Your Data in the Eyes of the Beholders

Design of a unified data valuation portal to estimate value of personal information from market perspective

Yonas Mitike Kassa[a,b], Jose Gonzalez[b], Ángel Cuevas[b], Rubén Cuevas[b], Miriarm Marciel[b,c], Roberto González[c]
[a]*IMDEA Networks Institute,* [b]*Universidad Carlos III de Madrid,* [c]*NEC Labs Europe*
yonas.kassa@imdea.org, {*jgcabana, acrumin, rcuevas*}*@it.uc3m.es,*
{*miriam.marciel, roberto.gonzalez*}*@neclab.eu*

***Abstract*—Nowadays Internet companies that offer valuable services "for free" are becoming ubiquitous. Users benefiting from these services have to expose their personal information through these services as they utilize them. On the other hand, personal information is becoming a merchandisable commodity, venues that sell personal information by auction are emerging. One of these markets is in the form of advertising systems. Despite being a lucrative business, the hoarding of user personal information by commercial companies is a growing issue primarily because of its non- transparent nature. In this paper we present a data valuation portal that shades light on what kinds of personal information is on market and the financial value of it.**

***Keywords*-online advertising; privacy; personal data; data transparency; advertising economy**

I. Introduction

Online advertising, which is the backbone for majority of "free" Internet services, is now a multibillion-dollar industry. In 2014 alone, online advertising generated $49.5B worth of revenue only in US, representing an increase of 16% with respect to 2013, which in turn exceeded 17% the revenue of 2012 [12]. It is this impressive size and growth that permits online advertising to emerge as the main source of revenue for most of Internet services such as search, social media and user generated content sites which are at the forefront of innovation in the Internet and have generated more than 3.4 million direct and indirect jobs in Europe in 2012 alone [11].

In parallel, the technology of online advertising has evolved from simple static display banners put to the sides of a web page by direct agreement between publishers and advertisers into a sophisticated dynamic system involving a multitude of intermediaries with the objective of matching relevant advertisements to a target user. To achieve this objective, these intermediaries leverage distributed tracking techniques to gather information about the online activities of each user and generate user profiles based on the collected personal data.

Even though this method is proven to be effective in identifying user demographics, interests and behaviors, there is an increasing concern among researchers and policy makers as little is known about how this personal data is used and more specifically, users have no control over what happens with their data. In addition, how this collected data is marketed is not transparent to users and they dont know the economic value of their data. In other words, knowing how much money companies in the online advertising market make out of users personal information is not upfront.

As the economic value of their data is not known to the average user, they are not able to evaluate the value of their data relative to the value they assign to literally "free" services like Google, Facebook and YouTube. For this reason, average users are not aware of the cost associated with their personal data from market perspective. On the contrary, many online service providers collect and monetize personal information in exchange for the free" services they offer. This collected personal data is very valuable to advertisers as it enables them identify and target potential customers. In fact, personal data is now a tradable asset that is integrated into advertising platforms.

We believe that a good way of educating people is by letting them know what is the actual value of their personal information across services on the web by providing tools that let them know the actual value of their personal information. Towards this end our goal is to construct a publicly available and user-friendly web-platform that allows end-users to configure a particular profile for a number of services (Google, FB, YouTube, LinkedIn) and inform them of the value of such a profile in a selected market. Using this web-portal users, regulators, and various stakeholders can inspect the monetary value assigned to specific demographic and behavioral profiles (age, gender, location, interests, etc.) in different major online services.

II. Related work

Despite the fact that there is a large amount of research spanning different aspects of internet services and online advertising such as its privacy implication([9], [4]), analysis([2], [8], [4]), and improvements[1], there is a very little work addressing the question of what is the actual quantitative value that the online advertising market assigns to end users' personal information.

To the best of our knowledge, there is very little work addressing the question of what is the actual quantitative value that the online advertising market assigns to end users personal information. There are two recent works [10], [13] that are partially related to the goal we want to achieve. In [10], the authors leverage the Facebook ad

978-1-5090-0990-9/16 $31.00 © 2016 IEEE
DOI 10.1109/ARES.2016.55

planner API to understand the auction system in OSNs. The work indirectly introduces some interesting elements related to the value of personal information since it analyses the evolution of the CPM median cost for 4 different groups of users and discusses about the price stability over time.

The work presented by Saez et al. [13] proposes a model that defines the value of a user according to the actions of that user in Facebook. In this case an action is summarized in three categories: write a post, upload a multimedia video and join a community (e.g., group, page, etc.). Hence, the value of users information depends on their capacity to spread information rather than their personal information in itself. The authors evaluate the proposed model with a dataset including information of +90k users from New Orleans, US. In addition, they leverage the FB ad planner API to evaluate the value of each user based on her profile interests. However, very few results are shown and the value of each user is computed in relative terms with respect to the most valuable users. Finally, all the users are located in the same city, and our initial investigations show that the location is one of a very important attributes that play a significant role on the value of a user.

We also want to mention that the individual research tools used in all of the above related works are not targeted for the general public in a user-friendly manner that can serve at a large scale, the lack of work in this area makes our work the first of its kind.

III. Background

In this section we give a quick background to online advertising and then proceed to types of advertising and ad monetization techniques.

Since the introduction of clickable banner ads in early 1990s the technology of online advertising has evolved to include many players including web publishers that sell slots (also called inventories) in their web pages or apps to advertisers that want to reach potential customers; and intermediaries that facilitate this process. Intermediaries include ad-networks and ad-exchanges that serve as a marketplace for automatic trading of ad inventory. Demand Side Platforms (DSPs) and Supply Side Platforms (SSPs) are other intermediaries lately incorporated into the ad-serving ecosystem, they operate as a unified interface for advertisers and publishers to interact with multiple ad-exchanges respectively. Figure 1 shows a bird's eye view of the complexity of the ecosystem[1]. Advances in advertising techniques has resulted in the following three major types of targeted advertising.

A. Contextual advertising: in this form of advertising advertisers select inventories based on the contents of the web page oblivious of the type of users visiting it.

B. Retargeting: this form of advertising is based on the shopping history of users across websites, using this technique an advertiser follows potential customers who showed interest in a particular product but left without conversion.

C. The other most advanced and effective targeting mechanism is behavioral advertising that utilizes personal information about users based on different user tracking techniques. it mainly leverages cookies and tracking pixels. In this scenario when a user consumes a given internet service the intermediaries integrated in the service expose information about her to advertisers which in turn decide the value of that particular user and offer a monetary value through Real Time Bidding (RTB) - if they find that user useful.

[1] www.lumapartners.com/resource-center/lumascapes-2

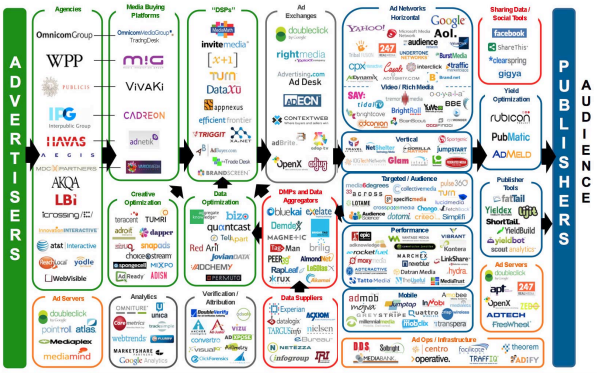

Figure 1. Digital advertising ecosystem (source LUMA partners)

Cost per click (CPC), cost per thousand impressions (CPM), and Cost Per Action (CPA) are few of the major Bidding techniques. CPC is a pricing model where advertisers pay for a user click, while in CPM payment is per thousand displays on the given slot independent on clicks. In CPA advertisers pay when a user acts following the ads shown, actions could be app installs, registration, or purchase.

IV. Web Portal Design Constraints

Our goal in this section is to present the design of a personal data valuation portal composed of front-end and back-end components which in turn have distinct sub-components that will be described later in this section.

The final goal of this design phase is to provide average users with a user friendly platform that will allow them to understand the value that the online advertising market gives to their personal information. This ambitious goal has to deal with the following challenges.

A) The online advertising market is not transparent enough, and it is not easy to access the bidding information associated by the advertisers. Therefore, the first challenge is to find appropriate and relevant sources of information that serve as input for the tools we plan to develop. To the best of our knowledge, the only sources providing information regarding aggregated bidding information of advertisers for particular audiences are major Internet Players such as Google and online social networks and social media like Facebook, LinkedIn or Twitter.

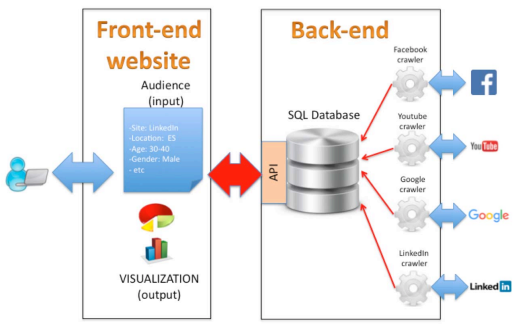

Figure 2. Architecture of the data valuation portal

B) Creating sophisticated crawlers that allow us to access first-hand bidding information available through the referred data sources. Note, that each company provides distinct mechanisms to access the bidding information, and thus each information source will require unique implementation of specific crawlers. Whenever it is possible we will exploit APIs made available by the information providers.

C) Storage of the crawled information has to be efficient and robust DBMSs that have to fulfil the following requirements are required: (i) large storage capacity to store time-based (e.g., every day, every hour, etc.) bidding information for a large number of audiences, (ii) quick resolution of queries, (iii) flexibility to extend the data sources in an easy manner, and (iv) efficient design that allows an easy integration with backend crawlers that periodically collect information and the front-end interface to end-users.

D) Design and implement simple and accessible front-end UIs for average users in order to maximize the adoption of the developed tools. The front-end of the tools developed in this task should be very simple and intuitive in order to maximize the number of users that can use the tool. In addition, it needs to be portable to many operative systems, web browsers, mobile environments, etc. so as to maximize the number of users with access to the platform.

V. Web portal implementation

To tackle the above constraints, the portal is currently implemented as a two component system with four back-end crawlers from the following major services: Facebook, YouTube, Google, and LinkedIn. Figure 2 shows the design of the architecture we are implementing for the data valuation web portal. The different major components of the platform are described as follows.

A. **User-facing front-end:** the front end component is implemented as a PHP module integrating two functionalities which are the following:

1) Website Profile Value Planner: This component is the user-facing component which lets the user select a target market from the above available services, the time window, and configure a desired profile by adjusting a set of parameters(e.g. location, gender, age, behavioral interests like games, drinks and fashion) via the web interface. We note that the parameters to define audiences are different across the four ad campaign planners. Therefore, this tool allows average users to understand the value associated to different audience profiles on different markets.

2) Web report generator: Using this same front end interface the user can retrieve one or more of the following values: Cost per Action (CPA), Cost per Mile (CPM), and Cost per Click (CPC) from the ads planner facilities of four big Internet players mentioned above. This interface handles the visualization of query result via a simple reporting chart that shows the temporal evolution of the selected value (CPC, CPM, and CPA) over the requested time window. The y- axis of this report visualization shows the monetary values of configured profile, while the x-axis shows the dates where bidding data was retrieved. This vizualization component is implemented with amCharts[2] charting library.

B. **Backend crawler #1 Facebook:** Facebook[3] offers a very complete ad planner to its advertising customers interested in finding a specific set of Facebook users with certain characteristics. Advertisers in Facebook can tune a large number of parameters. A non-exhaustive list of these parameters includes: Location (at country, region and city levels), Gender, Language, Education Level, Relationship Status, Ethnic Affinity, Life Events, type of Mobile Device, Interests, etc. This flexibility allows advertisers to define very specific audiences by combining multiple parameters together. However, collecting all possible configurations of audiences is an expensive process. This problem can be explained by the fact that finding the whole set of audiences Aud_p by combining set of parameter groups $P = [P_1, P_2..., P_N]$ is a multiplication of permutations and trying to retrieve the information of all Facebook parameter permutations is computationally unaffordable.

$$\#Aud_p = \prod_{i=1}^{N} |P_i|$$

C. **Backend crawler #2 YouTube:** YouTube[4] also offers to its advertisers the possibility to create campaigns targeting specific set of users that watch YouTube videos so that advertisers can place ads on different parts of the platform. The targeting

[2] www.amcharts.com/
[3] www.facebook.com
[4] www.youtube.com

options offered by YouTube include the following: Geographical information, devices (OS type, device model, and carriers), user preferences, Gender, age groups, and type of bidding. For a selected targeting option YouTube returns the recommended maximum and minimum bidding price based on past winning prices for the same audience. In order to crawl these prices, we use Selenium web driver [6] library used to perform automatic interactions with the platform and configure targeting options in order to retrieve the bidding price of the specified audiences.

D. **Backend crawler #3 Google:** Being one of the biggest players in online advertising business Google offers its advertising customers the option to select their targets on its different venues via its AdWords platform described as follows: Keyword Planner that provides relevant information for advertising campaign based on the keywords in search queries in Google search and this planner returns the suggested bid in terms of CPC. Display planner: helps plan and create target audiences that can be reached via a large collection of websites, mobile apps, and video content. Ads: this planner returns suggested bids in CPM and CPC. In order to define the different audiences, we will use the targeting settings that the Keyword Planner tool allows to use for a targeting audience: interests (for this we leverage the google interest category tree which is composed of 2000 interests as our initial set of interests.), location (at country level), Languages (the planner allows to target more than 40 languages). In summary, our initial list of targeting settings for google will be formed by more than 10000 audiences (250 interests * 40 locations).

E. **Backend crawler #4 LinkedIn:** LinkedIn[5], one of the biggest professional social networking services, also offers advertisers the possibility to target users in their social network. The targeting Options offered by LinkedIn are Location (86 countries), Gender, Age group, Type of bidding (CPC or CPM). For each configuration of ad Campaign LinkedIn returns the minimum and maximum price to reach these configured audiences along with the number of users of that category. As in the case of YouTube crawler we leverage the selenium library to fetch this data from LinkedIn.

F. **SQL Database:** is the central repository where all the collected information by the crawlers is stored. This database serves user requests via the API component.

G. **API:** the API module connects the front-end with the data stored in the database. This API accepts user requests via HTTP requests and maps these requests to distinct SQL queries to the DB, and transforms the database response into an HTTP response and sends back to the front end.

[5]www.linkedin.com

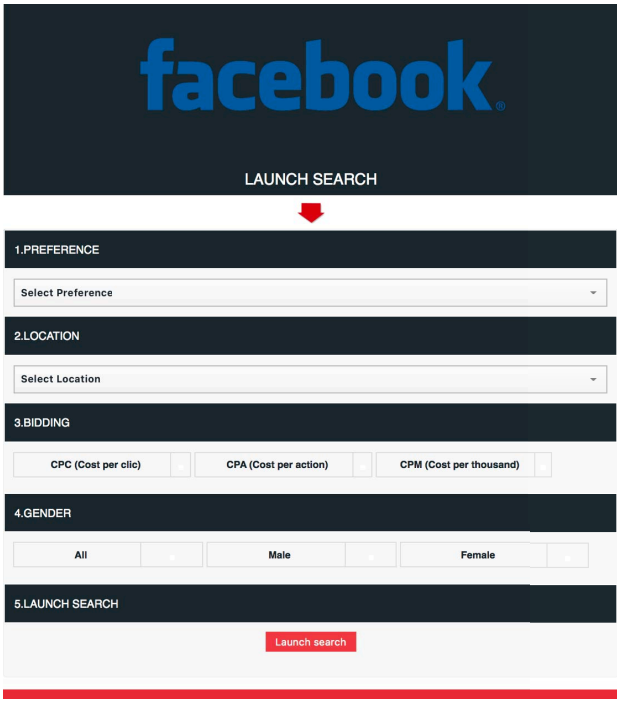

Figure 3. parameter configuration interface of Website Profile Value Planner

VI. Web portal mock-up

This section briefly describes the usability of the platform. Users can select one of the available markets from the menu displayed on top of the front page. Following this step, the value planner will generate the configuration parameters available in that given market. Sample configuration options for Facebook market is shown on Figure 3,. From these parameters users can build a profile by configuring a set of parameters (e.g., Country, Interest, Age, Gender, etc.) and then, depending on the chosen market, they will choose one of payment methods for the profile in terms of CPC, CPM or CPA.

Once the user has made her selection of the target market, the parameters she is interested in, and the type of costs, the web report generator displays a temporal diagram indicating the evolution of the cost in a temporal window of one week, with X and Y axes representing the day and cost value of the profile on that day respectively. A sample result showing CPM values of a user in United states interested in cosmetics is depicted in Figure 4, for comparison the value of a user located in India with the same interest is shown on Figure 5. The red lines on the report represent the minimum CPM bidding prices, while the black lines show maximum CPM bidding prices. The temporal window can be re-sized using a zoom functionality available on top of the graph to obtain a longer time window of the associated value to the chosen configuration.

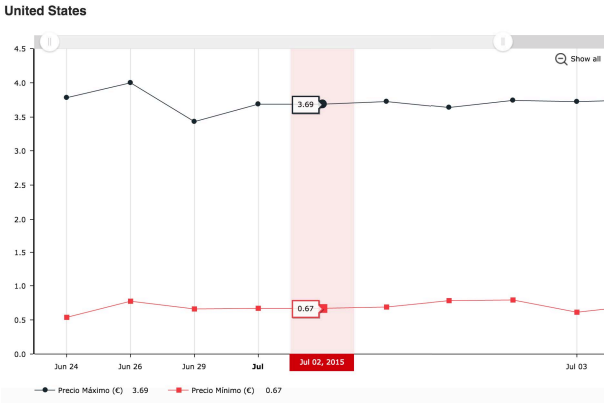

Figure 4. Report showing CPM value of configuration for US user

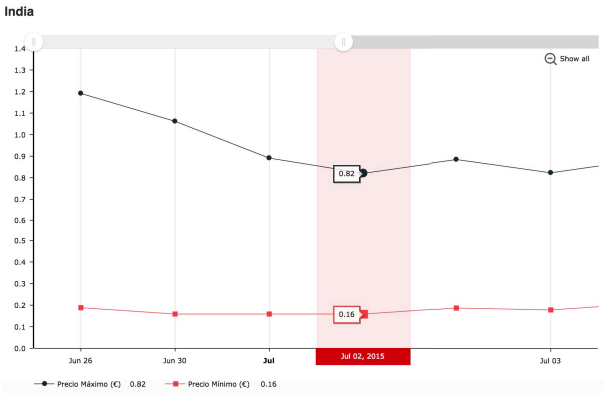

Figure 5. Report showing CPM value of configuration for Indian user

VII. Conclusion

In this paper we presented the architecture and a prototype implementation of a personal data valuation portal that shows the monetary values of different personal information from the actual market perspective. With the objective of developing a user-friendly portal to help users understand the economic value of their data, this work covered four different markets. As part of future work we plan to include other markets and also make the platform more extensible to include more demographic and behavioral information that we were not able to cover so far.

Acknowledgment

The research leading to these results has received funding from the Horizon 2020 Programme (H2020-DS- 2014-1) under Grant Agreement number 653449.

References

[1] W. Chanthaweethip, X. Han, N. Crespi, Y. Chen, R. Farahbakhsh, A. Cuevas. "Current City" Prediction for Coarse Location Based Applications on FB. IEEE GLOBECOM 13.

[2] Ruben Cuevas, Michal Kryczka, Angel Cuevas, Sebastian Kaune, Carmen Guerrero, and Reza Rejaie. Is content publishing in BitTorrent altruistic or profit-driven? ACM CoNEXT '10.

[3] https://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/99-ds-01-2014.html

[4] R. Farahbakhsh, X. Han, A. Cuevas, N. Crespi. Analysis of publicly disclosed information in FB profiles. IEEE/ACM ASONAM '13.

[5] Reza Farahbakhsh, Angel Cuevas, Ruben Cuevas, Reza Rejaie, Michal Fryzcka, Roberto Gonzalez and Noel Crespi. Investigating the reaction of BitTorrent content publishers to antipiracy actions. IEEE P2P '13.

[6] Selenium web-driver http://docs.seleniumhq.org/projects/webdriver/

[7] R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas. 2013. Google+ or Google-?: dissecting the evolution of the new OSN in its first year. WWW '13.

[8] X. Han, L. Wang, S. Park, A. Cuevas, N. Crespi. Alike People, Alike Interests? A Large-Scale Study on Interest Similarity in Social Networks. IEEE/ACM ASONAM 14. (Extended version published at Decission Support Systems).

[9] X. Han, L. Wang, J. Wen, A. Cuevas, C. Chen, N. Crespi. Is Your Hidden Location Undercover? Predicting Current City from Profile and Social Relationship. Tech. report available at http://arxiv.org/abs/1508.00784.

[10] Y. Liu, C. Kliman-Silver, R. Bell, B. Krishnamurthy, and A. Mislove. Measurement and analysis of OSN ad auctions. ACM COSN '14.

[11] http://www.iabeurope.eu/files/6713/6990/7349/IAB_Europe_study_-_Online_Jobs_Boosting_Europe_s_Competitiveness_-_Vlerick_Business_School.pdf

[12] http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_FY_2014.pdf

[13] D. Saez-Trumper, Y. Liu, R. Baeza-Yates, B. Krishnamurthy, and A. Mislove. 2014. Beyond CPM and CPC: determining the value of users on OSNs. ACM COSN 14