Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes

José González Cabañas, Ángel Cuevas, and Rubén Cuevas,
Department of Telematic Engineering, Universidad Carlos III de Madrid

https://www.usenix.org/conference/usenixsecurity18/presentation/cabanas

This paper is included in the Proceedings of the 27th USENIX Security Symposium.

August 15–17, 2018 • Baltimore, MD, USA

ISBN 978-1-931971-46-1

Open access to the Proceedings of the 27th USENIX Security Symposium is sponsored by USENIX.

Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes

José González Cabañas, Ángel Cuevas, and Rubén Cuevas
Department of Telematic Engineering
Universidad Carlos III de Madrid
{jgcabana, acrumin, rcuevas}@it.uc3m.es

Abstract

The recent European General Data Protection Regulation (GDPR) restricts the processing and exploitation of some categories of personal data (health, political orientation, sexual preferences, religious beliefs, ethnic origin, etc.) due to the privacy risks that may result from malicious use of such information. The GDPR refers to these categories as sensitive personal data. This paper quantifies the portion of Facebook users in the European Union (EU) who were labeled with interests linked to potentially sensitive personal data in the period prior to when GDPR went into effect. The results of our study suggest that Facebook labels 73% EU users with potential sensitive interests. This corresponds to 40% of the overall EU population. We also estimate that a malicious third party could unveil the identity of Facebook users that have been assigned a potentially sensitive interest at a cost as low as €0.015 per user. Finally, we propose and implement a web browser extension to inform Facebook users of the potentially sensitive interests Facebook has assigned them.

1 Introduction

The citizens of the European Union (EU) have demonstrated serious concerns regarding the management of personal information by online services. The 2015 Eurobarometer about data protection [21] reveals that: 63% of EU citizens do not trust online businesses, more than half do not like providing personal information in return for free services, and 53% do not like that Internet companies use their personal information in tailored advertising. The EU reacted to citizens' concerns with the approval of the General Data Protection Regulation (GDPR) [8], which defines a new regulatory framework for the management of personal information. EU member states were given until May 2018 to incorporate it into their national legislation.

The GDPR (and previous EU national data protection laws) defines some categories of personal data as sensitive and prohibits processing them with limited exceptions (e.g., the user provides explicit consent to process that data for a specific purpose). These categories of data are referred to as *"Specially Protected Data"*, *"Special Categories of Personal Data"* or *"Sensitive Data"*. In particular, the GDPR defines as sensitive personal data: *"data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation"*.

Due to the legal, ethical and privacy implications of processing sensitive personal data, it is important to know whether online services are commercially exploiting such sensitive information. If so, it is also essential to measure the portion of users/citizens who may be affected by the exploitation of their sensitive personal data. In this paper, we address these crucial questions focusing on *online advertising*, which represents the most important source of revenue for most online services. In particular, we consider Facebook (FB), whose online advertising platform is second only to Google in terms of revenue [2].

Facebook labels users with so-called ad preferences, which represent potential interests of users. FB assigns users different ad preferences based on their online activity within this social network and on third-party websites tracked by FB. Advertisers running ad campaigns can target groups of users assigned to a particular ad preference (e.g., target FB users interested in *"Starbucks"*). Some of these ad preferences suggest political opinions, sexual orientation, personal health, and other potentially sensitive attributes. In fact, an author of this paper received the ad shown in Figure 1 (left side). The author had not explicitly defined his sexual orientation, but he discovered that FB had assigned him the *"Homosexual-*

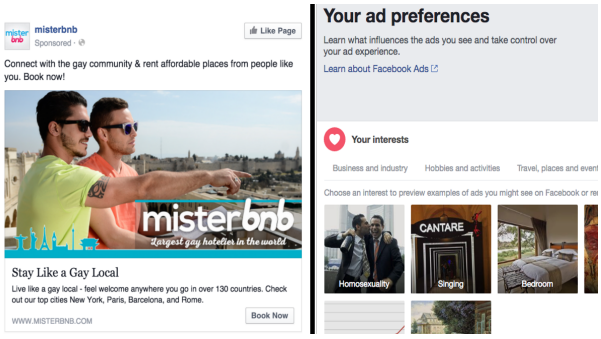

Figure 1: Snapshot of an ad received by one of the authors of this paper & ad preference list showing that FB inferred this person was interested in *Homosexuality*.

ity" ad preference (see Figure 1 right side). Our data suggests that similar assignment of potentially sensitive ad preferences occurs much more broadly. For example, landing pages associated with ads received by FB users in our study include: *iboesterreich.at* (political), *gay-dominante.com* (sexuality), *elpartoestuyo.com* (health).

This illustrates that FB may be actually processing sensitive personal information, which is now prohibited under the EU GDPR without explicit consent and also under some national data protection regulations in Europe. Recently, the Spanish Data Protection Agency (DPA) fined FB €1.2M for violating the Spanish data protection regulation [6]. The Spanish DPA argued that FB *"collects, stores and uses data, including specially protected data, for advertising purposes without obtaining consent."*

Motivated by these events and the enactment of the GDPR in the European Union, this paper examines Facebook's use of potentially sensitive data through January 2018, only months before the GDPR became enforceable. The main goal of this paper is *quantifying the portion of EU citizens and FB users that may have been assigned ad preferences linked to potentially sensitive personal data*. We leave analysis of Facebook data practices following the May 25, 2018 GDPR effective date (when violations could be enforceable) to future work.

To achieve our goal we analyze more than 5.5M ad preferences (126K unique) assigned to more than 4.5K FB users who have installed the Data Valuation Tool for Facebook Users (FDVT) browser extension [12]. The reason for using ad preferences assigned to FDVT users is that we can prove the ad preferences considered in our study have been indeed assigned to real users.

The first contribution of this paper is a methodology that combines natural language processing techniques and manual classification conducted by 12 panelists to obtain those ad preferences in our dataset potentially linked to sensitive personal data. These ad preferences may be used to reveal: ethnic or racial origin, political opinions, religious beliefs, health information or sexual orientation. For instance, the ad preferences *"Homosexuality"* and *"Communism"* may reveal the sexual orientation and the political preference of a user, respectively.

Once we have identified the list of potentially sensitive ad preferences, we use it to query the FB Ads Manager in order to obtain the number of FB users and citizens exposed to these ad preferences in the whole EU as well as in each one of its member states. This quantification is our second contribution, which accomplishes the main goal of the paper.

Finally, after illustrating privacy and ethics risks derived from the exploitation of these FB ad preferences, we present an extension of the FDVT that informs users of the potentially sensitive ad preferences FB has assigned them. This is the last contribution of this paper.

Our research leads to the following main insights:

- We have identified 2092 (1.66%) potentially sensitive ad preferences out of the 126k present in our dataset.
- FB assigns on average 16 potentially sensitive ad preferences to FDVT users.
- More than 73% of EU FB users, which corresponds to 40% of EU citizens, are labeled with at least one of the Top 500 (i.e., most popular) potentially sensitive ad preferences from our dataset.
- Women have a significantly higher exposure than men to potentially sensitive ad preferences. Similarly, The Early Adulthood group (20-39 years old) has the highest exposure of any age group.
- We perform a ball-park estimation that suggests that unveiling the identity of FB users labeled with potentially sensitive ad preferences may be as cheap as €0.015 per user.

2 Background

2.1 Facebook Ads Manager

Advertisers configure their ads campaigns through the Facebook (FB) Ads Manager.[1] It allows advertisers to define the audience (i.e., user profile) they want to target with their advertising campaigns. It can be accessed through either a dashboard or an API. The FB Ads Manager offers advertisers a wide range of configuration parameters such as (but not limited to): *location* (country, region, city, zip code, etc.), *demographic parameters* (gender, age, language, etc.), *behaviors* (mobile device, OS and/or web browser used, traveling frequency, etc.), and *interests* (sports, food, cars, beauty, etc.).

The *interest* parameter is the most relevant for our work. It includes hundreds of thousands of possibilities

[1] `https://www.facebook.com/ads/manager`

capturing users' interest of any type. These interests are organized in a hierarchical structure with several levels. The first level is formed by 14 categories.[2] In addition to the interests included in this hierarchy, the FB Ads Manager offers a *Detailed Targeting* search bar where users can type any free text and it suggests interests linked to such text. In this paper, we leverage the *interest* parameter to identify potential sensitive interests.

Advertisers can configure their target audiences based on any combination of the described parameters. An example of an audience could be *"Users living in Italy, ranging between 30 and 40 years old, male and interested in Fast Food"*.

Finally, the FB Ads Manager provides detailed information about the configured audience. The most relevant parameter for our paper is the *Potential Reach* that reports the number of registered FB users matching the defined audience.

2.2 Facebook ad preferences

FB assigns to each user a set of ad preferences, i.e., a set of interests, derived from the data and activity of the user on FB and external websites, apps and online services where FB is present. These ad preferences are indeed the interests offered to advertisers in the FB Ads Manager to configure their audiences.[3] Therefore, if a user is assigned *"Watches"* within her list of ad preferences, she will be a potential target of any FB advertising campaign configured to reach users interested in watches.

Any user can access and edit (add or remove) her ad preferences,[4] but we suspect that few users are aware of this option. When a user positions the mouse over a specific ad preference item, a pop-up indicates why the user has been assigned this ad preference. By examining 5.5M ad preferences assigned to FDVT users (see Subsection 2.3), we have found 6 reasons for the assignment of ad preferences: (i) *This is a preference you added*, (ii) *You have this preference because we think it may be relevant to you based on what you do on Facebook, such as pages you've liked or ads you've clicked*, (iii) *You have this preference because you clicked on an ad related to...*, (iv) *You have this preference because you installed the app...*, (v) *You have this preference because you liked a Page related to...*, (vi) *You have this preference because of comments, posts, shares or reactions you made related to...*

[2] Business and industry, Education, Family and relationships, Fitness and wellness, Food and drink, Hobbies and activities, Lifestyle and culture, News and entertainment, People, Shopping and fashion, Sports and outdoors, Technology, Travel places and events, Empty.

[3] Given that interests and ad preferences refer to the same thing, we use these two terms interchangeably in the rest of the paper

[4] Access and edit ad preference list: `https://facebook.com/ads/preferences/edit`

2.3 FDVT

The *Data Valuation Tool for Facebook Users (FDVT)* [12] is a web browser extension currently available for Google Chrome[5] and Mozilla Firefox.[6] It provides FB users with a real-time estimation of the revenue they are generating for Facebook according to their profile and the number of ads they see and click during a Facebook session. More than 6K users have installed the FDVT between its public release in October 2016 and February 2018. The FDVT collects (among other data) the ad preferences FB assigns to the user. We leverage this information to identify potentially sensitive ad preferences assigned to users that have installed the FDVT.

3 Legal considerations

3.1 General Data Protection Regulation

The EU General Data Protection Regulation (GDPR) [8] entered into force in May 2018 and is the reference data protection regulation in all 28 EU countries. The GDPR includes an article that regulates the use of *Sensitive Personal Data*. Article 9 is entitled *"Processing of special categories of personal data"* and states in its first paragraph: *"Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited"*.

After enumerating these particular prohibitions, the GDPR introduces ten exceptions to them (see Appendix A) for which the paragraph 1 of the article shall not apply. To the best of our knowledge none of these exemptions for processing sensitive personal data seem to apply to the case of FB ad preferences. Therefore, labeling FB users with ad preferences associated with sensitive personal data may contravene Article 9 of the GDPR.

3.2 Facebook fined in Spain

In September 2017 the Spanish Data Protection Agency (AEPD) fined Facebook €1.2M for violating the Spanish implementation of the EU data protection Directive 95/46EC [1] preceding the GDPR. In the fine's resolution [6] the AEPD claims that FB collects, stores and processes sensitive personal data for advertising purposes without obtaining consent from users. More details about the AEPD resolution are provided in Appendix B.

[5] `https://chrome.google.com/webstore/detail/fdvt-social-network-data/blednbbpnnambjaefhlocghajeohlhmh`

[6] `https://addons.mozilla.org/firefox/addon/fdvt`

The AEPD states that the use of sensitive data for advertising purposes through the assignment of ad preferences to users by FB violated the Spanish data protection regulation (and perhaps other EU member states' regulations which implemented into their national laws the EU data protection Directive 95/46EC [1], recently replaced by the GDPR).

3.3 Facebook terms of service

We have carefully reviewed FB's terms and policies. Although we are not attorneys, we found neither a clear disclosure to EU users that FB processes and stores sensitive personal data specifically nor a place where users can provide consent. To the best of our knowledge, both are required under GDPR. Furthermore, we have not found any general prohibition by FB on advertisers seeking to target ads based on sensitive personal data. More details about the analysis of FB terms of service are provided in Appendix C.

4 Dataset

To uncover potentially sensitive ad preferences and quantify the portion of EU FB accounts associated with them, we seek to collect a dataset of ad preferences linked to actual EU FB accounts. If we detect ad preferences that represent potentially sensitive personal data, this dataset would provide evidence that the preferences are assigned to real FB accounts. Based on this goal, our dataset is created from the ad preferences collected from real users who have installed the FDVT. We note that the number of ad preferences retrieved from the FDVT represents just a subset of the overall set of preferences, but we can guarantee that they have been assigned to real accounts. Our dataset includes the ad preferences from 4577 users who installed the FDVT between October 2016 and October 2017, from which 3166 users come from some EU country. These 4577 FDVT users have been assigned 5.5M ad preferences in total of which 126192 are unique.

Our dataset includes the following information for each ad preference:

-*ID of the ad preference*: This is the key we use to identify an ad preference independently of the language used by a FB user. For instance, the ad preference {Milk, Leche, Lait} that refers to the same thing in English, Spanish and French, is assigned a single FB ID. Therefore, we can uniquely identify each ad preference across all EU countries and languages.

-*Name of the ad preference*: This is the primary descriptor of the ad preference. FB returns a unified version of the name for each ad preference ID, usually in English. Hence, we have the English name of the ad preferences irrespective of the original language at collection. We note that in some cases translating the ad preference name does not make sense (e.g., the case of persons' names: celebrities, politicians, etc.).

-*Disambiguation Category*: For some ad preferences Facebook adds this in a separate field or in parenthesis to clarify the meaning of a particular ad preference (e.g., Violet (color); Violet: Clothing (Brand)) We have identified more than 700 different disambiguation category topics (e.g., Political Ideology, Disease, Book, Website, Sport Team, etc.). Among the 126K ad preferences analyzed, 87% include this field.

-*Topic Category*: In many cases, some of the 14 first level interests introduced in Section 2.1 are assigned to contextualize ad preferences. For instance, Manchester United F.C. is linked to Sports and Outdoors.

-*Audience Size*: This value reports the number of Facebook users that have been assigned the ad preference worldwide.

-*Reason why the ad preference is added to the user*: The reason why the ad preference has been assigned to the user according to FB. There are six possible reasons introduced in Subsection 2.2.

Figure 2 shows the CDF of the number of ad preferences per user. Each FDVT user is assigned a median of 474 preferences. Moreover, Figure 3 shows the CDF of the portion of FDVT users (x-axis) that were assigned a given ad preference (y-axis). We observe a very skewed distribution that indicates that most ad preferences are actually assigned to a small fraction of users. For instance, each ad preference is assigned to a median of only 3 (0.06%) FDVT users. However, it is important to note that many ad preferences still reach a reasonable portion of users. Our dataset includes 1000 ad preferences that reach at least 11% of FDVT users.

5 Methodology

We seek to quantify the number of EU FB users that have been assigned potentially sensitive ad preferences. To this end, we use the 126K unique ad preferences assigned to FDVT users and follow a two-step process. In the first step, we combine Natural Language Processing (NLP) techniques with manual classification to obtain a list of likely sensitive ad preferences from the 126K considered. In the second step, we leverage the FB Ads Manager API to quantify how many FB users in each EU country have been assigned at least one of the ad preferences labeled as potentially sensitive.

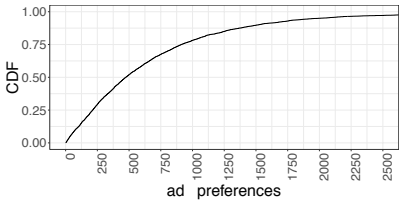

Figure 2: CDF of the number of ad preferences per FDVT user.

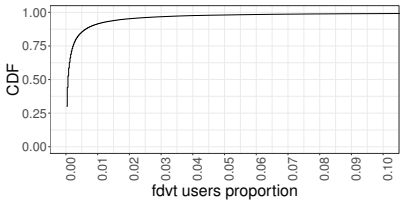

Figure 3: CDF of the portion of FDVT users (x-axis) per ad preference (y-axis).

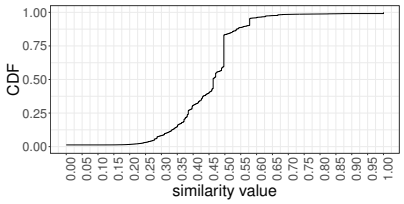

Figure 4: CDF of the semantic similarity score assigned to the 126K ad preferences from the FDVT dataset.

5.1 Identification of potentially sensitive ad preferences

We rely on a group of researchers with some knowledge in the area of privacy to manually identify potentially sensitive ad preferences within our pool of 126K ad preferences retrieved from FDVT users. However, manually classifying 126K ad preferences would be unfeasible.[7] To make this manual classification task scalable, we leverage NLP techniques to pre-filter the list of ad preferences more likely to be sensitive. This pre-filtering phase will deliver a subset of likely sensitive ad preferences that can be manually classified in a reasonable amount of time.

5.1.1 Pre-filtering

Sensitive categories: To identify likely sensitive ad preferences in an automated manner, we select five of the relevant categories listed as *Sensitive Personal Data* by the GDPR: (i) data revealing racial or ethnic origin, (ii) data revealing political opinions, (iii) data revealing religious or philosophical beliefs, (iv) data concerning health, and (v) data concerning sex life and sexual orientation. We selected these categories because a preliminary manual inspection indicated that there are ad preferences in our dataset that can likely reveal information related to them. For instance, the ad preferences *"Socialism"*,*"Islam"*,*"Reproductive Health"*,*"Homosexuality"* or *"Black Feminism"* may suggest *political opinion, religious belief, health issue, sexual orientation* or *ethnic or racial origin* of the users that have been assigned them, respectively. Note that all these examples of ad preferences have been extracted from our dataset; thus they have been assigned to actual FB users.

Our automated process will classify an ad preference as *likely sensitive* if we can semantically map that ad preference name into one of the five sensitive categories analyzed in this paper. To this end, we have defined a dictionary including both keywords and short sentences representative of each of the five considered sensitive categories. We used two data sources to create the dictionary: First, a list of controversial issues available in Wikipedia.[8] In particular, we selected the following categories from this list: politics and economics, religion, and sexuality. Second, we obtained a list of words with a very similar semantic meaning to the five sensitive personal data categories. To this end, we used the Datamuse API,[9] a word-finding query engine that allows developers to find words that match a set of constraints. Among other features, Datamuse allows *"finding words with a similar meaning to X"* using a simple query.

The final dictionary includes 264 keywords.[10] We leverage the keywords in this dictionary to find ad preferences that present high semantic similarity to at least one of these keywords. In these cases, we tag them as likely sensitive ad preferences. It is worth noting that this approach makes our methodology flexible, since the dictionary can be extended to include new keywords for the considered categories or other categories, which may uncover additional potentially sensitive ad preferences.

We next describe the semantic similarity computation in detail.

Semantic similarity computation: The semantic similarity computation process takes two inputs: the 126K ad preferences from our FDVT dataset and the 264 keyword dictionary associated with the considered sensitive categories. We compute the semantic similarity of each ad preference with all of the 264 keywords from the dictionary. For each ad preference, we record the highest similarity value out of the 264 comparison operations. As result of this process, each one of the 126K ad preferences is assigned a similarity score, which indicates its likelihood to be a sensitive ad preference.

To implement the semantic similarity comparison task, we leverage the Spacy package for python[11] (see

[7] If we consider 10s as the average time required to classify an ad preference as sensitive vs. non-sensitive, this task would require 44 full eight-hour days.

[8] https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues

[9] https://www.datamuse.com/api/

[10] https://fdvt.org/usenix2018/keywords.html

[11] https://spacy.io

details about Spacy in Appendix D). We chose Spacy because it has been previously used in the literature for text processing purposes offering good performance [15][22]. Moreover, Spacy offers good scalability. It computes the 33314688 (126192 x 264) semantic similarity computations in 7 min using a server with twelve 2.6GHZ cores and 96GB of RAM. To conduct our analysis we leverage the *similarity* feature of Spacy. This feature allows comparing words, text spans or documents, and computes the semantic similarity among them. The output is a semantic similarity value ranging between 0 and 1. The closer to 1 the higher the semantic similarity is.

This process revealed very low similarity values for some cases in which the analyzed ad preference closely matched the definition of some of the sensitive personal data categories. Some of these cases are: physical persons such as politicians (which may reveal the political opinion of the user); political parties with names that do not include any standard political term; health diseases or places of religious cults that may have names with low semantic similarity with health and religious related keywords in our dictionary, respectively. Three examples illustrating the referred cases are: <name: "Angela Merkel", disambiguation: Politician>; <name: "I Love Italy", disambiguation: Political Party>; <name: "Kegel" exercise, disambiguation: Medical procedure>. In most of these cases the disambiguation category is more useful than the ad preference name when performing the semantic similarity analysis. For instance, in the case of politicians' names, political parties and health diseases the disambiguation category field includes the term *"politician"*, *"Political Party"* and *"disease"*, respectively. This field is also very useful for determining the definition of ad preference names that have multiple meanings.

Overall, we found that for classifying ad preferences, the disambiguation category, when it is available, is a better proxy than the ad preference name. Therefore, if the ad preference under analysis has a disambiguation category field, we used the disambiguation category string instead of the ad preference name to obtain the semantic similarity score of the ad preference.

Selection of likely sensitive ad preferences: The semantic similarity computation process assigns a similarity score to each one of the 126K ad preferences in our dataset. This similarity score represents the anticipated likelihood for an ad preference to be sensitive.

In this step of the process, we have to select a relatively high similarity score threshold that allows us to create a subset of likely sensitive ad preferences that can be manually labeled with reasonable manual effort.

Figure 4 shows the CDF for the semantic similarity score of the 126K ad preferences. The curve is flat near 0 and 1, with a steep rise between similarity values 0.25 and 0.6. This steep rise implies that setting our threshold to values below 0.6 would result in a rapid growth of the number of ad preferences to be manually tagged. Therefore, we set the semantic similarity threshold to 0.6 because it corresponds to a relatively high similarity score. The resulting automatically filtered subset includes 4452 ad preferences (3.5% of the 126K), which is a reasonable number to be manually tagged.

Note that the CDF has two jumps at similarity scores equal to 0.5 and 0.58. The first one is linked to the disambiguation category *"Local Business"* while the second one refers to the disambiguation category *"Public Figure"*. Overall, we do not expect to find a significant number of potentially sensitive ad preferences within these disambiguation categories. Hence, this observation reinforces our semantic similarity threshold selection of 0.6.

5.1.2 Manual classification of potentially sensitive ad preferences

We recruited twelve panelists. All of them are researchers (faculty and Ph.D. students) with some knowledge in the area of privacy. Each panelist manually classified a random sample (between 1000 and 4452 elements) from the 4452 ad preferences included in the automatically filtered subset described above. We asked them to classify each ad preference into one of the five considered sensitive categories (Politics, Health, Ethnicity, Religion, Sexuality), in the category "Other" (if it does not correspond to any of the sensitive categories), or in the category "Not known" (if the panelist does not know the meaning of the ad preference). To carry out the manual labeling, the researchers were given all the contextual information Facebook offers per ad preference: name, disambiguation category (if available) and topic (if available).[12]

Each ad preference was manually classified by five panelists. We use majority voting [20] to classify each ad preference either as sensitive or non-sensitive. That is, we label an ad preference as sensitive if at least three voters (i.e., the majority) classify it in one of the five sensitive categories and as non-sensitive otherwise.

Table 1 shows the number of ad preferences that received 0, 1, 2, 3, 4 and 5 votes classifying them into a

[12]The provided instructions to panelists were: *"Assign only one category per ad preference. If you think that more than one category applies to an ad preference use only the one you think is most relevant. If none of the categories match the ad preference, classify it as 'Other'. In case you do not know the meaning of an ad preference please read the disambiguation category and topic that may help you. If after reading them you still are unable to classify the ad preference, use 'Not known' to classify it."*

votes	0	1	2	3	4	5
#preferences	1054	767	539	422	449	1221

Table 1: Number of ad preferences that received 0, 1, 2, 3, 4 or 5 votes classifying them into one sensitive data categories.

sensitive category. 2092 out of the 4452 ad preferences are labeled as sensitive, i.e., have been classified into a sensitive category by at least 3 voters. This represents 1.66% of the 126K ad preferences from our dataset.

An ad preference classified as sensitive may have been assigned to different sensitive categories (e.g., politics and religion) by different voters. We have evaluated the voters' agreement across the sensitive categories assigned to ad preferences labeled as sensitive using the Fleiss' Kappa test [10][11]. The Fleiss' Kappa coefficient obtained is 0.94. This indicates an almost perfect agreement among the panelists' votes that link an ad preference to a sensitive category [16]. Hence, we conclude that (almost) every ad preference classified as sensitive corresponds to a unique sensitive category among the 5 considered.

The 2092 ad preferences manually labeled as sensitive are distributed as follows across the five sensitive categories: 58.3% are related to politics, 20.8% to religion, 18.2% to health, 1.5% to sexuality, 1.1% to ethnicity and just 0.2% present discrepancy among votes. The complete list of the ad preferences classified as sensitive can be accessed via the FDVT site.[13] We refer to this subset of 2092 ad preferences as the *suspected sensitive subset*.

5.2 Retrieving the number of FB users assigned potentially sensitive ad preferences from the FB Ads Manager

We leverage the FB Ads Manager API to retrieve the number of FB users in each EU country that have been assigned each of the 2092 potentially sensitive ad preferences from the suspected sensitive subset. We collected this information in January 2018. Following that, we sorted these ad preferences from the most to the least popular in each country. This allows us to compute the number of FB users assigned at least one of the Top N potentially sensitive ad preferences (with N ranging between 1 and 2092). To obtain this information we use the OR operation available in the FB Ads Manager API to create audiences. This feature allows us to retrieve how many users in a given country are interested in *ad preference 1* OR *ad preference 2* OR *ad preference 3*... OR *ad preference N*. An example of this for N = 3 could be *"how many people in France are interested in Communism OR Islam OR Veganism"*.

Although the number of users is a relevant metric, it does not offer a fair comparative result to assess the importance of the problem across countries because we can find EU countries with tens of millions of users (e.g., France, Germany, Italy, etc) and some others with less than a million (e.g., Malta, Luxembourg, etc). Hence, we use the portion of users in each country that have been assigned potentially sensitive ad preferences as the metric to analyze the results. Beyond FB users we are also interested in quantifying the portion of citizens assigned sensitive ad preferences in each EU country. We have defined two metrics used in the rest of the paper:

-***FFB(C,N)***: This is the percentage of FB users in country C that have been assigned at least one of the top N potentially sensitive ad preferences from the suspected sensitive subset. We note C may also refer to all 28 EU countries together when we want to analyze the results for the whole EU. It is computed as the ratio between the number of FB users that have been assigned at least one of the top N potentially sensitive ad preferences and the total number of FB users in country C, which can be retrieved from the FB Ads Manager.

-***FC(C,N)***: This is the percentage of citizens in country C (or all EU countries together) that have been assigned at least one of the top N potentially sensitive ad preferences. It is computed as the ratio between the number of citizens that have been assigned at least one of the top N potentially sensitive ad preferences and the total population of country C. We use World Bank data to obtain EU countries' populations.[14]

The criterion to select the top N ad preferences out of the 2092 potentially sensitive ad preferences identified is popularity. This means that we select the N ad preferences assigned to the most users according to the FB Ads Manager API. Note that FFB(C,N) and FC(C,N) will likely report a lower bound concerning the total percentage of FB users and citizens in country C tagged with potentially sensitive ad preferences for two reasons. First, these metrics can use at most N = 2092 potentially sensitive ad preferences, which (assuming that our voters are accurate) is very likely a subset of all sensitive ad preferences available on FB. Second, the FB Ads Manager API only allows creating audiences with at most N = 1000 interests (i.e., ad preferences). Beyond N = 1000 interests the API provides a fixed number of FB users independently of the defined audience. This fixed number is 2.1B which to the best of our knowledge refers the total number of FB users included in the Ads Manager. Therefore, in practice, the maximum value of N we can use in FFB and FC is 1000.

[13] `https://fdvt.org/usenix2018/panelists.html`

[14] `https://data.worldbank.org`

reason of assignment	all ad preferences	potentially sensitive ones
due to a like	71.64%	81.36%
due to an ad click	21.51%	15.85%
FB suggests it could be relevant	4.83%	2.45%
due to an app installation	1.78%	0.04%
due to comments or reaction buttons	0.18%	0.26%
added by user	0.04%	0.03%
unclear or not gathered by FDVT	0.01%	0.01%

Table 2: Frequency of the six reasons why ad preferences are assigned to FDVT EU users according to Facebook explanations.

6 Quantifying the exposure of EU users to potentially sensitive ad preferences

In this section, we first analyze the exposure of the FDVT users to the 2092 potentially sensitive ad preferences included in the suspected sensitive subset. Afterwards, we use the FFB and FC metrics to analyze the exposure of EU FB users and citizens to those ad preferences. Finally, we perform a demographic analysis to understand whether users from certain gender or age groups are more exposed to sensitive ad preferences.

6.1 FDVT users

4121 (90%) FDVT users are tagged with at least one sensitive ad preference. Overall, the 2092 unique sensitive ad preferences have been assigned more than 146K times to the FDVT users. If we focus only on EU users, which are the focus of this paper, 2848 (90%) have been tagged with potentially sensitive ad preferences. Overall, they have been assigned more than 100K sensitive interests (1528 unique). The median (avg) number of potentially sensitive ad preferences assigned to FDVT users is 10 (16). The 25th and 75th percentiles are 5 and 21, respectively.

Our FDVT dataset includes the reason why, according to FB, each ad preference has been assigned to a user. Table 2 shows the frequency of each reason for both all ad preferences and only the potentially sensitive ones. The results indicate that most of the sensitive ad preferences are derived from *users likes* (81%) or *clicks on ads* (16%). There are very few cases (0.03%) in which users proactively include potentially sensitive ad preferences in their list of ad preferences using the configuration setting offered by FB. As a reminder, according to the EU GDPR, FB should obtain explicit permission to process and exploit sensitive personal data. Users likes and clicks on ads do not seem to meet this requirement.

6.2 EU FB users and citizens

Figure 5 shows the FFB (C,N) for values of N ranging between 1 and 1000. The figure reports the max, min and avg values across the 28 EU countries.[15] We observe that even if we consider a low number of sensitive ad preferences, the fraction of affected users is very significant. For instance, on average 60% of FB users from EU countries are tagged with some of the top 10 (i.e., most popular) potentially sensitive ad preferences.

Moreover, we observe that FFB is stable for values of N ranging between 500 and 1000. We note that we have obtained the same stable result for each individual EU country. This indicates that any user tagged with potentially sensitive ad preferences outside the top 500[16] has likely been already tagged with at least one potentially sensitive ad preference within the top 500. We conjecture that this asymptotic behavior may indicate that the lower bound represented by FFB(C, N=500) is close to the actual fraction of FB users tagged with sensitive ad preferences.

Table 3 shows FFB(C,N=500) and FC(C,N=500) for every EU country. The last row in the table shows average results for the 28 EU countries together (EU28).

We observe that 73% of EU FB users, which corresponds to 40% of EU citizens, are tagged with some of the top 500 potentially sensitive ad preferences in our dataset. If we focus on individual countries, FC(C,N=500) reveals that in 7 of them more than half of their citizens are tagged with at least one of the top 500 potentially sensitive ad preferences: Malta (66.37%), Cyprus (64.95%), Sweden (54.53%), Denmark (54.09%), Ireland (52.38%), Portugal (51.33%) and Great Britain (50.28%). In contrast, the 5 countries least impacted are: Germany (30.24%), Poland (31.62%), Latvia (33.67%), Slovakia (35%) and Czech Republic (35.98%). Moreover, FFB(C,N=500) ranges between 65% for France and 81% for Portugal. This means that approximately 2/3 or more of FB users in any EU country are tagged with some of the top 500 potentially sensitive ad preferences.

These results suggest that a very significant part of the EU population can be targeted by advertising campaigns based on potentially sensitive personal data.

6.3 Expert-verified sensitive ad preferences

To confirm that our set of potentially sensitive ad preferences contains ones likely relevant under GDPR, we examined a subset of 20 ad preferences that all panelists classified as sensitive. An expert from the Spanish DPA reviewed and confirmed the sensitivity of each of the 20

[15] The average across EU countries has been computed by summing the average of each EU country and dividing it by 28 since the Top N preference for each country changes from country to country.

[16] The top 500 list by country can be accessed at `https://fdvt.org/usenix2018/top500.html`

country	C	FFB(C,500)	FC (C,500)	country	C	FFB(C,500)	FC (C,500)
Austria	AT	75.00	37.73	Ireland	IE	80.65	52.38
Belgium	BE	70.27	45.82	Italy	IT	79.41	44.55
Bulgaria	BG	72.97	37.88	Latvia	LV	72.53	33.67
Croatia	HR	80.00	38.36	Lithuania	LT	75.00	41.78
Cyprus	CY	79.17	64.95	Luxembourg	LU	72.22	44.60
Czech Republic	CZ	71.70	35.98	Malta	MT	80.56	66.37
Denmark	DK	77.50	54.09	Netherlands	NL	74.55	48.18
Estonia	EE	66.67	36.46	Poland	PL	75.00	31.62
Finland	FI	70.97	40.04	Portugal	PT	81.54	51.33
France	FR	65.79	37.37	Romania	RO	75.76	38.06
Germany	DE	67.57	30.24	Spain	ES	74.07	43.06
Great Britain	GB	75.00	50.28	Slovakia	SK	70.37	35.00
Greece	GR	77.19	40.94	Slovenia	SI	78.00	37.78
Hungary	HU	75.44	43.80	Sweden	SE	73.97	54.53
				European Union	EU	73.25	40.63

Table 3: Percentage of EU FB users (FFB) and citizens (FC) per EU Country that have been assigned some of the Top 500 potentially sensitive ad preferences within their country. The last row reports the aggregated number of all 28 EU countries together.

ad preferences in that subset according to the GDPR. We note this subset is not necessarily representative of all potentially sensitive ad preferences (or preferences that EU citizens may find objectionable), but it represents an expert-validated subset we use for further analysis.

Tables 4 and 5 show the percentage of FB users (FFB) and citizens (FC) tagged with each of the 20 expert-verified sensitive ad preferences per EU country. Note that the last row presents the aggregate results for the 20 in each country, and the last column presents the aggregate results for the 28 EU countries together.

We observe that 42.9% of EU FB users, which corresponds to 23.5% of EU citizens, are tagged with at least one of the expert-verified sensitive ad preferences. Hence, around one-quarter of the EU population has been tagged in FB with at least one of the expert-verified sensitive ad preferences. If we analyze the results per country, we observe that the fraction of the population affected ranges between 15% in Estonia (EE), Latvia (LV) and Poland (PL) and 38% in Malta (MT). These findings suggest that FB may have used GDPR-relevant data for a large percentage of EU citizens in the period prior to when the GDPR became enforceable.

6.4 Age and gender analysis

We analyze the association of different demographic groups (based on gender and age) with potentially sensitive ad preferences. The gender analysis considers two groups, men vs. women, while the age analysis considers four age groups following the division proposed by Erikson et al. [7]: 13-19 (Adolescence), 20-39 (Early Adulthood), 40-64 (Adulthood) and 65+ (Maturity). For each group, we compute FFB(C = EU28, N = 500) from the 2092 suspected sensitive ad preferences subset and FFB(C = EU28, N = 20) using exclusively expert-verified sensitive ad preferences. Figures 6 and 7 report the results for age and gender groups, respectively.

The Early Adulthood group is clearly the most exposed age group to suspected (20-expert-verified) sensitive ad preferences. 61% (45%) of users in this group have been tagged with some of the Top 500-suspected (20-expert-verified) sensitive ad preferences. Following the Early Adulthood group we find the Adolescence, Adulthood and Maturity groups with 55% (42%), 40% (32%) and 39% (28%) of its users tagged with some of the Top 500-potentially (20-expert-verified) sensitive ad preferences, respectively. Although the difference in the exposure to sensitive ad preferences is substantial across groups, all of them present a considerably high exposure. In particular, more than one-quarter of the users within every group is exposed to expert-verified sensitive ad preferences.

The gender-based analysis shows that 78% (49%) of women are exposed to the Top 500-suspected (20-expert-verified) ad preferences. The exposure is notably smaller for men, where the fraction of tagged users with some of the Top 500-suspected (20-expert-verified) sensitive ad preferences shrinks by 10 (18) percentage points to 68% (31%). This result suggests the existence of a gender bias, which despite its obvious interest is out of the scope of this paper.

7 Commercial exploitation of sensitive ad preferences with real FB ad campaigns

Our analysis shows that Facebook labeled a significant portion of EU citizens using potentially sensitive personal data. In this section, we demonstrate that FB allowed ads to be targeted to users assigned to expert-verified sensitive ad preferences. Between October 6 and October 15, 2017 we ran three FB ad campaigns using expert-verified sensitive ad preferences such as: *"religious beliefs"* (targeting users interested in Islam OR Judaism OR Christianity OR Buddhism), *"political opinions"* (targeting users interested in Communism OR Anarchism OR Radical feminism OR Socialism) and *"sexual orientation"* (targeting users interested in Transsexualism OR Homosexuality).[17] The 3 campaigns focused on four EU countries: Germany, Spain, France and Italy.

Overall, with a budget of €35 we were able to reach 26458 users tagged with some of the previous sensitive ad preferences. Our credit card was charged and we received the bills and summary reports associated with our campaigns (see Figure 8). This experiment provides substantial evidence that FB generated (before May 25) revenue from the commercial exploitation of expert-verified sensitive personal data according to the GDPR definition of *sensitive data*.

[17] *"Anarchism"* and *"Transsexualism"* were not explicitly verified by the expert but closely mirror verified ad preferences.

name	AT	BE	BG	HR	CY	CZ	DK	EE	FI	FR	DE	GR	HU	IE	IT	LV	LT	LU	MT	NL	PL	PT	RO	SK	SI	ES	SE	GB	EU28
COMMUNISM	0.48	0.61	1.35	1.30	1.67	3.21	0.38	0.61	0.52	2.29	0.43	0.81	0.74	0.52	1.15	0.56	0.94	0.64	0.39	0.24	2.19	0.94	1.90	1.74	1.70	0.56	0.30	0.41	0.93
ISLAM	8.18	7.16	4.59	5.50	13.54	4.91	6.75	2.22	4.19	7.89	7.57	4.21	2.28	4.19	4.12	2.75	2.38	5.00	6.67	5.36	2.44	3.69	3.50	3.11	6.50	4.07	6.58	6.82	5.71
QURAN	3.41	3.38	1.08	1.00	4.48	0.45	1.90	0.65	1.16	3.95	3.24	1.18	0.74	1.35	1.71	1.01	0.51	1.83	1.86	2.45	0.45	0.62	0.77	0.56	2.00	0.96	2.74	3.64	2.46
SUICIDE PREVENTION	0.14	0.15	0.20	0.32	0.21	0.12	0.12	0.10	0.09	0.16	0.14	0.23	0.12	1.10	0.28	0.13	0.15	0.28	0.27	0.15	0.14	0.22	0.13	0.44	0.26	0.44	0.15	0.27	0.28
SOCIALISM	1.00	0.78	0.57	0.48	1.15	2.45	3.00	0.76	0.48	0.47	0.43	0.91	1.93	1.10	3.53	0.34	0.94	2.78	1.08	0.28	0.50	2.15	0.35	2.33	0.82	1.48	1.37	0.93	1.21
JUDAISM	2.50	1.16	0.86	0.70	2.29	0.72	2.17	1.01	0.61	1.26	1.38	1.30	1.16	1.26	2.29	1.76	1.81	1.19	3.06	1.00	1.19	1.69	1.40	0.93	0.74	1.15	0.64	0.95	1.32
HOMOSEXUALITY	6.14	5.54	2.97	6.50	4.38	5.47	5.00	3.89	5.16	7.37	5.68	5.09	4.21	9.03	7.65	4.62	3.19	5.00	7.50	6.18	3.56	4.46	3.80	4.44	7.60	8.15	4.93	8.64	6.79
ALTERNATIVE MEDICINE	5.00	2.97	8.38	6.00	5.62	4.15	4.00	4.17	4.19	2.89	3.24	7.19	4.21	9.68	6.18	3.96	2.56	5.56	7.50	3.64	2.25	8.00	3.90	2.93	5.00	5.56	3.84	6.14	4.29
CHRISTIANITY	10.68	7.43	6.22	7.50	9.69	3.77	15.00	2.22	4.19	5.53	6.49	6.67	9.30	10.97	12.65	3.19	3.81	7.22	18.89	5.18	6.25	12.46	10.00	4.81	4.60	10.00	4.66	7.50	8.21
ILLEGAL IMMIGRATION	0.17	0.07	0.10	0.02	0.07	0.68	0.05	0.01	0.07	0.05	0.06	0.26	0.26	0.06	0.08	0.02	0.06	0.01	0.08	0.02	0.02	0.02	0.02	0.11	0.36	0.14	0.33	0.05	0.09
ONCOLOGY	0.23	0.27	0.62	0.44	3.96	0.57	0.15	0.10	0.08	0.17	0.16	0.49	0.30	1.29	0.94	0.70	1.62	0.19	0.78	0.45	1.25	1.09	0.73	0.59	0.21	0.70	0.08	0.66	0.61
LGBT COMMUNITY	6.36	6.62	5.14	6.50	6.56	6.04	6.50	5.14	6.45	7.11	5.95	5.79	4.39	11.94	8.53	5.27	5.88	6.67	9.44	6.36	5.88	7.85	6.30	4.81	6.00	7.04	6.44	11.14	8.21
GENDER IDENTITY	0.03	0.08	0.01	0.08	0.88	0.02	0.03	0.02	0.02	0.07	0.03	0.56	0.07	0.23	0.07	0.20	0.10	0.10	0.14	0.03	0.05	0.05	0.04	0.01	0.08	0.07	0.09	0.55	0.10
REPRODUCTIVE HEALTH	0.01	0.07	0.20	0.40	0.02	0.14	0.05	0.02	0.06	0.01	0.01	0.04	0.10	0.71	0.04	0.07	0.05	0.01	0.24	0.03	0.01	0.04	0.01	0.03	0.00	0.03	0.05	0.13	0.07
BIBLE	17.95	10.81	8.65	10.50	11.46	7.17	12.75	4.31	4.84	7.63	15.41	8.25	10.00	19.03	17.65	5.71	6.25	14.44	20.28	10.91	14.38	12.31	8.70	6.67	7.40	7.04	5.48	15.68	12.14
PREGNANCY	15.68	12.97	9.19	17.00	13.54	16.23	14.50	10.00	11.29	10.79	11.89	13.51	11.23	20.97	12.35	13.19	18.75	12.78	9.72	14.55	15.00	18.46	9.70	18.89	13.00	14.07	13.42	18.41	14.29
NATIONALISM	0.86	0.78	1.65	1.85	2.19	2.45	1.00	0.58	0.45	1.08	1.00	1.74	2.11	2.00	1.32	2.42	0.94	2.19	2.78	0.70	3.00	1.69	2.50	1.37	0.61	1.11	0.99	0.91	1.39
VEGANISM	14.55	10.27	7.30	10.50	10.21	9.25	12.75	9.86	15.16	8.68	11.35	9.82	9.82	14.84	13.53	9.23	8.12	13.06	13.33	10.91	8.12	11.23	6.70	8.52	14.00	10.37	16.44	13.64	11.43
BUDDHISM	3.18	3.38	1.62	3.55	3.33	2.26	2.08	1.53	1.13	2.61	1.43	2.63	3.33	3.87	2.94	1.98	1.88	3.33	4.17	2.45	1.31	6.92	1.90	1.67	3.00	2.19	1.51	2.50	2.39
FEMINISM	4.55	3.78	3.51	3.80	5.52	2.08	5.50	2.78	6.77	5.00	3.78	3.68	2.46	9.35	5.88	3.19	3.56	5.83	8.61	3.64	3.44	8.15	2.40	4.07	3.90	8.89	13.70	7.27	7.50
UNION	45.45	39.19	32.43	41.50	45.83	37.74	45.00	27.78	35.48	34.21	40.54	36.84	36.84	51.61	44.12	32.97	36.25	41.67	47.22	40.00	36.88	44.62	34.34	35.56	39.00	40.74	41.10	47.73	42.86

Table 4: Percentage of FB users (FFB) per EU country that have been assigned each of the 20 expert-verified sensitive ad preferences listed in the table. The last row reports the aggregated FFB value for all 20 ad preferences per EU country. The last column reports the aggregated FFB value across all 28 EU countries.

name	AT	BE	BG	HR	CY	CZ	DK	EE	FI	FR	DE	GR	HU	IE	IT	LV	LT	LU	MT	NL	PL	PT	RO	SK	SI	ES	SE	GB	EU28
COMMUNISM	0.24	0.40	0.70	0.62	1.37	1.61	0.26	0.33	0.29	1.30	0.19	0.43	0.43	0.34	0.64	0.26	0.52	0.39	0.32	0.15	0.92	0.59	0.96	0.87	0.82	0.32	0.22	0.27	0.51
ISLAM	4.12	4.67	2.39	2.64	11.11	2.46	4.71	1.22	2.37	4.48	3.39	2.23	1.32	2.72	2.31	1.28	1.32	3.09	5.49	3.47	1.03	2.32	1.78	1.55	3.15	2.37	4.85	4.57	3.13
QURAN	1.71	2.20	0.56	0.48	3.67	0.23	1.33	0.36	0.66	2.24	1.45	0.62	0.43	0.88	0.96	0.47	0.28	1.13	1.53	1.59	0.19	0.39	0.39	0.28	0.97	0.56	2.02	2.44	1.35
SUICIDE PREVENTION	0.07	0.10	0.10	0.15	0.17	0.06	0.08	0.05	0.05	0.09	0.06	0.12	0.07	0.71	0.16	0.06	0.08	0.17	0.22	0.09	0.06	0.14	0.07	0.22	0.13	0.26	0.11	0.18	0.15
SOCIALISM	0.50	0.51	0.29	0.23	0.94	1.23	2.09	0.42	0.27	0.27	0.19	0.48	1.12	0.71	1.98	0.16	0.52	1.72	0.89	0.18	0.21	1.36	0.18	1.16	0.40	0.86	1.01	0.62	0.66
JUDAISM	1.26	0.76	0.45	0.34	1.88	0.36	1.52	0.55	0.35	0.72	0.62	0.69	0.67	0.82	1.29	0.82	1.01	0.74	2.52	0.65	0.50	1.07	0.71	0.46	0.36	0.67	0.47	0.64	0.72
HOMOSEXUALITY	3.09	3.61	1.54	3.12	3.59	2.75	3.49	2.13	2.91	4.19	2.54	2.70	2.44	5.87	4.29	2.14	1.78	3.09	6.18	4.00	1.50	2.81	1.93	2.21	3.68	4.74	3.64	5.79	3.71
ALTERNATIVE MEDICINE	2.52	1.94	4.35	2.88	4.61	2.08	2.79	2.28	2.37	1.64	1.45	3.82	2.44	6.29	3.47	1.84	1.43	3.43	6.18	2.35	0.95	5.04	1.98	1.46	2.42	3.23	2.83	4.11	2.34
CHRISTIANITY	5.37	4.85	3.23	3.60	7.95	1.89	10.47	1.22	2.37	3.14	2.90	3.54	5.40	7.12	7.10	1.48	2.12	4.46	15.56	3.35	2.64	7.85	5.07	2.39	2.23	5.81	3.43	5.03	4.49
ILLEGAL IMMIGRATION	0.09	0.04	0.05	0.01	0.06	0.34	0.03	0.00	0.04	0.03	0.03	0.14	0.15	0.04	0.04	0.01	0.03	0.01	0.07	0.01	0.01	0.01	0.01	0.05	0.17	0.08	0.24	0.04	0.05
ONCOLOGY	0.11	0.18	0.32	0.21	3.25	0.28	0.10	0.06	0.05	0.10	0.07	0.26	0.17	0.84	0.53	0.33	0.91	0.12	0.64	0.29	0.53	0.69	0.37	0.29	0.10	0.41	0.06	0.44	0.33
LGBT COMMUNITY	3.20	4.32	2.67	3.12	5.38	3.03	4.54	2.81	3.64	4.04	2.66	3.07	2.55	7.75	4.79	2.45	3.27	4.12	7.78	4.11	2.48	4.94	3.20	2.39	2.91	4.09	4.75	7.47	4.49
GENDER IDENTITY	0.01	0.05	0.01	0.04	0.72	0.01	0.02	0.01	0.01	0.04	0.01	0.30	0.04	0.15	0.04	0.09	0.06	0.06	0.12	0.02	0.02	0.03	0.02	0.00	0.04	0.04	0.06	0.37	0.05
REPRODUCTIVE HEALTH	0.00	0.05	0.11	0.19	0.02	0.07	0.04	0.01	0.04	0.01	0.00	0.02	0.06	0.46	0.02	0.03	0.03	0.01	0.19	0.02	0.01	0.02	0.01	0.01	0.00	0.02	0.03	0.09	0.04
BIBLE	9.03	7.05	4.49	5.04	9.40	3.60	8.90	2.35	2.73	4.34	6.90	4.37	5.81	12.36	9.90	2.65	3.48	8.92	16.71	7.05	6.06	7.75	4.42	3.32	3.58	4.09	4.04	10.51	6.64
PREGNANCY	7.89	8.46	4.77	8.15	11.11	8.14	10.12	5.47	6.37	6.13	5.32	7.16	6.52	13.62	6.93	6.12	10.44	7.89	8.01	9.40	6.32	11.62	4.92	9.39	6.30	8.18	9.90	12.34	7.82
NATIONALISM	0.43	0.51	0.86	0.89	1.79	1.23	0.70	0.32	0.25	0.61	0.45	0.92	1.22	1.30	0.74	1.12	0.52	1.36	2.29	0.45	1.26	1.07	1.27	0.68	0.30	0.65	0.73	0.61	0.76
VEGANISM	7.32	6.70	3.79	5.04	8.38	4.64	8.90	5.39	8.55	4.93	5.08	5.21	5.70	9.64	7.59	4.28	4.53	8.06	10.99	7.05	3.43	7.07	3.40	4.24	6.78	6.03	12.12	9.14	6.25
BUDDHISM	1.60	2.20	0.84	1.70	2.73	1.14	1.45	0.84	0.64	1.48	0.64	1.40	1.94	2.51	1.65	0.92	1.04	2.06	3.43	1.59	0.55	4.36	0.96	0.83	1.45	1.27	1.11	1.68	1.31
FEMINISM	2.29	2.47	1.82	1.82	4.53	1.04	3.84	1.52	3.82	2.84	1.69	1.95	1.43	6.08	3.30	1.48	1.98	3.60	7.09	2.35	1.45	5.13	1.22	2.03	1.89	5.17	10.10	4.88	4.10
UNION	22.86	25.55	16.84	19.90	37.60	18.94	31.41	15.19	20.02	19.43	18.14	19.54	21.39	33.52	24.75	15.30	20.19	25.73	38.91	25.85	15.55	28.09	17.25	17.68	18.89	23.68	30.29	31.99	23.45

Table 5: Percentage of citizens (FC) per EU country that have been assigned each of the 20 expert-verified sensitive ad preferences listed in the table. The last row reports the aggregated FC value for all 20 ad preferences per EU country. The last column reports the aggregated FC value across all 28 EU countries.

To the best of our knowledge the conducted ad campaigns were compliant with the Terms of service of Facebook introduced in Section 3.3 and Appendix E.

8 Ethics and privacy risks associated with sensitive personal data exploitation

The possibility of reaching users labeled with potentially sensitive personal data enables the use of FB ads campaigns to attack specific groups of people based on sensitive personal data (race, sexual orientation, religious beliefs, etc.). Below, we illustrate two specific examples of potential attacks.

Hate campaigns: An attacker could create hate speech campaigns using sensitive ad preferences representative of a specific sensitive social group within its target audience. For instance, a neo-Nazi organization could create ads campaigns with offensive messages targeting people interested in *Judaism* or *Homosexuality*. Hate speech campaigns can reach thousands of users at a very low cost (e.g., we reached more than 26K FB users spending only €35 in FB ads campaigns).

Identification attack: An attacker can use FB to identify citizens belonging to a sensitive social group defined by its religious belief, sexual orientation, political preference, etc. To this end, an attacker just needs to replicate a phishing-like attack [14]. The attacker would configure a campaign targeting a sensitive audience (e.g., people interested in *homosexuality*) using a fancy advertisement that serves as bait to attract the targeted users to the attacker's webpage (e.g., the ad promises the user will win an iPhone X if she clicks on the ad). If the user clicks on the ad, she will be redirected to the attacker's webpage. Once there, the attacker can use different techniques exploited in phishing attacks [14] persuading the user to provide some personal data that would reveal her identity. For instance, in the example of the iPhone X giveaway, the landing page can show a message congratulating the user for winning the phone requesting that the user provides personal data (name, address, phone number, etc.) for shipping purposes.

A recent study [13] ran experiments implementing email-based phishing attacks in which 9% of the users posted their credentials (username and password) to the phishing site (i.e., attacker's landing page). Using as a reference this success rate for phishing attacks and the results from the ad campaigns described in Section 7, we can make a ball-park estimation of the cost of identifying users tagged with expert-verified sensitive ad preferences. We spent €35 on our ad campaigns to reach 26K

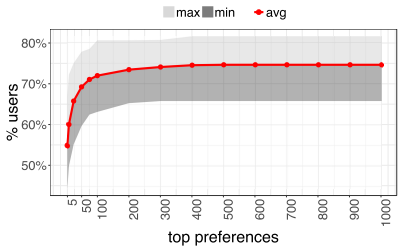

Figure 5: FFB (C,N) for values of N ranging between 1 and 1000. The figure reports the min, average and max FFB value across the 28 EU countries.

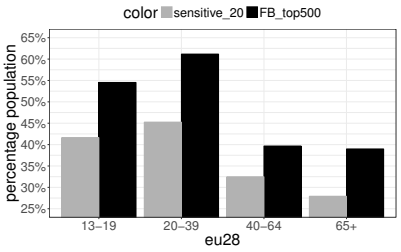

Figure 6: Percentage of EU FB users assigned at least one of the Top 500 (black) and 20-very sensitive (grey) ad preferences in the following age groups: 13-19, 20-39, 40-64, 65+.

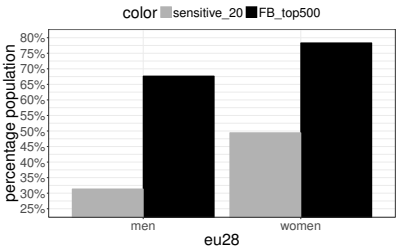

Figure 7: Percentage of EU FB users assigned at least one of the Top 500 (black) and 20-very sensitive (grey) ad preferences in the following gender groups: Men, Women.

Ad Set Name	Reach	Impressions	Amount Spent	Location (Ad Set Settings)
Religion	7,630	7,985	€5.00 of €5.00	IT, ES, FR and DE
Political	11,025	16,537	€10.00 of €10.00	IT, ES, FR and DE
Sexuality	7,314	7,367	€20.00 of €20.00	IT, ES, FR and DE
Results from 3 ad sets	26,458 People	31,889 Total	€35.00 Total Spent	

Figure 8: FB report from the 3 ad campaigns we ran targeting users based on sensitive ad preferences.

users from which 2.34K (according to the 9% reference success rate) may provide personal information on the attacker's webpage that could reveal their identity. Based on this, identifying an arbitrary member of the group may be as cheap as €0.015. Even if we consider a success rate two orders of magnitude smaller (0.09%), the cost would be €1.5 per user.

The estimated cost to reveal the identity of users based on potentially sensitive personal data is rather low considering the serious privacy risks users may face. For instance, (i) in countries where homosexuality is considered illegal or immoral governments or other organizations could obtain the identity of people that are likely homosexual (e.g., interested in *homosexuality, LGBT*, etc.); (ii) neo-Nazi organizations could identify people in specific regions (by targeting a town or even a zip code) that are likely Jewish (e.g., interested in *Judaism, Shabbat*, etc.); (iii) health insurance companies could try to identify people that may have non-profitable habits (e.g., interested in *tobacco, fast food*, etc.) or health problems (e.g., *food intolerance*) to reject them as clients or charge them more for health insurance. Users may face the negative consequences of such phishing-like attacks even if FB has wrongly labeled them with some sensitive ad preference.

In summary, although Facebook does not allow third parties to identify individual users directly, ad preferences can be used as a very powerful proxy to perform identification attacks[18] based on potentially sensitive personal data at a low cost. Note that we have simply described this ad-based phishing attack but have not implemented it due to the ethical implications.

9 FDVT extension to inform users about their potentially sensitive ad preferences

The results reported in previous sections motivate a need for solutions that make users aware of the use of sensitive personal data for advertising purposes. To this end, we have extended the FDVT browser extension to inform users about the potentially sensitive ad preferences that FB has assigned them: (i) we have built a classifier to automatically tag ad preferences assigned to FDVT users as sensitive or non-sensitive; (ii) we have modified the FDVT back-end and front-end to incorporate this new feature.

9.1 Automatic binary classifier for sensitive ad preferences

We rely on the methodology described in Section 5 to compute the semantic similarity between ad preferences and sensitive personal data categories (i.e., politics, religion, health, ethnicity and sexual orientation). Recall that each ad preference is assigned a semantic similarity score that ranges between 0 (lowest) and 1 (highest). To build an automatic binary classifier we have to define a threshold so that ad preferences over (below) it are classified as sensitive (non-sensitive).

[18] The described attack can be implemented on any advertising platform allowing advertisers to target users based on sensitive personal data.

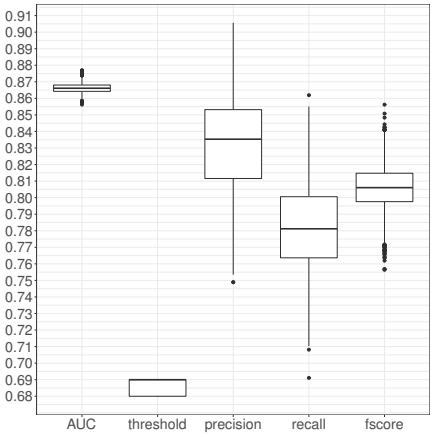

Figure 9: AUC, precision, recall and F-score for the optimal threshold to automatically classify an ad preference as sensitive or non-sensitive. The figures shows the results obtained from 5000 iterations across different randomly chosen training and validation data subsets.

To set this threshold, we use the automatically filtered dataset from Section 5.1.2. It includes 4452 ad preferences, where 2092 were classified as sensitive from the votes of 12 panelists (i.e., suspected sensitive subset). We follow a standard training-testing model approach. We randomly split our dataset in training and validation subsets that include 80% and 20% of the samples, respectively. The training subset is used to find the optimal threshold. In turn, we use the validation subset to assess the performance of the selected threshold. The optimal threshold is selected as the one maximizing the F-score for the training subset [24]. Moreover, we validate the performance of the selected threshold computing the precision, recall and F-score on the validation subset. We performed 5000 iterations of this process, each using different randomly chosen testing and validation subsets, to prove the robustness of the proposed binary classifier.

Figure 9 presents boxplots showing the AUC, precision, recall and F-score for the optimal threshold across the 5000 iterations. The optimal threshold remains quite stable ranging between 0.68 and 0.69. Similarly, the AUC derived from the ROC curve for our binary classifier presents a very stable result around 0.86, which is associated with good performance according to standard quality metrics [9][28].

The median precision of our binary classifier is 0.835 (min = 0.75, max = 0.90) and the median recall is 0.78 (min = 0.70, max = 0.86).

Even though the classifier may be imperfect, it still may achieve the goal of increasing collective awareness among FB users regarding the potential use of sensitive personal data for advertising purposes.

Potentially sensitive interests in your profile:

Preference Name	Addition	Deletion	Description	Status
Democracy	2017-06-12	--	You have this preference because you liked a Page related to Democracy.	Active
Homosexuality	2017-09-25	--	You have this preference because you liked a Page related to Homosexuality.	Active
Socialism	2017-09-28	--	You have this preference because you liked a Page related to Socialism.	Active
Veganism	2017-11-18	--	You have this preference because you clicked a Page related to Veganism.	Active
Bible	2017-12-23	--	This is a preference you added.	Active
Pregnancy	2017-05-20	2017-07-10	You have this preference because you installed an app related to Pregnancy.	Deleted
Quran	2017-05-20	2017-08-30	You have this preference because you liked a Page related to Quran.	Deleted

Figure 10: Webpage displaying sensitive preferences.

9.2 System implementation

FDVT Backend: We computed the semantic similarity score for all ad preferences stored in our database. For ad preferences with a similarity score ≥ 0.69, we classify them as sensitive and add them to a blacklist.[19] Each time a FDVT user starts a session in FB we retrieve her updated set of ad preferences and compare them with the blacklist to obtain a list of ad preferences linked to potentially sensitive personal data. We store the history of potentially sensitive ad preferences assigned to the user to notify her of those preferences that FB has removed. Finally, every time a user is assigned a new ad preference that is not already in our database, we compute its semantic similarity score and include it in the blacklist if the ad preference is classified as sensitive.

FDVT User Interface: We have introduced a new button in the FDVT extension interface with the label *"Sensitive FB Preferences"*. When a user clicks on that button, we display a web page listing the potentially sensitive ad preferences included in the user's ad preference set. Figure 10 shows an example of this webpage. We provide the following information for each potentially sensitive ad preference: (i) Ad preference name, (ii) Addition date, (iii) Deletion date (only for removed ad preferences), (iv) Description, which indicates the reason why FB has assigned that ad preference to the user, and (v) Status, either active (highlighted in green) or deleted (highlighted in red).

10 Related work

We focus on prior work that addresses issues associated with sensitive personal data in online advertising, as well as recent work that analyzes privacy and discrimination issues related to FB advertising and ad preferences.

Carrascosa et al. [4] propose a new methodology to quantify the portion of targeted ads received by Internet users while they browse the web. They create bots, referred to as *personas*, with very specific interest profiles (e.g., persona interested in cars) and measure how many of the received ads actually match the specific interest of the analyzed persona. They create personas based on sensitive personal data (e.g., health) and demonstrate that

[19] The value of the optimal threshold may change over the time since it will be recomputed periodically.

they are also targeted with ads related to the sensitive information used to create the persona's profile. Castelluccia et al.[5] show that an attacker that gets access (e.g., through a public WiFi network) to the Google ads received by a user could create an interest profile that could reveal up to 58% of the actual interests of the user. The authors state that if some of the unveiled inserts are sensitive, it could imply serious privacy risks for users.

Venkatadri et al. [26] and Speicher et al. [25] exposed privacy and discrimination vulnerabilities related to FB advertising. In [26], the authors demonstrate how an attacker can use Facebook third-party tracking JavaScript to retrieve personal data (e.g., mobile phone numbers) associated with users visiting the attacker's website. Moreover, in [25] they demonstrate that sensitive FB ad preferences can be used to apply negative discrimination in advertising campaigns (e.g., excluding people based on their race). The authors also show that some ad preferences that initially may not seem sensitive could be actually used to discriminate in advertising campaigns (e.g., excluding people interested in *Blacknews.com* that are potentially black people).

Finally, Andreou et al. [3] analyze whether the reasons FB uses to explain why a user is targeted with an ad are aligned with the actual audience the advertiser is targeting. To do this, they analyze the explanation that Facebook includes in each delivered ad referred to as *"Why Am I Seeing this Ad"*. This explanation describes the target audience associated with the delivered ad. Out of the analysis of 79 ads, they conclude that in many cases the provided explanations are incomplete and sometimes misleading. They also perform a qualitative analysis related to the ad preferences assigned to FB users based on a small dataset including 9K ad preferences distributed across 35 users. They conclude that the reasons why ad preferences are assigned are vague.

In summary, the existing literature suggests that the online advertising ecosystem (beyond Facebook) exploits sensitive personal information for commercial purposes. In addition, previous work highlights several privacy, discrimination and transparency issues associated with FB ad preferences. Our work complements this body of literature quantifying the number of users in FB that may be exposed to the commercial exploitation of their sensitive personal data.

11 IRB and FDVT users' consent

The Ethics committee of the authors' institution has provided IRB approval to conduct the implementation of the FDVT and the research activities derived from it.

To comply with the most rigorous ethics and legal standards, during the installation process of the FDVT, a user has to: (i) read and accept the Terms of use[20] and privacy policy,[21] and (ii) grant explicit permission to use the information stored (in an anonymous manner) for research purposes.

Finally, it is also worth noting that we did not gather any information (neither personal nor non-personal) from those users who clicked on the ads we used in the FB advertising campaigns described in Section 7.

12 Conclusion

Our findings suggest that Facebook commercially exploited potentially sensitive personal data for advertising purposes through the ad preferences that it assigns to its users. Facebook has already been fined in Spain for this practice. The GDPR became enforceable on May 25, 2018. We studied the potentially sensitive personal data that FB assigned to EU users in the period prior to this date. The results reveal that the portion of affected EU FB users is as high as 73% (40% of EU citizens). We illustrate how FB users that have been assigned sensitive ad preferences could face risks, like low-cost targeted attacks seeking to identify such users. The results of our paper urge a quick reaction from Facebook to eliminate all ad preferences that can be used to infer the political orientation, sexual orientation, health conditions, religious beliefs or ethnic origin of a user for two reasons: (i) this may avoid Facebook running afoul of Article 9 of the GDPR, and (ii) it may protect users from threats that exploit this sensitive data.

Acknowledgements

J.G. Cabañas acknowledges funding from the Ministerio de Economía, Industria y Competitividad (Spain) through the project TEXEO (TEC2016-80339-R) and the Ministerio de Educación, Cultura y Deporte (Spain) through the FPU Grant (FPU16/05852). A. Cuevas acknowledges funding from the Ministerio de Economía, Industria y Competitividad (Spain) and the European Social Fund (EU) through the Ramón Y Cajal Grant (RyC-2015-17732). R. Cuevas acknowledges funding from the European H2020 project SMOOTH (786741). We would also like to thank legal experts that have provided very valuable feedback for this work.

References

[1] Directive 95/46/EC. Eur-lex.europa.eu. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:31995L0046.

[20] https://www.fdvt.org/terms_of_use/

[21] https://www.fdvt.org/privacy_agreement.html

[2] Google and Facebook tighten grip on us digital ad market. Emarketer.com, Sep 2017. https://www.emarketer.com/Article/Google-Facebook-Tighten-Grip-on-US-Digital-Ad-Market/1016494.

[3] ANDREOU, A., VENKATADRI, G., GOGA, O., GUMMADI, K. P., LOISEAU, P., AND MISLOVE, A. Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations. In *NDSS 2018, Network and Distributed Systems Security Symposium 2018, 18-21 February 2018, San Diego, CA, USA* (San Diego, ÉTATS-UNIS, 02 2018).

[4] CARRASCOSA, J. M., MIKIANS, J., CUEVAS, R., ERRAMILLI, V., AND LAOUTARIS, N. I always feel like somebody's watching me: Measuring online behavioural advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies* (New York, NY, USA, 2015), CoNEXT '15, ACM, pp. 13:1–13:13.

[5] CASTELLUCCIA, C., KAAFAR, M.-A., AND TRAN, M.-D. Betrayed by your ads! In *International Symposium on Privacy Enhancing Technologies Symposium* (2012), Springer, pp. 1–17.

[6] DE PROTECCIÓN DE DATOS, A. E. The spanish dpa fines facebook for violating data protection regulations, 11 September 2017. http://www.agpd.es/portalwebAGPD/revista_prensa/revista_prensa/2017/notas_prensa/news/2017_09_11-iden-idphp.php.

[7] ERIKSON, E. H., AND ERIKSON, J. M. *The life cycle completed (extended version)*. WW Norton & Company, 1998.

[8] EU. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 27 April 2016. http://eur-lex.europa.eu/eli/reg/2016/679/oj.

[9] FAWCETT, T. An introduction to roc analysis. *Pattern recognition letters 27*, 8 (2006), 861–874.

[10] FLEISS, J. L. Measuring nominal scale agreement among many raters. *Psychological bulletin 76*, 5 (1971), 378.

[11] FLEISS, J. L., LEVIN, B., AND PAIK, M. C. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[12] GONZÁLEZ CABAÑAS, J., CUEVAS, Á., AND CUEVAS, R. FDVT: Data Valuation Tool for Facebook users. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, CO, USA, 2017), ACM, pp. 3799–3809.

[13] HAN, X., KHEIR, N., AND BALZAROTTI, D. Phisheye: Live monitoring of sandboxed phishing kits. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2016), CCS '16, ACM, pp. 1402–1413.

[14] HONG, J. The state of phishing attacks. *Commun. ACM 55*, 1 (Jan. 2012), 74–81.

[15] KORPUSIK, M., COLLINS, Z., AND GLASS, J. Semantic mapping of natural language input to database entries via convolutional neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* (2017), IEEE, pp. 5685–5689.

[16] LANDIS, J. R., AND KOCH, G. G. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[17] MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[18] MIKOLOV, T., SUTSKEVER, I., CHEN, K., CORRADO, G. S., AND DEAN, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), pp. 3111–3119.

[19] MIKOLOV, T., YIH, W.-T., AND ZWEIG, G. Linguistic regularities in continuous space word representations. In *hlt-Naacl* (2013), vol. 13, pp. 746–751.

[20] NARASIMHAMURTHY, A. Theoretical bounds of majority voting performance for a binary classification problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*, 12 (2005), 1988–1995.

[21] OPINION, T., AND SOCIAL. Special eurobarometer 431 data protection, 2015. http://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_431_en.pdf.

[22] PANCHENKO, A. Best of both worlds: Making word sense embeddings interpretable. In *LREC* (2016).

[23] PENNINGTON, J., SOCHER, R., AND MANNING, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.

[24] RICCI, F., ROKACH, L., SHAPIRA, B., AND KANTOR, P. B. *Recommender Systems Handbook*, 1st ed. Springer-Verlag New York, Inc., New York, NY, USA, 2010.

[25] SPEICHER, T., ALI, M., VENKATADRI, G., RIBEIRO, F. N., ARVANITAKIS, G., BENEVENUTO, F., GUMMADI, K. P., LOISEAU, P., AND MISLOVE, A. Potential for discrimination in online targeted advertising.

[26] VENKATADRI, G., LIU, Y., ANDREOU, A., GOGA, O., LOISEAU, P., MISLOVE, A., AND GUMMADI, K. P. Privacy risks with Facebook's PII-based targeting: Auditing a data broker's advertising interface. In *S&P 2018, IEEE Symposium on Security and Privacy, 20-24 May 2018, San Francisco, CA, USA* (San Francisco, ÉTATS-UNIS, 05 2018).

[27] WEISCHEDEL, R., PALMER, M., MARCUS, M., HOVY, E., PRADHAN, S., RAMSHAW, L., XUE, N., TAYLOR, A., KAUFMAN, J., FRANCHINI, M., ET AL. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA* (2013).

[28] ZHU, W., ZENG, N., WANG, N., ET AL. Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. *NESUG proceedings: health care and life sciences, Baltimore, Maryland 19* (2010).

Appendix

A GDPR exceptions for processing sensitive personal data

Below we list the exceptions included in GDPR Article 9 that allow processing sensitive information. In the exceptions text, the term data subject refers to users in the context of FB and the term data controller refers to FB itself. To the best of our knowledge none of the GDPR exemptions for processing sensitive personal data would apply to FB sensitive ad preferences.

(a) *"the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject"*.

(b) *"processing is necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of employment and social security and social protection law in so far as it is authorised by Union or Member State law or a collective agreement pursuant to Member State law providing for appropriate safeguards for the fundamental rights and the interests of the data subject"*.

(c) *"processing is necessary to protect the vital interests of the data subject or of another natural person where the data subject is physically or legally incapable of giving consent"*.

(d) *"processing is carried out in the course of its legitimate activities with appropriate safeguards by a foundation, association or any other not-for-profit body with a political, philosophical, religious or trade union aim and on condition that the processing relates solely to the members or to former members of the body or to persons who have regular contact with it in connection with its purposes and that the personal data are not disclosed outside that body without the consent of the data subject"*.

(e) *"processing relates to personal data which are manifestly made public by the data subject"*.

(f) *"processing is necessary for the establishment, exercise or defense of legal claims or whenever courts are acting in their judicial capacity"*.

(g) *"processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject"*.

(h) *"processing is necessary for the purposes of preventive or occupational medicine, for the assessment of the working capacity of the employee, medical diagnosis, the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union or Member State law or pursuant to contract with a health professional and subject to the conditions and safeguards referred to in paragraph 3"*.[22]

(i) *"processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of healthcare and of medicinal products or medical devices, on the basis of Union or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject, in particular professional secrecy"*.

(j) *"processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject"*.

B Spanish DPA resolution related to FB fine

In this appendix, we list the main elements included in the Spanish DPA resolution associated with the €1.2M fine imposed on FB for violating the Spanish data protection regulation.

- *The Agency notes that the social network collects, stores and uses data, including specially protected data, for advertising purposes without obtaining consent.*

- *The data on ideology, sex, religious beliefs, personal preferences or browsing activity are collected directly, through interaction with their services or from third party pages without clearly informing the user about how and for what purpose will use those data.*

- *Facebook does not obtain unambiguous, specific and informed consent from users to process their data since the information it offers is not adequate*

- *Users' personal data are not totally canceled when they are no longer useful for the purpose for which they were collected, nor when the user explicitly requests their removal.*

[22] Paragraph 3 can be found in [8]

- *The Agency declares the existence of two serious and one very serious infringements of the Data Protection Law and imposes on Facebook a total sanction of 1,200,000 euros.*

- *The AEPD is part of a Contact Group together with the Authorities of Belgium, France, Hamburg (Germany) and the Netherlands, that also initiated their respective investigation procedures to the company.*

C Facebook terms of service and advertising policy

FB users agree to the Facebook Terms of Service[23] when opening a FB account. This is the entry document where users are informed what FB is doing with their personal data. However, in order to better understand the details regarding FB data management users are redirected to another document referred to as Data Policy.[24] We found three sections very relevant for our research in the Terms of Service document:

Section 16. Special Provisions Applicable to Users Outside the United States. This section includes the following clause *"You consent to have your personal data transferred to and processed in the United States."* While this grants FB sufficient permission to process and store personal data, the GDPR and prior data protection regulations in some EU countries establish a clear difference between personal data and *"specially protected"* or *"sensitive"* personal data. To the best of our knowledge, FB does not obtain explicit permission specifically to process and store sensitive personal data.

Section 9. About Advertisements and Other Commercial Content Served or Enhanced by Facebook. In this section, users are informed that FB can use the user information, name, picture, etc. for advertising and commercial purposes.

Section 10. Special Provisions Applicable to Advertisers . Advertisers are forwarded to two more documents: Self-Serve Ad Terms[25] (not very relevant for our research) and Advertising Policies.[26] The latter document includes 13 sections from which Section 4.12[27] (4-Prohibited Content; 12- Personal attributes) is very relevant for our paper. Section 4.12 states: *"Ads must not contain content that asserts or implies personal attributes. This includes direct or indirect assertions or implications about a person's race, ethnic origin, religion, beliefs, age, sexual orientation or practices, gender identity, disability, medical condition (including physical or mental health), financial status, membership in a trade union, criminal record, or name."*. Examples of what content is allowed and what content is prohibited are provided in the Advertising Policies.

D Spacy

Spacy is a free open source package for advance NLP operations. Spacy offers multiple NLP features such as information extraction, natural language understanding, deep learning for text, semantic similarity analysis, etc., which are accomplished through different predefined models. To conduct our analysis, we leverage the "similarity" feature of Spacy that allows comparing two words or short text providing a semantic similarity value ranging between 0 (lowest) and 1 (highest). This feature computes similarity using the so-called Glove (Global vectors for word representation) method [23]. Gloves are multi-dimensional meaning representations of words computed using word2vec [17][18][19].

Spacy word vectors are trained using a large corpus of text incorporating a rich vocabulary. In addition, Spacy also takes into account context to define the representation of a word, which allows Spacy to better identify its meaning considering the surrounding words. Spacy offers different models to optimize the semantic similarity computation. We have chosen the model *en_core_web_md*[28] because it optimizes the similarity analysis between words and short sentences, which matches the nature of ad preferences names. The chosen model is an English multi-task Convolutional Neural Network (CNN) trained on OntoNotes [27] with GloVe vectors that are in turn trained on Common Crawl.[29] Common Crawl is an open source repository for crawling data. The model uses word vectors, context-specific token vectors, POS (part-of-speech) tags, dependency parse and named entities.

E Ad campaigns compliance with Facebook terms of service

Figures 11 and 12 show the two ads we used in our campaigns. These ads refer to our FDVT browser extension

[23] `https://www.facebook.com/terms.php` (accessed December 19, 2017)

[24] `https://www.facebook.com/about/privacy/` (accessed December 19, 2017)

[25] `https://www.facebook.com/legal/self_service_ads_terms` (accessed December 19, 2017)

[26] `https://www.facebook.com/policies/ads/` (accessed December 19, 2017)

[27] `https://www.facebook.com/policies/ads/prohibited_content/personal_attributes` (accessed December 19, 2017)

[28] `https://spacy.io/models/en#en_core_web_md`

[29] `http://commoncrawl.org/`

Figure 11: FDVT ad 1

Figure 12: FDVT ad 2

and thus they do not include content that asserts or implies personal attributes. Indeed, the landing page where users were redirected in case they clicked on any of these ads is the webpage of the FDVT project.[30]

In the experiments, we did not record any information from those users clicking the ads and visiting our landing page. The only information we use in this paper is that provided by FB through the reports it offers to advertisers related to their ad campaigns.

[30] `https://www.fdvt.org/`