

FDVT: Data Valuation Tool for Facebook Users

José González Cabañas
Universidad Carlos III de
Madrid
jgcabana@it.uc3m.es

Ángel Cuevas
Universidad Carlos III de
Madrid
acrumin@it.uc3m.es

Rubén Cuevas
Universidad Carlos III de
Madrid
rcuevas@it.uc3m.es

ABSTRACT

The OECD, the European Union and other public and private initiatives are claiming for the necessity of tools that create awareness among Internet users about the monetary value associated to the commercial exploitation of their online personal information. This paper presents the first tool addressing this challenge, the Facebook Data Valuation Tool (FDVT). The FDVT provides Facebook users with a personalized and real-time estimation of the revenue they generate for Facebook. Relying on the FDVT, we are able to shed light into several relevant HCI research questions that require a data valuation tool in place. The obtained results reveal that (i) there exists a deep lack of awareness among Internet users regarding the monetary value of personal information, (ii) data valuation tools such as the FDVT are useful means to reduce such knowledge gap, (iii) 1/3 of the users testing the FDVT show a substantial engagement with the tool.

ACM Classification Keywords

H.3.5. Online Information Services: Data sharing

Author Keywords

FDVT; Facebook; Personal Data; Privacy; Data Valuation

INTRODUCTION

There are several public and private initiatives exposing the necessity of research activities that develop technologies to create awareness among Internet users regarding the value of their personal information. For instance, the OECD acknowledged the importance of having tools that allow measuring the monetary value associated to online personal data [16], but also highlighted that: (i) it is an extremely complex task, (ii) the existing methodologies are still in a very preliminary stage. Similarly, the European Commission launched in 2014 an open call for projects [6] that among other elements stated: *"Data protection and privacy frameworks in Member States and Associated Countries need to be implemented in a transparent and user-friendly way to help users understand how their personal data might be used, including the economic value of their data."* In this line, the Data Transparency Lab (DTL) [12], a private initiative that promotes transparency in

the management of personal information, included the following research topic in its 2015 Grant Program, *"Raising User and Societal Awareness - Measuring the value of personal information"*. Therefore, there is an increasing demand requesting tools that allow Internet users to know the monetary value of their personal information, in other words, the revenue they generate for online services that commercially exploit their personal information. To the best of our knowledge there is nothing close to such tool available nowadays.

We can find some effort, mostly in the economics literature, addressing the question of what is the economic value of personal information. The most adopted methodology uses the contingent valuation method widely applied in economics and marketing research. Basically, they rely in surveys/interviews where they ask users their willingness-to-pay (WTP) to protect/recover some personal information and/or their willingness-to-accept (WTA) to sell some personal information for a given amount of money. A second methodology uses market cap of online services to measure the average value of a user in that service, i.e., it divides the yearly revenue or net benefit of the service by the number of active users to obtain the average value of a user profile. These methodologies are rather limited if we think of them as data valuation tools. Basically, they just provide a static picture of an overall average value common to all the users in the system that does not consider: (i) different users generate different monetary value for online services depending on their personal information and online activity, (ii) users generate value for online services in a continuous basis. The referred methodologies do not capture the actual way in which users generate monetary value for online services. Hence, they are not valid to develop a data valuation tool.

A comprehensive data valuation tool should be able to provide Internet users with: (i) data values aligned to actual market prices, (ii) personalized feedback per user to let each user know an estimation of how much money she is generating out of her personal information, (iii) real-time information of the value generated over the time.

In this paper we present the first personal data valuation tool that meets those three requirements. This tool is based on a disruptive approach that measures the monetary revenue a user generates for an online service in real-time out of her activity in that service. This novel approach targets services that generate their revenue by commercially exploiting the personal information of Internet users through tailored advertising. In particular we have focused our effort on creating a tool that applies this methodology to one of the most popular online services, i.e., Facebook (FB), which obtains the vast

majority of its revenue through tailored advertising. Therefore, the first contribution of this paper is the so-called *FDVT: Facebook Data Valuation Tool*. The FDVT has been implemented as a Google Chrome extension that informs users of a personalized and real-time estimation of the revenue they are generating for Facebook while browsing in this system.

Although the FDVT has itself an inherent value for the research community, it also allows us to address a number of research questions that could not be handled without a data valuation tool in place. In this paper we will focus on three research questions that can be answered relying on the FDVT.

First, the FDVT provides us with a ground truth that can be used to evaluate what is the actual lack of awareness of Internet users regarding the value they generate out of their personal information. To address that question we performed a lab experiment with skilled Internet users, i.e., BSc, MSc and PhD students in the area of Computer Science and Telecommunications. The idea of using advanced Internet users is that if they show an important lack of awareness, it may suggest an important knowledge gap in the society as well. In the experiment, we exposed the students to a number of questions about the business model of Facebook and the value they consider they generate for FB per session and per month. Later we introduced them the FDVT and asked them to login to Facebook with their own account and run a regular session in a computer with the FDVT installed. After obtaining the FDVT feedback, we used a closed question to assess whether their own perception about the money they generate for FB is aligned to the FDVT estimation, and thus implicitly find their potential lack of knowledge regarding the monetary value they are generating for an online service like FB. Finally, they evaluated whether the FDVT is an appropriate tool to create awareness in the society about the value of online personal information.

Second, although there are several public and private initiatives highlighting the necessity of data valuation tools, they cannot assess in advance whether these tools will actually be able to capture the interest (i.e., engagement) of Internet users. We evaluate that question by analyzing the interaction of 59 beta-testers with the FDVT Chrome extension during a period of 5 months from March to July 2016.

Third, the FDVT forces users to undergo a registration process the first time they use it. In this process, we request them to fill 4 personal information items: Country, Gender, Age+birthday and Relationship status. The only compulsory parameter is the Country while the remaining ones are optional. In this paper we analyze whether users are reluctant to provide optional (personal) information when they are using an informative tool such as the FDVT.

The main findings of this paper are summarized as follow:

- After using the FDVT 3/4 of the lab experiment participants confirmed that they were surprised by the session revenue reported by the FDVT. This result, together with the wide discrepancy among their answers regarding the revenue they generate, confirms an important lack of awareness of skilled Internet users about the value associated to online personal information.

- More than 85% of the lab experiment participants highly value the FDVT as an appropriate tool to create social awareness about the value of online personal information.

- 1/3 of the beta-testers show a relevant engagement to the FDVT during most part of the 5 months evaluation period in which they have frequently (at least once a month in median) interacted with the FDVT.

- Users are not reluctant to provide personal information for an informative tool such as the FDVT. In particular, 71% of the beta-testers filled all the 3 optional parameters in the FDVT registration process and 95% filled at least two of them.

- The average FDVT revenue estimations just diverge 12% from the numbers reported by Facebook in the 2016 second quarter results report.

In a nutshell, this paper presents the first comprehensive steps towards the increasing demand of creating awareness among Internet users about the economic value generated out of their personal information. The FDVT is based on a novel approach that aims to provide users real-time and personalized feedback of the revenue they generate for Facebook. Finally, this research opens an opportunity to the research community to replicate the proposed approach in other online services.

BACKGROUND

The goal of this section is to briefly describe the business model of FB and how advertisers can easily create tailored advertising campaigns to well-defined audiences through the FB ads campaign manager¹. This brief background will serve the reader to better understand the FDVT operation.

Facebook exploits personal information of the users registered in the platform to offer advertisers the possibility to define advertising campaigns targeting well-defined audiences. An audience is defined as a set of demographic (e.g., location, age, gender), behavioural (e.g., mobile OS used) and users' interest (e.g., food, sports, etc) parameters relevant for an advertiser. Once an advertiser defines the targeted audience Facebook ensures to display the ads of that campaign to users whose profile matches the defined audience. Then Facebook charges the advertisers for the ads displayed to those users. In 2015 FB reported \$17.92B revenue from which more than 95% was obtained through advertising².

We can find two dominant models to charge advertisers in the online advertising market³. In the first one, known as Cost Per Mile (CPM), advertisers are charged based on the number of impressions of their ads. The CPM refers to the price an advertiser has to pay for 1000 impressions of an ad. In the second model, known as Cost Per Click (CPC), advertisers pay for each click of the user on the ad. FB ads campaign manager offers advertisers a reference of the CPC/CPM range (i.e., min, median and max values) that other advertisers have paid in the recent past for the audience they are targeting.

¹<https://www.facebook.com/ads/manager/creation>

²<http://bit.ly/2cu4jnj>

³Note that nowadays there are other models such as the Cost per Action (CPA), Cost per View (CPV) of a video, etc. However, CPC and CPM are the most widely used.

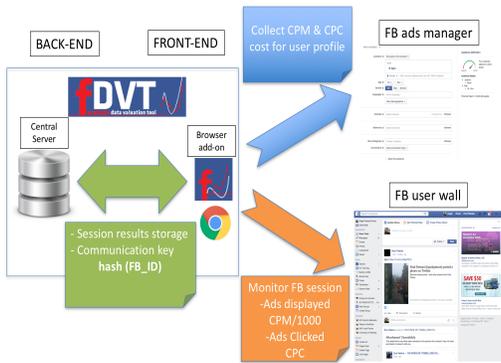


Figure 1. FDVT design



Figure 2. FDVT interface



Figure 3. FDVT registration window



Figure 4. FDVT extension icon

This reference is very important because in the case of FB (and many other adtech players) there is a real-time auction algorithm that decides which ad is displayed to an online user matching an audience among all the advertisers competing for that audience. Simplifying the Facebook auction model the advertiser with the highest CPC/CPM bid will win the auction. Therefore, FB CPC/CPM references for the recent winning bids will allow advertisers to optimize the price they have to pay for each impression or click.

FB CPC and CPM references establish the actual market price of specific audiences and in turn audiences can be linked to user profiles. Hence, if we can construct a more or less simple profile for a specific user in Facebook, we will be able to know what is the actual market value of that user in terms of CPC and CPM at a particular time. Therefore, FB CPC and CPM references are roughly revealing the value of users for FB based on the personal information included in their FB profile. The FDVT uses median CPC and CPM references from FB as the ground to estimate the monetary value users generate based on some profile information of the user (i.e., audience). In order to retrieve CPC and CPM prices associated to a particular audience we developed a software able to automatically query the FB ads manager API following a previous work that already exploited this API [13].

FACEBOOK DATA VALUATION TOOL (FDVT)

The FDVT provides end users with a real-time and personalized estimation of the monetary value they generate for Facebook based on the commercial exploitation of its personal information through tailored advertising. The FDVT is divided into two parts: a front-end running in the end user premises and a central back-end that stores anonymous information associated to FB sessions. Figure 1 depicts a diagram of the FDVT design. Following we describe in detail the FDVT.

Front-end

In this subsection we will explain in detail the FDVT interface and the main functionalities associated to the FDVT front-end. The FDVT front-end is available as a Google Chrome extension and can be downloaded from the Chrome web store⁴.

⁴<http://fdvt.org>

FDVT Interface

Our goal was to create a tool valid for average Internet users. This implies to design a tool easy to install and use with a friendly and intuitive interface. Therefore, the FDVT front-end has been developed as a web-browser extension that: (i) can be installed with one click from an online store very similarly to the way mobile apps or desktop widgets are installed; (ii) users can access the personalized feedback by simply clicking on the web extension while they are browsing in a FB session; (iii) that click will display an interface that informs the user of the monetary value she is generating.

Figure 2 shows a snapshot of the FDVT interface. The information displayed in the interface is: (i) *TOTAL VALUE*, which indicates the revenue generated by the user since she installed the FDVT, (ii) *Session ads*, which refers to the number of ads displayed during the session together with the value generated due to those impressions; (iii) *Ads Clicked*, which indicates the number of ads the user has clicked on during the session together with the value generated due to those clicks; (iv) *Value Generated*, where the user is informed about the revenue she is generating in the current session as well as the accumulated value generated during the current day, the last 7 days and the last 30 days. Even more, the FDVT extension icon incorporates by default a small red box including the accumulated revenue generated by the user as depicted in Figure 4. This allows FDVT users to obtain their overall accumulated revenue without even interacting with the tool. The interface also includes an option to share the overall revenue in the user's FB and Twitter walls.

Registration process

One of the main functionalities implemented in the FDVT front-end is the user registration process. The first time the user clicks on the FDVT browser extension during a FB session the FDVT displays a registration window (depicted in Figure 3) where she is asked to provide: Country, Age+birthday, Gender and Relationship Status. The only compulsory parameter the user has to fill is the Country, because it is the minimum (and obligatory) parameter to define an audience in the ads campaign planner of Facebook. The remaining three parameters are optional. The parameters pro-

vided by the user in the registration are used to define the audience associated to the user profile. To conclude the registration process the user has to obligatorily check-in the following checkboxes: (i) confirm that she has read and accepts the terms of use of the FDVT⁵, (ii) confirm that she has read and accepts the privacy agreement of the FDVT⁶, (iii) confirm that she grants us permission to use the collected data for research purposes. Once the user has completed the registration process the front-end sends the registration profile of the user to the FDVT back-end using an anonymous FDVT user ID. This ID is computed as a hash of the FB user ID.

Real-Time revenue computation

Once a user has completed the registration process every time she opens a FB session the front-end queries the FB ads campaign API⁷. We follow the FB API query structure introduced in [13]. The query includes the demographic parameters provided by the user in the registration process (location, age, gender and relationship status) to configure the audience from which we retrieve the real-time CPC and CPM values associated to the user. The FB API returns a JSON file that includes, among other things, the min, median and max CPC and CPM references for the requested audience. In turn, the front-end notifies the back-end of the median CPM, CPC and the initial timestamp of the new session.

In parallel to the start-up process, the FDVT extension begins to monitor and account for the number of ads displayed during the session and the number of clicks the user performs on those ads. In order to compute an estimation of the real-time session revenue generated by a user we apply the following formula:

$$Session_Value = \frac{CPM}{1000} * n_{ads} + CPC * n_{clicks} \quad (1)$$

where n_{ads} refers to the number of ads displayed in the session and n_{clicks} refers to the number of ads clicks. Every time new ads are displayed or the user clicks on an ad, the session value is updated.

We note that the front-end locally stores some information related to the ads displayed to the user. In particular, the FDVT stores for each ad: (i) FB add id, which is an identifier that FB assigns to each add, (ii) the location of the ad (either news-feed or right side of the wall), (iii) the url associated to the ad that will inform us about the landing page of the ad, and (iv) timestamps associated to users' clicks on ads. The front-end also registers the timestamps associated to the clicks of the user on the FDVT browser extension (i.e., interactions with the FDVT) and the number of posts displayed in the news-feed of the user during the session. This information is also transmitted to the back-end. The reason to store all this information locally is to being able to inform users of the value generated in the current session even in the case the back-end cannot be accessed.

⁵http://www.fdvt.org/terms_of_use

⁶http://www.fdvt.org/privacy_agreement.html

⁷We use a distributed approach in which users grant permission to the FDVT to query the FB API using their FB account.

Back-end

The FDVT back-end was designed to store all the information associated to FB sessions of a user once she has installed the FDVT. This creates a valuable anonymous dataset that registers the following information per FDVT user session: duration, ads displayed and clicked, median CPC and CPM associated to the user audience, revenue generated and interactions of the user with the FDVT extension.

The front-end communicates the back-end all the information locally stored within a session. This communication happens: (i) at the beginning of the session to notify that the user has started a new session, (ii) every 10 minutes after the beginning of the session, (iii) at the end of the session. If the session lasts less than 10 minutes there will not be any intermediate communication. The information is codified in JSON format by the front-end and is stored in a SQL database in the back-end. Every time the front-end of a user notifies the beginning of a new session a PHP process running in the back-end computes the accumulated revenue generated by that user in the last 7 days, last 30 days and since the moment she installed the FDVT and sends that information back to the front-end. The front-end will eventually display this information if the user clicks on the FDVT extension. To compute the accumulated revenue of a user we just need to sum the value of the sessions registered for that user in a specific time window (e.g., last 7 days).

Privacy and Security considerations

The FDVT has been designed as a privacy-preserving tool so that FDVT users cannot be identified with the information stored in the back-end. Towards this end we do not store any Personal Identifiable Information (PII) in the back end. The only personal information stored in the back-end is related to the parameters provided by the user in the registration process. In addition, we note that the FDVT extension only operates when the user browses in the domain facebook.com and does not collect any information from any other domain. In addition, we are currently improving the FDVT to support https to encrypt the communications between the front-end and back-end.

LAB EXPERIMENT: SAVVY USERS' AWARENESS

In this section we aim at answering the first research question posed in the introduction: are Internet skilled users aware of the monetary value associated to their personal information?

Methodology

In order to address the first research question we used a lab-based experiment⁸ with advanced Internet users that have a deep knowledge on how Internet works, i.e., Computer Science and Telecommunication Engineering Bachelor, MSc and PhD students. Our assumption is that if these skilled users do not understand the Internet business model to exploit their personal data or the actual monetary value they generate out of such exploitation, it may suggest that there is lack of awareness in the society (i.e., average Internet users) as well.

⁸The questionnaire used for this experiment is available at <http://fdvt.org/chi2017>

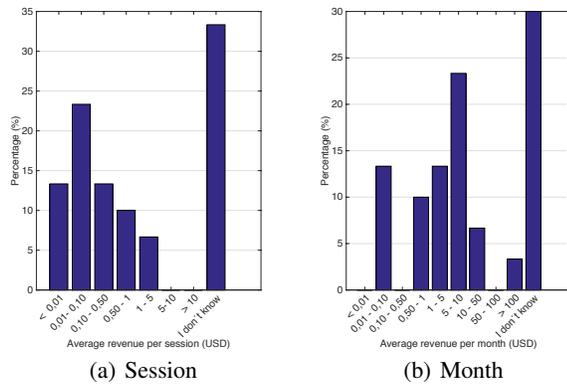


Figure 5. Barplot depicting the distribution of answers provided by the lab experiment participants about the average revenue that they estimate they generate for FB per session and per month.

The lab experiment proceeded in three steps: First, the students were exposed to a survey in which they were asked to: (i) provide some demographic information, (ii) provide some information about their Internet expertise and use of ad blockers, (iii) answer some questions related to the business model of Facebook, and (iv) provide an estimation of the value they think they generate for Facebook in an average session and in a month. Second, we introduced the participants to the FDVT and asked them to login to FB with their FB account in a computer with the FDVT installed, complete the FDVT registration and run a regular session in FB. Third, the users were exposed to another set of questions where they were asked to: (i) indicate whether the value reported by the FDVT associated to their sample FB session was surprising or actually was aligned to their expectation, (ii) provide their opinion about the potential of the FDVT as a tool to create awareness among Internet users of the value associated to online personal information. We note that all the questions were subject to a limited number of predefined answers.

Participants

The lab experiment was carried out during July, August and September 2016. The experiment was completed by 30 students, from which 8 were women and 22 men, with the following age distribution according to the option (i.e., age interval) they selected in the questionnaire: 15 (18-24 years old), 14 (25-34) and 1 (35-44). The participants come from 7 different countries: Spain (23), Iran (2), Ethiopia (1), Italy (1), Sri Lanka (1), United States (1) and Uruguay (1). We note that the participants did not receive any compensation.

Results

We have selected 2 questions (Q1, Q2 below) from the survey to discuss the technical skills of the students, 4 questions (Q3-Q6) to discuss what is the actual awareness of Internet skilled users regarding the revenue they generate for online services out of the exploitation of their online personal information, and 1 question (Q7) to derive the potential of the FDVT as a useful tool to create social awareness according to the participants opinion.

Q1- What is your Internet user level?

At the beginning of the experiment we wanted to know what is the actual Internet level that the participants assign themselves in a scale 1 (basic level) to 5 (expert). We note that users self-assess their Internet level and thus the evaluation is not objective. 90% of the students classified themselves as advanced users either with a level 5 (53.3%) or 4 (36.7%). Therefore, as we intended, most participants could be considered as skilled Internet users.

Q2- Have you installed an ad blocker in your computer?

This question aims at understanding up to what extend the participants are concerned by online advertising and have decided to install and ad blocker in their computer. Interestingly, 2/3 of them have installed an adblocker. This result suggests that most of the users participating in the experiment prefer to avoid ads while browsing in the Internet.

Q3 - How does Facebook earns money?

We asked the students to select at most 2 answers among the 7 available options. Following we add between brackets the number of students selecting each answer: Through ads (27), FB resells data to third parties (14), Through private investment (4), Companies paying a fee to use FB (3), Through public funding (3), Through merchandising (3), Premium Users (1). All students except three of them selected that Facebook earns money through online advertising. This clearly demonstrates that advanced Internet users know that FB exploits personal information for making money through tailored ads. Therefore, they are aware that their personal information generates revenue for online services.

Q4- How much money do you think you generate for Facebook in a standard session? (in USD)

Figure 5(a) shows the distribution of answers across the options chosen by the participants. Interestingly, 1/3 of them, which form the most numerous group, directly recognize that they ignore the answer. All the remaining answers are sparsely distributed across other options. If we use the median session value across the +8000 sessions registered in the FDVT back-end, i.e., \$0.007, as a ground truth reference, only 13% of the users were close to that estimation by choosing the answer <\$0.01. The high discrepancy across students' answers reveals a clear lack of consensus among advanced Internet users about the value they generate in a FB session, which can be translated into a global lack of awareness. We acknowledge that we were expecting this discrepancy since a session is a non-usual time reference for assigning monetary value. That is why we decided to repeat the question using a very standard time metric such as the month.

Q5- How much money do you think you generate for Facebook in average per month? (in USD)

Figure 5(b) presents the distribution of answers for this question. We again observe a high discrepancy among users answers, and again 1/3 of the students directly acknowledge that they do not know the answer. As we will show later in the paper the average revenue per month and user in Facebook is a bit higher than \$1 both according to FB market cap and FDVT estimations. If we use this value as reference, only 23% of the participants provide a close answer either selecting the option

\$1-\$5 or \$0.50-\$1. In this case, the discrepancy across the answers (within a very standard time reference such as one month) reveals a noticeable global lack of awareness among Internet advanced users of the value that personal data generates for one of the most popular online services such as FB.

Q6- Are you surprised by the economic value you have generated in this session?? (in USD)

Although the results discussed so far already reveal a clear lack of knowledge, we wanted to achieve that the participants acknowledged themselves this statement. In order to accomplish that objective we first introduced the FDVT to the students and asked them to use it in a regular session to retrieve the FDVT feedback, to later question them whether they were surprised by the revenue they generated in the sample FB session. A positive answer to this question represents an implicit acknowledgement from the participants about a wrong perception of the monetary value they generate through the commercial exploitation of their personal information. A major part of the students, 73.3%, recognized to be surprised by the result. Interestingly, 1/2 of the users were expecting to generate less revenue and 1/4 thought they were generating more revenue for FB.

Q7- What is the value of the FDVT for creating awareness among society regarding the value associated to their personal online information?

We asked the participants to rank the FDVT in a scale 1 (useless) to 5 (awesome tool) regarding its potential to create awareness in the society about the monetary value of personal information. Most of them agreed that the FDVT has an important potential to achieve that objective since they choose either 4 (50%) or 5 (36.7%) as answer, while the remaining users ranked the FDVT with a value of 3. To complement this question we asked the participants whether they would install the tool after the experiment and 3/4 show their willingness to do so, while the remaining 25% show some uncertainty since they choose as answer *Maybe, I am not sure*.

Overall, the results of the lab experiment provide solid evidences to answer the first research question: Internet advanced users are still far from having a clear knowledge of the monetary value associated to their personal information. This supports the hypothesis that the society is not aware about the value of on-line personal information. This situation urges to create attempts such as the FDVT to try to diminish the lack of awareness so that Internet users begin to know what is the actual revenue they generate for online services out of the commercial exploitation of their personal information. In addition, the Internet savvy users participating in the lab experiment have validated the FDVT as a useful tool to create social awareness.

FDVT FIELD STUDY

In this section we aim at answering the second and third questions posed in the introduction: (i) what is the potential engagement that Internet users may have with data valuation tools? (ii) are users reluctant to provide personal information when they use an informative tool such as the FDVT? In addition, we briefly discuss some relevant data valuation insights derived from the information stored in the FDVT back-end.

Methodology

To answer these questions we required that actual Internet users installed and used the FDVT over a long period of time in order to register their interaction (i.e., clicks) with the FDVT browser extension to measure their engagement. Hence, we had to recruit beta-testers to evaluate the FDVT engagement. The beta-testers were recruited from five main sources: colleagues from the authors' University, collaborators from an European research project in which the authors participate, people that contacted us after the FDVT was featured in several Spanish language news media, people that contacted us after the FDVT was featured as a reference tool in the Mozilla Take The Web Back campaign, and people that contacted us after we presented the tool in several dissemination activities aiming to approach science to society. In all the cases, FDVT beta-testers were users that proactively shown an interest in testing the tool. We note that the FDVT beta-testers group used in the field study differs from the users following the lab experiment since we wanted to get native users interested in the FDVT to avoid an artificial use of the tool. In this line, FDVT beta-testers were neither asked to implement any specific action nor to provide any feedback. In order to be sure we do not take into account lab experiment participants we did not consider in the field study analysis any user installing the FDVT after July 1st.

We have measured the engagement (interest) of the beta-testers to the FDVT through their interactions (i.e., clicks) with the FDVT extension. We remind that the extension incorporates by default the accumulated value generated by the user in a small red box below the FDVT icon (see Figure 4). Hence, all FDVT users are informed about the accumulated revenue they have generated without requiring to interact with the tool. This may discourage some users from obtaining the complete FDVT feedback since knowing the accumulated value may be enough for them. Then, it is reasonable to assume that if a user clicks on the extension is because she shows a high level of interest on the FDVT complete feedback, and by extension on acquiring deep understanding regarding the way she generates revenue for Facebook.

Finally, to answer the third question, we just had to analyze the parameters that each beta-tester filled during the registration process in order to understand whether they are reluctant to provide optional personal information (i.e., gender, age+birthday and relationship status).

Participants

Our field study includes 59 beta-testers from which we monitored their FDVT activity from the moment they installed the tool until July 31st. Most beta-testers installed the tool during March 2016 which means our engagement analysis covers a period of 5 months. Beta-testers are divided into 10 women, 48 men and 1 user that did not specify his/her gender. Beta-testers' age ranges from 19 to 57 with a median age of 30. Finally, the beta-testers come from 19 different countries (according to the information they provided in the registration process): Spain (27), Switzerland (4), Germany (4), France (3), Greece (3), Australia (2), Belgium (2), Brazil (2), United States (2), Andorra (1), Afghanistan (1), Argentina

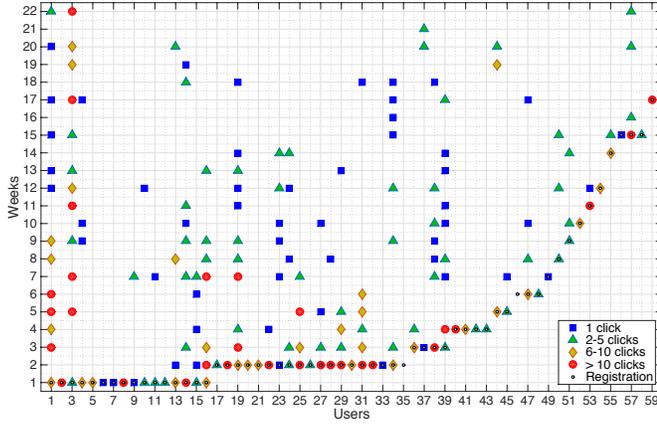


Figure 6. Number of clicks of beta-testers on the FDVT browser extension per user (x-axis) and week (y-axis) from March 1st to July 31st.

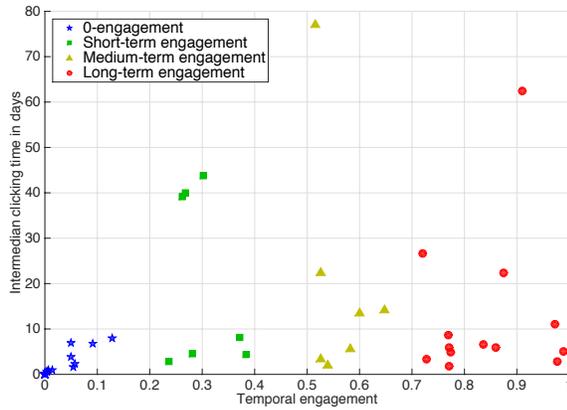


Figure 7. Scatter plot showing the engagement of FDVT beta-testers and classifying them into one of the four engagement clusters. The x-axis refers to the *Temporal Engagement* and the y-axis to the *intermedian Clicking Time* in days.

(1), Ecuador (1), Ethiopia (1), India (1), Italy (1), Mexico (1), UK (1), and Venezuela (1). The major presence of Spanish users is due to the dissemination of the FDVT in some Spanish media. We note that the participants did not receive any compensation.

Results: Users' engagement

We find 2 users that never clicked on the FDVT extension and 5 that just clicked once. In contrast, there are 30 users that have clicked at least 10 times on the extension, and 6 very active users that have clicked more than 50 times. In average the beta-testers interacted 20 times with the FDVT, being the median 10 clicks. Although, these results reveal a high degree of interest from the beta-testers, we need to analyze the temporal pattern of those clicks in order to understand whether the FDVT engages users over the time or not.

Figure 6 shows the number of clicks per beta-tester and week. We find an interesting discrepancy across beta-testers behaviour. We can find some users such as 2, 5, 7, 8, 18 or 33 that just interacted with the FDVT during the first week after installing it, but they did not click on the FDVT icon

Group	Time in Facebook (hours)
0-engagement	0.27
Short-term engagement	55.8
Medium-term engagement	86.95
Long-term engagement	170.5

Table 1. Median time spent in Facebook during the engagement evaluation period in each of the four groups obtained after applying the k-means clustering algorithm using beta-testers *Temporal Engagement*.

anymore. We could conclude that these users present a negligible engagement. Contrary, there are users such as 1, 3, 14, 19, 34, 38 or 57 that show a long-term engagement because they interact frequently with the tool since the moment they installed it. In order to carry out a more pragmatic analysis we clustered the beta-testers into different groups according to their engagement level. For this purpose, we run the k-means clustering algorithm [14] using as clustering parameter a metric referred to as *Temporal Engagement*. Given a user A we compute her *Temporal Engagement* as the time passed from the moment user A installed the FDVT until the last click of user A on the FDVT browser extension divided by the total time the FDVT has been running on user's A computer (i.e., from the FDVT installation until July 31st). The closer the *Temporal Engagement* is to 1 the more recent has been the last click, and thus we can consider the user is still engaged since she recently interacted with the FDVT. Contrary, a *Temporal Engagement* close to 0 indicates that the user only interacted with the FDVT during the first days (or weeks) after she installed the tool.

We forced the k-means clustering algorithm to classify users into four different groups (according to their temporal engagement) that we define as: 0 engagement, Short-term engagement, Medium-term engagement and Long-term engagement. Table 1 shows the median time that users in each cluster spent in Facebook during the evaluation period, and Figure 7 shows a scatter plot of the intermedian clicking time (y axis) Vs. the temporal engagement (x axis) where each point refers to one beta-tester. Following, we discuss the engagement associated to each cluster:

- *0-engagement*: This group is formed by 32 users with a *Temporal Engagement* < 0.2 . These users never clicked on the FDVT or just clicked few times in the days following the FDVT installation. The median time spent in FB (0.27 hours) reported in Table 1 for this group demonstrates that actually the users in this group are not engaged to Facebook either (at least not via Google Chrome). Therefore, it seems that these users installed the tool to simply test it but they are not very active FB users, thus they did not become interested on obtaining the FDVT feedback over the time.

- *Short-term engagement*: It is formed by 7 users that only interacted with the FDVT during the first half of the evaluation period. The users in this group are engaged to the FDVT during a short period after they installed the tool, but that interest disappears quickly. Half of the users in this group present an intermedian-clicking time below 10 days that depicts a relevant interest during the short interaction period, whereas the remaining ones present long intervals beyond 40 days between two consecutive FDVT clicks. Finally, it is important to note that although the users in this group spent a consider-

able amount of time in FB during the evaluation period, 55 hours in median, they are still far from the time spent by the users in the next two groups. Therefore, this demonstrates a rather moderate FB engagement.

- *Medium-term engagement*: This group is formed by 7 users that present a *Temporal Engagement* between 0.5 and 0.7. Roughly speaking these users are showing a significant engagement since all of them have shown their interest in retrieving the FDVT feedback during the second half of the evaluation period. Therefore, it is very likely that many of these users are still engaged and will eventually interact again with the FDVT at some point in the future. The intermedian-clicking time reveals that most of the users in this group retrieves the FDVT feedback at least once a month, except one particular case showing a value close to 80 days. Finally, the users in this group also show a significant FB engagement since they spent in median 87 hours in FB during the evaluation period.

- *Long-term engagement*: This group is formed by 13 users that have been engaged at least 70% of the time since they installed the FDVT. These users have recently clicked on the FDVT extension and in all the cases except one they execute that action at least once a month, and in most of the cases once a week. Therefore, this group is showing a high degree of interest in the FDVT. This interest is aligned to the time they spent in FB that multiplies by 2×, 3× and 600× the use of Facebook from the users in the Medium-term, Short-term and 0-engagement groups, respectively.

Overall, the results show that 1/3 of the beta-testers have demonstrated a relevant engagement to the FDVT, and in particular 1/5 a long-term engagement. In addition, the FDVT engagement is highly correlated to the FB engagement measured as the time users spent in FB during the evaluation period. This result suggests that a data valuation tool for an online service will mostly engage users that are in turn highly engaged to the service. Hence, it seems that data valuation tools such as the FDVT should focus on attracting active users in the system they are targeting. In the case of FB, the company has recently reported 1.71 billion monthly active users. This number envisions an optimistic future for the FDVT that could potentially attract a large number of users to install it.

Results: Users' concerns to provide personal information

The results derived from the registration process show that 95% of the registered users filled at least 2 optional parameters, and 71% filled all the requested parameters. Surprisingly, only 1 user did not provide any optional parameter. Going a bit deeper into the results only 1 user rejected to fill his/her gender, while 9 and 11 users did not fill the age+birthday and the relationship status, respectively. These results suggest that users tend to trust the FDVT and accept the trade-off of providing some personal information in exchange of knowing the monetary value associated to the commercial exploitation of their personal data.

Results: Data valuation insights

The revenue generated by a user depends on three factors: (i) the time she spends in Facebook which increases the opportu-

nity to receive more ads, (ii) the number of clicks on ads, and (iii) the price (i.e., CPM, CPC) advertisers are willing to pay for the audience matching the user profile. The average and median revenue generated by the beta-testers during the field study were \$4.9 (STD=\$11.9) and \$0.3 (IQR=\$3.95), respectively. Similarly, the average and median time spent in FB were 127.3 hours (STD=293 hours) and 6.7 hours (IQR=90.9 hours). The high discrepancy between median and average values denotes that beta-testers present a heterogeneous FB activity and thus the revenue they generate for FB is quite different. The results show that those users spending more time in FB tend to generate more revenue as the high pearson correlation (i.e., 0.68) between these two parameters demonstrates. A very important aspect that increases a lot the revenue users generate for FB is the number of ads they click on. Using as reference the average CPC and CPM prices of beta-testers we find that an ad click (i.e., CPC) generates 170× more value than an ad impression (i.e., CPM/1000). Only 21 beta-testers clicked at least once on an ad. This group generated almost 2.5× more revenue in average than the group formed by the users that never clicked on an ad (\$7.8 Vs \$3.2, respectively). Finally, the location of the users has an important impact on their associated CPC and CPM, and thus on the potential revenue they generate for FB. For instance, the average CPM of Australian, European, US, Asian, Latin American and African users within the beta-testers is \$420, \$279, \$193, \$108, \$101 and \$78, respectively. If we take as example European Vs. Latin American users the CPM difference is roughly 3×. This means that an European user would generate the same revenue than a Latin American user visualizing one third of the ads.

FDVT ACCURACY ASSESSMENT

In order to assess the quality of the FDVT we analyze whether the estimations it provides are aligned to the revenue reported by Facebook. To do that we have first obtained the average quarterly revenue that a user generates according to the FB results for the 2nd quarter of 2016⁹, which are the most recent ones at the time of writing this paper. In that period, Facebook reports \$6.239B of revenue and 1.71 billion monthly active users. By simply dividing both quantities we obtain that the average revenue generated per user in the referred period is \$3.65. In parallel, we have estimated what is the average quarterly revenue generated per user based on the information stored for the 59 beta-testers using the FDVT. To this end, we compute for each user the average revenue generated per week and multiply it by 13 weeks forming a quarter, which offers an average quarterly revenue for each beta-tester. In turn, we obtain the average quarterly revenue across the 59 beta-testers. The average quarterly revenue per user based on the FDVT reported values is \$3.21. Therefore, the FDVT underestimates the actual revenue per user in only 12%. Although the methodology employed to measure the FDVT accuracy is a balk-park approach, it is the only existing ground-truth information we can currently use for validation.

⁹<http://bit.ly/2aeqFSR>

RELATED WORK

We can find a large body of literature studying the value of information assets and privacy from a macroscopic economic perspective [1, 18, 23, 24]. However, it has been only recently when researchers have addressed the question of what is the monetary value of personal information from a microscopic point of view in the context of online services [16]. The most adopted methodology to answer that question has been retrieving directly from users through survey, interviews, economics experiments, etc, the value they assign to personal data. Particularly, most of the authors have used the contingent valuation method widely applied in economics and marketing research [15, 7, 25, 21]. This methodology measures users willingness-to-pay (WTP) to protect/recover their personal information and/or their willingness-to-accept (WTA) to sell their personal information and apply different mechanisms (e.g., conjoint analysis [11, 17]) to conclude the monetary value of some particular aspects of users personal information [19, 22, 3, 10, 4, 8, 2]. The main drawback of this methodology is that it relies on users estimations to define the monetary value of personal information, however there exists already well-defined market values for the personal data value. Therefore, this methodology is useful to understand the perception of Internet users regarding the value of their personal information, but it is useless to inform Internet users of the actual monetary value associated to their personal information. In contrast to this methodology, the FDVT directly informs Facebook users of the revenue they generate according to market prices. We note that there are two works applying this methodology that, similarly to the FDVT, aims to retrieve the value that users assign to their own Facebook profiles [22, 3]. The results depict a very large discrepancy among users valuation. This discrepancy is aligned to the one observed in our lab experiment and reinforces our conclusion regarding the lack of awareness of Internet users about the monetary value associated to their personal information.

A second methodology proposes to use the aggregated market cap (revenues, net income, etc) of companies exploiting personal information to quantify the monetary value of personal data records (e.g., dividing the revenue of a company by the number of active users in order to get the average monetary value per user) [9]. Basically, they apply the same computation we have used to obtain the average revenue generated per user relying on Facebook results for the 2nd quarter of 2016. This methodology can be used to compute an average value per user common to all the users in the system. However, the revenue generated by the FDVT beta-testers in our field experiment denotes that there is a considerable heterogeneity among the revenue generated by different users. Therefore, the referred methodology cannot provide personalized and real-time revenue estimations, as it is the case of the FDVT.

More closely to our work, in [5] the authors analyse the prices that advertisers bid to display ads using as reference 100 users. To this end they exploit a vulnerability of the Real-Time Bidding (RTB) that was just present in a limited number of ads delivery. Therefore, although the experiment was interesting to get some understanding on advertisers bidding on real users, it got access to very little information related to some few ads

during the user browsing. This invalidates this methodology to provide the actual revenue generated by the user. Instead the FDVT approach is valid to generate revenue estimations for all the ads delivered/clicked during a Facebook session, which allows to provide a complete estimation of the actual revenue each user generates for Facebook.

There are two previous works in the literature using the FB Ads Campaign API [13][20]. Liu et al. [13] quantitatively analyze the bidding prices available through the FB API per country and for different audiences over the time. However, they are monitoring CPM prices of global audiences without mapping them to real FB users. Therefore, in contrast to the FDVT they are not looking at the revenue generated per users but analyzing how CPM prices changes based on time and location. Authors in [20] try to infer the value of Facebook users. The authors generate a model to reflect how FB users' activity (e.g., likes, shares) is propagated to friends together with a second simplistic model that guesses the number of ad impressions received per user. To validate their work the authors rely on a dataset from 2009 that only includes users from New Orleans and a second dataset with CPM and CPC prices from 2014. However, they lack an actual ground truth because they neither know the actual number of ads impressions displayed to each user nor the ads clicks performed by each user. Due to these limitations, they do not provide the overall revenue generated per user but just a comparative value among users assigning the value 1 to the user that according to their model generates more money. In addition, similarly to the other methodologies, this work uses a static dataset that is useless to generate real-time information. In contrast to this work the FDVT adopts a real-time approach that measures the revenue that users generate while browsing in Facebook according to real market prices references paid by advertisers in Facebook.

To the best of our knowledge the FDVT is the first tool that provides a real-time and personalized feedback about the revenue users generate for an online service such as FB.

DISCUSSION

Implications of FDVT estimations on users data valuation perception

The long term goal of this research is to create awareness among average Internet users about the monetary value of their personal information using Facebook as reference. We acknowledge upfront that the FDVT only provides estimations, since we do not know what is the actual price an advertiser has paid to display and ad or gather a click from a user. Even in the case the error of FDVT estimations is high they would still be relevant. For instance, let us assume one case in which the FDVT estimates that a user has generated \$1 per month, even if the FDVT is incurring in an error of $5\times$ and the user has actually generated \$5 for Facebook, she will still globally understand that she is not generating hundreds of dollars per month out of her activity in Facebook. Then, FDVT estimations will be in most cases informative enough to get at least a rough knowledge of the value associated to the user's personal information.

Data valuation impact on privacy decisions

One of the factors that Internet users may consider when making privacy decisions is the economic benefit of the company that will exploit that information. This will allow users to evaluate the trade-off between the added value of the service and the economic benefit extracted by that company out of the commercial exploitation of personal data. Therefore, creating awareness using data valuation tools such as the FDVT becomes an important element to allow Internet users making better informed decisions around privacy.

FDVT Interface improvement

The goal of this research is to create a simple interface with few information highlighting at the top the accumulated revenue generated by the user. Although we could have included in the interface more detailed information on how the revenue estimation is obtained, we believe that this could have an overwhelming effect for many users that may not be interested in detailed information. Therefore, the FDVT leaves open an interesting HCI challenge regarding how the interface could be improved to not only inform users about the generated revenue but also: (i) let users understand in a simple way how that revenue was estimated, and (ii) let users understand the potential sources of discrepancy with the actual revenue generation.

FDVT Limitations

First, the FDVT is providing estimations of the actual value the user generates for Facebook based on the median CPM and CPC values reported by FB for the audience matching the user registration profile in the FDVT. We do not know whether the demographic attributes registered in the FDVT are the same used in the actual FB profile of the user, then if the user provides us fake attributes she will be receiving revenue estimations related to the user profile she has registered in the FDVT.

Second, the FDVT is currently only available for desktops and laptops through a Google Chrome extension. We are aware that some FDVT users may access FB from mobile devices in addition to their laptops/desktops. Also, a user may access FB from different computers and just use Google chrome in one of them. In those cases the FDVT will be only providing partial information regarding the value those users are generating for FB. Then, for those users the FDVT will be actually generating a lower bound estimation of the actual revenue they generate.

Third, a FB user profile matches a large number of audiences since a user can be targeted based on her demographic parameters but also based on her interests, the mobile device she uses, etc. However, the FDVT only uses the demographic parameters provided in the registration process to create the audience associated to the user, and retrieve the CPM and CPC values associated to that specific audience to compute the revenue estimations generated by the user for FB. It is likely that FB may target users based on other parameters, e.g., behavioral data, in addition to the parameters used by the FDVT. In those cases the FDVT will be providing inaccurate estimations that may impact the perception of the user on her own

data valuation since those estimations can overestimate or underestimate the actual revenue she is generating. However, as we have discussed at the beginning of this section the received feedback will be still informative enough since the user will get a rough global estimation of the revenue she is generating.

Fourth, the conclusions extracted from this paper are derived from a field experiment including only 59 beta-testers. Those 59 beta-testers cannot be considered as a representative sample of the whole FB ecosystem that is formed by more than 1.7B users. Also those beta-testers are users showing a proactive interest in the FDVT that may imply that the engagement ratio reported in this paper overestimates the actual engagement ratio once the FDVT is publicly launched and installed by hundreds of users. Similarly, the reported 95% of beta-testers that were not reluctant to provide optional personal information during the registration process may decrease once the FDVT attracts a large number of real users installations. In a nutshell, the extracted conclusions in this paper are limited to the beta-testers sample participating in the field experiment and cannot be extrapolated to the whole Facebook ecosystem.

CONCLUSIONS

We align to the demand of the OECD, the European Commission and private initiatives like the Data Transparency Lab regarding the necessity of providing Internet users with data valuation tools that allow them to understand what is the actual monetary value of their personal information. We believe that this is a natural pedagogic way to introduce average Internet users into more complex privacy concepts and help to construct a global social demand for more transparent online services with respect to the management of personal information. In this line, this paper presents the first data valuation tool that provides Internet users with a personalized and real-time estimation of the revenue they generate for Facebook out of the commercial exploitation of their personal information. Relying on the FDVT we conclude that: (i) Internet users are far from understanding what is the actual monetary value of their personal information, (ii) users that are very active in an online service are very likely to engage to data valuation tools that inform them of the revenue they generate for that service, (iii) data valuation tools such as the FDVT are worthy attempts to let Internet users understand that their personal information has an associated value that generates revenue for online services. We note that the FDVT has received a noticeable attention after its public release in Oct. 1st 2016. Currently, the FDVT has been installed more than 3000 times by the end of 2016.

ACKNOWLEDGEMENTS

This work has been partially supported by the European Commission through the H2020 TYPES (653449) and ReCRED (653417) projects, the Ministerio de Economía y Competitividad of Spain through the project BigDatAAM (FIS2013-47532-C3-3-P), the Ministerio de Economía y Competitividad of Spain and the European Social Fund through the Ramón y Cajal grant (RYC-2015-17732) and the Data Transparency Lab.

REFERENCES

1. Alessandro Acquisti, Curtis R Taylor, and Liad Wagman. 2016. The economics of privacy. *Available at SSRN 2580411* (2016).
2. George Loewenstein Alessandro Acquisti, Leslie K. John. 2013. What Is Privacy Worth? *The Journal of Legal Studies* 42, 2 (2013), 249–274. <http://www.jstor.org/stable/10.1086/671754>
3. Christine Bauer, Jana Korunovska, and Sarah Spiekermann. 2012. On the value of information-what Facebook users are willing to pay. *ECIS 2012 proceedings* (2012).
4. Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. 2013. Your Browsing Behavior for a Big Mac: Economics of Personal Information Online. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 189–200.
5. Claude Castelluccia, Lukasz Olejnik, and Tran Minh-Dung. 2014. Selling Off Privacy at Auction. In *Network and Distributed System Security Symposium (NDSS) (NDSS)*. ISOC, San Diego, California, United States. <https://hal.inria.fr/hal-01087557>
6. European Commission. 2014. Secure societies - Protecting freedom and security of Europe and its citizens. Topic DS1 - Privacy. (2014). http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-security_en.pdf.
7. Don L Coursey, John L Hovis, and William D Schulze. 1987. The disparity between willingness to accept and willingness to pay measures of value. *The Quarterly Journal of Economics* 102, 3 (1987), 679–690.
8. Dan Cvreck, Marek Kumpost, Vashek Matyas, and George Danezis. 2006. A Study on the Value of Location Privacy. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society (WPES '06)*. ACM, New York, NY, USA, 109–118.
9. Claudio Feijóo, José Luis Gómez-Barroso, and Peter Voigt. 2014. Exploring the economic value of personal information from firms? financial statements. *International Journal of Information Management* 34, 2 (2014), 248–256.
10. Jens Grossklags and Alessandro Acquisti. 2007. When 25 Cents is Too Much: An Experiment on Willingness-To-Sell and Willingness-To-Protect Personal Information.. In *WEIS*.
11. Anders Gustafsson, Andreas Herrmann, and Frank Huber. 2013. *Conjoint measurement: Methods and applications*. Springer Science & Business Media.
12. Data Transparency Lab. 2015. (2015). <http://datatransparencylab.org/>.
13. Yabing Liu, Chloe Kliman-Silver, Robert Bell, Balachander Krishnamurthy, and Alan Mislove. 2014. Measurement and Analysis of OSN Ad Auctions. In *Proceedings of the Second ACM Conference on Online Social Networks (COSN '14)*. ACM, New York, NY, USA, 139–150. DOI: <http://dx.doi.org/10.1145/2660460.2660475>
14. Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE transactions on information theory* 28, 2 (1982), 129–137.
15. Robert Cameron Mitchell and Richard T Carson. 1989. *Using surveys to value public goods: the contingent valuation method*. Resources for the Future.
16. OECD. 2013. Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value. *OECD Digital Economy Papers* 222 (2013). <http://dx.doi.org/10.1787/5k486qtx1dmq-en>
17. V. Srinivasan Paul E. Green. 1990. Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. *Journal of Marketing* 54, 4 (1990), 3–19. <http://www.jstor.org/stable/1251756>
18. Richard A. Posner. 1981. The Economics of Privacy. *The American Economic Review* 71, 2 (1981), 405–409. <http://www.jstor.org/stable/1815754>
19. Dimitris Potoglou, Sunil Patil, Covadonga Gijn, Juan Francisco Palacios, and Claudio. Feijo. 2013. The Value Of Personal Information Online: Results From Three Stated Preference Discrete Choice Experiments In The UK. In *ECIS 2013*.
20. Diego Saez-Trumper, Yabing Liu, Ricardo Baeza-Yates, Balachander Krishnamurthy, and Alan Mislove. 2014. Beyond CPM and CPC: Determining the Value of Users on OSNs. In *Proceedings of the Second ACM Conference on Online Social Networks (COSN '14)*. ACM, New York, NY, USA, 161–168.
21. Jason F Shogren, Seung Y Shin, Dermot J Hayes, and James B Kliebenstein. 1994. Resolving differences in willingness to pay and willingness to accept. *The American Economic Review* (1994), 255–270.
22. Sarah Spiekermann and Jana Korunovska. 2016. Towards a value theory for personal data. *Journal of Information Technology* (2016). DOI: <http://dx.doi.org/10.1057/jit.2016.4>
23. George J. Stigler. 1980. An Introduction to Privacy in Economics and Politics. *The Journal of Legal Studies* 9, 4 (1980), 623–644. <http://www.jstor.org/stable/724174>
24. Hal R Varian. 1996. Economic aspects of personal privacy. *Privacy and Self-regulation in the Information Age* (1996).
25. Jinhua Zhao and Catherine L Kling. 2001. A new explanation for the WTP/WTA disparity. *Economics Letters* 73, 3 (2001), 293–300.