicly available over the Kaggle platform.[2]

All the datasets are available online (anonymized) [9]: most of the results of this manuscript can therefore be validated, repeated and extended by any external researcher.

B. Privacy

It is fundamental to find a trade-off between the desire to obtain knowledge for shaping new technologies and the need to not violate the privacy of individuals. Both the data collection processes and the collected datasets have been discussed, reviewed and approved by the ethical board of my university and by the ISP security board. I took all possible actions to protect leakages of private information and the identity of users itself. In particular, the IP addresses of clients are anonymized using a technique based on irreversible hash functions, and only the data that is strictly needed for my studies is retained. ISP home Internet installations are identified by anonymized keys, and browsers by *user-agent* strings. Privacy requirements

[1]https://smartdata.polito.it/computing-facilities/
[2]https://www.kaggle.com/c/avazu-ctr-prediction

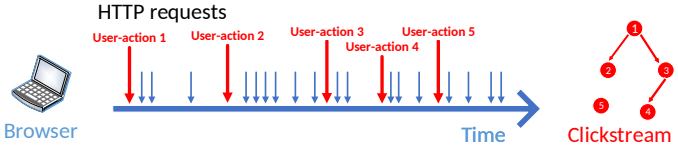

Fig. 2. Example of a client browser activity, where its user-actions are highlighted. User-actions create clickstreams.

limit any different granularity. In sections III-B and IV I limit the data to i) the anonymized client IP address, ii) the name of the contacted server, and iii) the timestamp of the TCP connection.

In [10] I deepen the problem of online privacy. I study the entities than can collect and access these kind of data, other than researchers, highlighting what are the privacy and ethical issues that arise for users, companies, scientists and governments and presenting the current legislation.

C. Identification of user-actions

A *user-action* is the explicit action of requesting a URL by a user to fetch a web-page, triggered by an interaction with a browser [2]. The fundamental technical challenge is to extract user-actions from raw HTTP logs. Indeed, rendering a web-page is a complex process [11] that requires the browser to download HTML files, JavaScript, multimedia objects and dynamically generated content. All these objects are retrieved via independent HTTP requests. Furthermore, non-interactive web applications (e.g., cloud storage clients and OS updates) rely on HTTP to exchange data too, and all those requests are mixed together with users' activity. User-actions thus correspond to *web-pages* explicitly visited by a user. Fig. 2 depicts the timeline of a user surfing the web. The user visits five web-pages, whose corresponding user-actions are marked by tall red arrows. I call *clickstream* the list of user-actions. The clickstream, shown on the right-hand side of Fig. 2 is typically modeled as a directed graph, where web-pages constitute the vertices, and edges represent the movement of a user through web-pages, i.e., following hyperlinks.

My classification problem consists in identifying the visited URLs that are user-actions. In the past this problem has been faced by designing ad-hoc heuristics driven by domain knowledge, e.g., by rebuilding the web-page structure [11], [4], or manually building blacklists and simple tests [12]. Machine learning approach allows automatic tuning of parameters, learns which features are the best candidates for solving the problem, and adapts to different scenarios.

I collect browsing histories containing user-actions of 10 volunteers for several months, while also recording all HTTP requests of their web navigation through Tstat. I next label entries in HTTP logs that match browsing history entities as user-actions, with care to manage redirections and avoid requests coming from non-considered web browser. At the end of this process, about 2% of all HTTP requests are labelled as actual user-actions. I then extract a large number of features to feed a clasiffier. I consider 17 features that can be grouped into

four categories: (i) based on referring relations among URLs; (ii) based on timestamps; (iii) describing properties of objects; and (iv) describing properties of URLs. I compute their information gain, and analyze and select the most informative ones for the problem. I considered four different classification algorithms, i.e., Bayesian networks, decision trees, random forests and neural networks, using stratified 10-fold cross-validation. Among the four, the decision tree and random forest perform the bests, with F-Measure equal to 90.6%. Given a decision tree is simpler and readable, I decide to use it.

Results show that my approach generalizes ad-hoc designed heuristics, with both precision and recall over 90%. I show that models built with machine learning algorithms are robust, presenting consistent performance in different scenarios, also with smartphone traffic [5]. Aiming to foster further researches and validations of my results, I make the datasets and the classifier code available at [9].

D. Identification of Core domains

A similar methodology has been deployed for encrypted traffic [3]. The idea is to identify user-actions from flow level measurements. Here, I aim at building a list of domains that typically contain user-actions since they host actual web services. When visiting a web-page, the browser application first downloads the main HTML document and then fetches all the objects of the page (images, scripts, advertisements, etc.). These are often hosted on external servers (e.g., CDNs) having different domains. Given the nature of encrypted pages, here I want to identify the domain name originally contacted to download the main HTML document of a page, here called *Core domain*. Core domains are important since they are intentionally visited by users, like *www.facebook.com* and *en.wikipedia.org*. I call *Support domains* those domains automatically contacted by visiting a Core domain, or by background applications, like *static.10.fbcdn.net* and *dl-client.dropbox.com*. Support domains do not contain useful information about user intention.

Given a domain, I visit through active browsing the home-page it hosts to extract page features.[3]. Based on the response, I classify it as a Core or Support domain. I solve the classification problem with a machine learning approach using a decision tree classifier again. I consider an extensive list of features guided by domain knowledge. These include the length and the content type of the main HTML document, the number of objects of the page, the domains contacted by the browser to fetch all objects, and whether the browser has been redirected to an external domain. I manually build a labeled dataset that I use for training and testing, publicly available [9].

Interestingly, the final decision tree results in a simple, efficient, and descriptive model. Despite its simplicity, overall accuracy is higher than 96% when tested against 1 000 labeled domains, using 10–fold cross validations.

[3]Selenium automatic browser, http://www.seleniumhq.org/

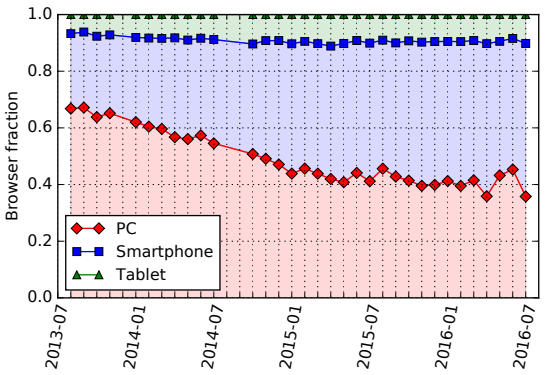

Fig. 3. Evolution of share of active browsers category from July 2013 to July 2016.

III. Characterize and Model Browsing Habits

A. Longitudinal characterization

In [5] I thoroughly investigate browsing habits of internauts, providing the evolution over three years. I apply the classifier of Section II-C to extract users' clickstreams. I answer the following two questions:

- How are the clickstream graphs affected by the web evolution over the past years?
- What are the differences between clickstream graphs from different browsing devices (e.g., PCs and smartphones)?

I provide a longitudinal characterization of the clickstream graphs. Fundamental to answer these questions is the availability of data. I leverage a three-year long anonymized dataset (July 2013 - July 2016). Analyses are performed in a *per-browser* level, i.e., the combination of the household identifier and the user-agent string. People may use several browsers to explore the web, and several persons may be aggregated in a household. The probes monitored 25 000 *households* and observed more than 64 billion HTTP requests. From this dataset, I extract the user-actions and build clickstream graphs for each browser in a household. In total, I construct 5.5 million graphs corresponding to over 1 billion visited web-pages. To the best of my knowledge, this is one of the largest datasets available online [9] that includes clickstream graphs from regular Internet users, browsing with multiple devices.

I present a characterization of clickstreams that confirms and precisely quantifies many intuitions about the way people navigate the web, besides leading to a number of interesting findings. Here I outline the main results.

Firstly, web-page complexity has continuously increased from 2013 to 2016, with URLs intentionally visited by users going from 2% to 1.5% of the total number of URLs requested by browsers.

The number of devices and applications used to browse the web at home has increased significantly, with smartphones and tablets accounting for 29% and 9% of the visited web-pages in 2016, respectively. Users are interacting more frequently with the web from their smartphones at home than in the past [4].

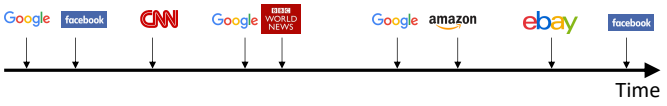

Fig. 4. Example of the temporal sequence of Core domains visited by a user, called a user trajectory.

For example, see Fig. 3, where I report the fraction of browsers per category. However, in a session on a mobile browser only 5 web-pages are visited on average, in a time span of only 2 minutes.

When considering the number of visited web-pages, we observe that 50% of the clickstream graphs include less than 27 web-pages per day for PCs (8 for smartphones), belonging to less than 9 domains (4 for smartphones). Considering consecutive visited web-pages, i.e., a path, we observe that people visit very few domains, even when navigating through hundreds of web pages. These numbers have mostly remained constant over the years, despite changes in devices and applications used to browse the web.

Search Engines (SEs) and Online Social Networks (OSNs) are among the preferred means to discover content. As of 2016, 54% of web domains were visited starting from Google, and 9% (6% in 2013) starting from Facebook. SEs are starting point of longer and deeper navigation, while content promoted by OSNs typically generates visits to a single or very few web-pages. Interestingly, OSNs are much more important to discover content on smartphones than on PCs, a result previously not highlighted.

Encryption has gained momentum in the web with many popular domains migrating to HTTPS. We can see the impact of HTTPS on properties of the clickstream graphs. Still, in June 2016, only around 13% of the domains were served (partly or totally) in HTTPS, and 85% of the encrypted traffic was related to the top 20 content providers. Transitions *from* popular encrypted domains *to* the unencrypted ones were still visible in the analysis. Thus, an encrypted domain appears as a single vertex in a clickstream graph, connected to all vertices representing plain domains visited from it.

B. Modeling web trajectories

Users' browsing activities can be also described by means of the paths they follow when navigating through web-sites. Even with nowadays widespread encryption, a passive network observer can still obtain valuable information about the trajectory a user follows. For instance, the DNS queries of domains of the web-sites contacted during browsing are still not encrypted and easily accessible from passive probes. In this part of the work, published in [6], I refer to the sequence of domains visited by a user as the user *trajectory*. Both user's circumstances and preferences affect such trajectories. Here, I consider only the domains intentionally visited, i.e., Core

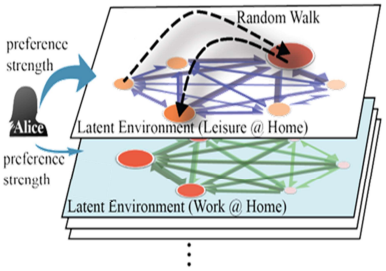

Fig. 5. Each user performs random walks over latent environments, with different probability of interest (preference strength) towards environment.

(a) Environment with computer-science related domains.

(b) Environment with many travel related domains.

Fig. 6. Examples of top-10 domains in different environments, showed as word-clouds.

domains (see Section II-D). In Fig. 4 I show an example of a trajectory of Core domains. Notice how the same domain can appear multiple times.

Armed with these sequences of visited domains, i.e., user's trajectories, I analyze them by modelling each user as a random surfer over latent environments. User trajectories are the outcome of a combination of latent user preferences and the latent environment that users are exposed to in their browsing [13]. It is expected that real user behaviour will be (i) non-stationary, and (ii) time heterogeneous. In other words, user behaviour change and evolve over time, and is different for each user. The model I use is designed to cope with the complex challenges of learning personalized predictive models of non-stationary, time heterogeneous, and transient (Markovian) user trajectories. Each environment captures a latent factor that leads to a user visiting a domain. See the representation in FIg. 5 for a visual representation.

I build this model and analyze its results using traffic summaries of $\approx 7\,500$ anonymized users in my university campus and in an ISP (as in Section III-A) during 4 weeks in 2017. I collect and log information about each TCP connection and I extract all the Core domains, for each user (using the methodology explained in Section II-D). Focusing on these domains, I reconstruct meaningful trajectories over time. At this point, I use the possibly best fitting model on such data. A big data approach must be considered for retrieving, processing and managing such amount of data. As usual, the anonymized trajectories of domains and their models are

publicy available [9].

Thanks to this model, my methodology and analysis shows that it is possible to:

- Model accurately the users trajectories, by simply considering domains names.
- Automatically extract environments with similar or likely connected web-sites; For example Figure 6 illustrates the word-clouds of the top 10 domains of specific environments. Observe how expressive are the word-cloud in describing the topic of each environment.
- Highlight differences in terms of popularity and content of environments.
- Extract the interests of communities of people.

IV. Users' fingerprints from visited domains

Privacy and user tracking are hot topics that impact everyone who uses the web. Encryption limits access to exchanged information, yet a lot of information can be extracted. I explore different techniques for profiling and fingerprinting users by using only the information about visited domains. Would it be possible to build an accurate user profile by simply considering the set of domains she visits during her browsing session? And would it be possible to re-identify her in a future time, e.g., when she is connected in a different network? Real-case scenarios include applications for tracking users in different networks, e.g., tracking users from both mobile and house traffic from a certain area, in which we may want to associate the two datasets. Or, when users change their IPs due to dynamic assignment.

Armed with the large datasets used in Section III-B ($\approx 7\,500$ anonymized users during 4 weeks in 2017, trace available at [9]), I answer the previous questions. I investigate the use of three metrics, considering (i) a simple *Jaccard index*, (ii) an information theory *Maximum Likelihood* approach [14], and (iii) a text mining methodology based on *TFIDF* (Term Frequency - Inverse Document Frequency). I evaluate their performance, highlighting their strengths, weaknesses and trade-offs. Results unveil that TFIDF offers overall the best performance, identifying a given user in different scenarios with up to 94% of accuracy. The rationale of this surprising result is the fact that among the hundreds of domains visited during few days, many are persistent in time and create a peculiar and unique mix of traffic.

To give an intuition about the discriminative power of the built profiles, I report in Fig. 7 the Cumulative Distribution Function (CDF) of TFIDF metric between the same user (called self-similarity) and between two different users, considering all the visited domains from two consecutive weeks. Self-similarity is much higher than the similarity with a different user, thus allowing us to correctly identify the target user.

To get more insights, I investigate which domains are more useful for such purpose, in particular considering those intentionally visited by the users, the *Core domains*, or those contacted by the browser to fetch objects that compose a web-page or by other background applications, the *Support domains*. Results show that intentionally visited web-services prove to better characterize the user than Support domains; however users are better re-identified when all the traffic is taken into account, suggesting that even Support domains help in characterizing users.

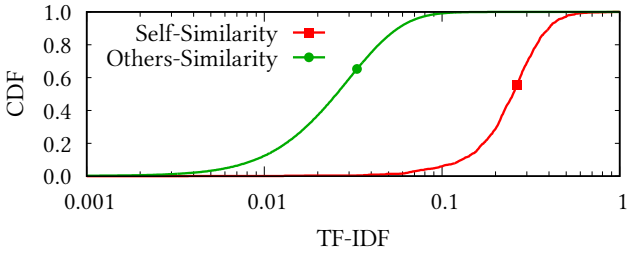

Fig. 7. CDF of similarity in two consecutive weeks between the same user, or two different users.

My study shows on the one hand how complicated is to protect privacy when online; on the other hand, the potential of good similarity metrics and machine learning applications links to, e.g., forensic.

V. User interaction with advertisements: modelling and optimization of ads Placement

In recent years we have seen a proliferation of online platforms offering different types of services. Profits are often obtained through ad sales, i.e., the insertion of advertisements within the content displayed to users. In this section I focus on the influence of online advertisements on the users and their interaction with recommendation systems [3].

I consider an online system for targeted advertising. A publisher provides available ads slots, i.e., portions of the user's navigation experience where ads can be inserted. At the same time, an advertiser provides ads that can potentially fill in those slots. The match between available slots and ads generates impressions shown to each given user, who might decide to perform valuable actions on them, such as clicks. An ad server has to decide which impressions to deliver to each specific user, and at which time instants. Both the publisher and advertiser's revenues grow with the increase of these users' valuable actions. As an example, we can consider advertisements for a pay-to-play online game, where each time the user plays, she pays for the service. Therefore, the user is pushed to resume playing (and paying) through advertisements. In their attempt to maximize their revenues, publisher and advertiser should take into account that the likelihood of a user to perform a valuable action may be impacted by the history of shown impressions. For example, a user overwhelmed and annoyed with impressions arriving too close in time might be less likely to perform actions on them. Thus, the number and temporal spacing of impressions can be optimized, as it has already been recognized (e.g., in [15]).

I study the detailed temporal dynamics of the explained advertising system by developing a model that incorporates the user's reaction. Then I estimate the likelihood that the

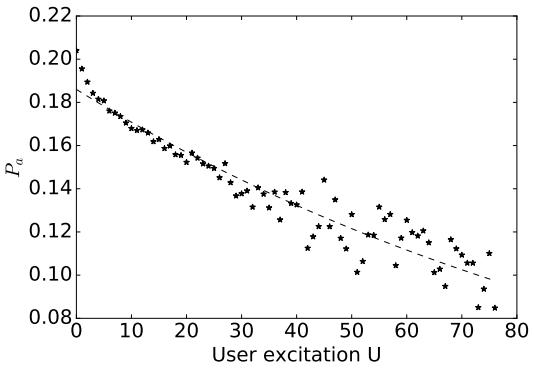

Fig. 8. Experimental P_a obtained from Avazu dataset. User response and excitation are correlated.

user will perform a valuable action on a particular impression. Finally, I try to maximize the number of clicks to ads per time unit, called Click-Through-Intensity (CTI). To the best of my knowledge, this is first optimization of the frequency capping of an ad campaign using a behavioural model, capturing the main features of the above system.

In my stochastic model, the user performs a valuable action on an impression also depending on the history of past impressions. I introduce the probability P_a that the user performs such action, and a user excitation U. The user excitation U depends on the previous seen ads and, if no new ad arrives, decreases in time. The Avazu dataset reports the click/no-click actions performed by 9 million users on on-line ads, over 10 days. I tune the parameters of the model on the trace and compute the evolution of the excitation U for each user using the impression arrival times reported in the trace. Fig. 8 reports the obtained empirical P_a. The dashed line in the plot shows the best least-square fitting of the data, revealing a significant correlation between the user response P_a and the user excitation U. This suggests that my methodology can be effectively employed to model the system and there is possibility to maximize the CTI. If P_a were independent, or very weakly correlated with U, the best strategy would simply be to overwhelm the users with ads.

Then, I identify different regimes of the proposed system and devise analytical and numerical strategies to optimize the CTI metric. Such strategies can provide useful theoretical benchmarks for the deployment of better advertising platforms. Applying these strategies to the real trace shows that impressions can be delayed, and thus better spread out over time, obtaining significant improvements of CTI, namely 7%. Therefore, my model allows to optimize the sequence of impressions itself, achieving significant gains in terms of of profits.

VI. Conclusions

My findings have several implications to the Internet actors. For example: (i) network operators can find anomalies in behaviour of users or group of users; (ii) models of trajectories can be used to propose personalized recommendation systems for users browsing; (iii) clickstreams analysis can help advertisers to make informed decisions on whether to target ads campaigns on specific device users; (iv) my fingerprint study should stimulate researchers to investigate privacy aspects and find countermeasures; (v) advertisement companies can use my model to optimize the sequence of impressions, achieving gains in terms of profits.

Finally, the current digital transformation implicates that everyone and everything produce data that can be exploited to create new disruptive capabilities. Data analytics allows us to realize incredible transformations not only in the web. Exploiting the knowledge of the users' behaviour from these data, modelling and optimizing system performances as I did in my work, will be a key factor for designing future architectures in many fields.

References

[1] L. Vassio, "Data analysis and modelling of users' behaviour on the web," Ph.D. dissertation, Politecnico di Torino (Italy), 2018. [Online]. Available: https://lucavassio.files.wordpress.com/2018/09/vassio_tesi.pdf

[2] L. Vassio, I. Drago, and M. Mellia, "Detecting user actions from HTTP traces: Toward an automatic approach," in *Proceedings of 2016 IWCMC Conference*, 2016, pp. 50–55.

[3] L. Vassio, D. Giordano, M. Trevisan, M. Mellia, and A. P. C. da Silva, "Users' Fingerprinting Techniques from TCP Traffic," in *Proceedings of ACM SIGCOMM 2017 Big-DAMA Workshop*, 2017, pp. 49–54.

[4] G. Xie, M. Iliofotou, T. Karagiannis, M. Faloutsos, and Y. Jin, "Resurf: Reconstructing web-surfing activity from network traffic," in *Proceedings of 2013 IFIP Networking Conference*, 2013, pp. 1–9.

[5] L. Vassio, I. Drago, M. Mellia, Z. B. Houidi, and M. L. Lamali, "You, the web, and your device: Longitudinal characterization of browsing habits," *ACM Transactions on the Web*, vol. 12, no. 4, pp. 1–30, 2018.

[6] L. Vassio, F. Figueiredo, A. P. C. da Silva, M. Mellia, and J. M. Almeida, "Mining and modeling web trajectories from passive traces," in *Proocedings of 2017 IEEE International Conference on Big Data*, 2017, pp. 4016–4021.

[7] L. Vassio, M. Garetto, C. Chiasserini, and E. Leonardi, "Rethinking online advertising: from click through rate to click through intensity," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, vol. Under Review, 2018.

[8] M. Trevisan, A. Finamore, M. Mellia, M. Munafo, and D. Rossi, "Traffic analysis with off-the-shelf hardware: Challenges and lessons learned," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 163–169, 2017.

[9] L. Vassio, "Three-year long dataset of anonymized clickstreams. Anonymized visited domains, core domains, trajectories and their models are available." https://smartdata.polito.it/category/open-datasets/.

[10] L. Vassio, H. Metwalley, and D. Giordano, "The exploitation of web navigation data: Ethical issues and alternative scenarios," in *Blurring the Boundaries Through Digital Innovation*. Springer International Publishing, 2016, pp. 119–129.

[11] S. Ihm and V. S. Pai, "Towards understanding modern web traffic," in *Proceedings of 2011 IMC*, 2011, pp. 295–312.

[12] Z. Ben-Houidi, G. Scavo, S. Ghamri-Doudane, A. Finamore, S. Traverso, and M. Mellia, "Gold mining in a river of internet content traffic," in *Proceedings of 2014 TMA Workshop*, 2014, pp. 91–103.

[13] F. Figueiredo, B. Ribeiro, J. M. Almeida, and C. Faloutsos, "Tribeflow: Mining and predicting user trajectories," in *Proceedings of 2016 WWW*, 2016, pp. 695–706.

[14] J. Su, A. Shukla, S. Goel, and A. Narayanan, "De-anonymizing web browsing data with social networks," in *Proceedings of 2017 WWW*, 2017, pp. 1261–1269.

[15] N. Gupta, A. Das, S. Pandey, and V. K. Narayanan, "Factoring past exposure in display advertising targeting," in *Proceedings of 2012 ACM SIGKDD*, 2012, pp. 1204–1212.