Collaborative Ad Transparency: Promises and Limitations

Eleni Gkiouzepi, Athanasios Andreou, Oana Goga, Patrick Loiseau

Collaborative Ad Transparency: Promises and Limitations

Eleni Gkiouzepi*, Athanasios Andreou†, Oana Goga‡, Patrick Loiseau§
*Technical University of Berlin; gkiouzepi@tu-berlin.de
†Algorithmic Transparency Institute; athanasios.andreou@ati.io
‡CNRS, Inria, Institut Polytechnique de Paris; oana.goga@cnrs.fr
§Inria, FairPlay team; patrick.loiseau@inria.fr

***Abstract*—Several targeted advertising platforms offer transparency mechanisms, but researchers and civil societies repeatedly showed that those have major limitations. In this paper, we propose a *collaborative ad transparency* method to infer, without the cooperation of ad platforms, the targeting parameters used by advertisers to target their ads. Our idea is to ask users to donate data about their attributes and the ads they receive and to use this data to infer the targeting attributes of an ad campaign. We propose a Maximum Likelihood Estimator based on a simplified Bernoulli ad delivery model. We first test our inference method through controlled ad experiments on Facebook. Then, to further investigate the potential and limitations of collaborative ad transparency, we propose a simulation framework that allows varying key parameters. We validate that our framework gives accuracies consistent with real-world observations such that the insights from our simulations are transferable to the real world. We then perform an extensive simulation study for ad campaigns that target a combination of two attributes. Our results show that we can obtain good accuracy whenever at least ten monitored users receive an ad. This usually requires a few thousand monitored users, regardless of population size. Our simulation framework is based on a new method to generate a synthetic population with statistical properties resembling the actual population, which may be of independent interest.**

I. Introduction

Online advertising platforms have access to massive amounts of user data, which allows them to provide advertisers with powerful ways to target specific users according to a detailed range of characteristics and delivery optimization techniques. Advertisers can target users interested in specific topics such as tennis and rock climbing (*attribute-based targeting*), that visited their website (*retargeting*), or that are similar to their customers (*lookalike audiences*), to name a few [1].

Governments and NGOs around the world [2]–[7], have been pushing online platforms to make the inner workings of their advertising systems more transparent to the public. As a result, several ad platforms implemented transparency mechanisms, such as Facebook's "Why am I seeing this ad?". While this is a positive step, recent reports emphasize that the information provided in these explanations is limited and only shows one (of many) targeting parameters used by the advertiser [8]–[11]. Similar limitations exist in Political Ad Libraries (public repositories aggregating political ads run on the platform). For instance, Facebook only provides information about the location, age, and gender of users who received the ad and gives no information on the actual targeting parameters used by the advertisers. Hence, despite positive developments, *we still lack grounded information about the targeting parameters used to deliver ads*.

To increase the level of transparency of online platforms, researchers and lawmakers have called for *independent auditing systems* that allow citizens to participate in the process of building a healthier web by donating data about the content they see online [12], [13]. There are several existing tools such as AdAnalyst [14], Ad Observer [15], and WhoTargetsMe [16] that collect ads from volunteers and have been installed by thousands of users. These tools showed their utility on multiple occasions by providing data essential to show limitations with Ad Libraries [17], [18], or the existence of discriminating, manipulatory and illegal advertising [17], [19]–[21].

In this paper, *we investigate whether we can infer, through independent auditing and without the cooperation of ad platforms, the targeting parameters used by advertisers to target their ads*. Specifically, we focus on attribute-based targeting where an advertiser targets users that satisfy a combination of selected attributes (called the *targeting formula*), and we investigate whether data collected from a set of *volunteer monitored users* (their attributes and the ads they receive) can be used to infer the targeting formula—we term this *collaborative ad transparency*. Such a solution would be practical because we can easily monitor the attributes of users (from their Ad Preferences page where platforms list all attributes they inferred about a user) and the ads they receive (through browser extensions such as AdAnalyst and Ad Observer). However, we currently do not know whether such inference can achieve reasonable accuracy and in which conditions.

Our methodological and experimental contributions are:

(1) We propose an estimator that infers the targeting formula by analyzing the attributes of users who received the ad and of users who did not (Section II). Our estimator uses the maximum likelihood principle based on a simplified Bernoulli model of ad delivery where all users satisfying the attributes in the targeting formula receive the ad with equal probability (encoding the ad campaign budget). To handle realistic cases where the campaign budget is unknown, we also propose an estimation procedure based on Expectation-Maximization.

(2) We propose an experimental setup that allows us to test the accuracy of our estimator in the real world (Section III). For this, we recruited 420 Facebook users and asked them to install

a browser extension we developed that can monitor the ads they receive and the attributes Facebook inferred about them (their Ad Preference page). Then, we assume the role of the advertiser and create ad campaigns that target users with various attributes; this gives us access to the ground truth targeting formula. We performed 79 real-world controlled experiments on Facebook, out of which 45 reached three or more of our monitored users. Our results show that we infer the targeting formula correctly in 17 cases. However, if we only look at experiments where our ad has been received by ten or more monitored users, the accuracy increases to 65%. Hence, despite making a number of simplifying assumptions, the real-world experiments show an interesting potential for collaborative ad transparency, provided enough monitored users receive the ad.

(3) Providing a detailed understanding of the accuracy of collaborative ad transparency under different conditions (e.g., number of monitored users, prevalence of the targeting attributes) entails a significant difficulty: it is too costly to do it through controlled experiments on the real platform. To bypass this difficulty, we propose a *simulation framework* that enables us to analyze the impact of various parameters on inference accuracy extensively. The framework is based on a synthetic users population and a simplified Bernoulli model to generate ground truth data (i.e., simulate which users receive the ad).

A key challenge for obtaining results with external validity is to base the simulations on a synthetic population that resembles the real population (Section IV). We propose a method based on correlated binary vectors to generate a synthetic population of users such that the distribution of attributes matches a predefined distribution at second-order (probability of having any pair of attributes). We apply it to generate populations of 10^6 users that match the distribution of Facebook's population over 322 attributes. Our proposed method generates users that are i.i.d. and is suitable for generating high-dimensional data. Further, the population statistical characteristics match beyond second-order (third-order and number of attributes per user).

(4) To validate that our simulation framework captures advertising on the real platform soundly, we recreate the 45 real-world controlled ad experiments (Section V). Then, we confront the estimator's accuracy in the wild with the estimator's accuracy in our simulations. The accuracy is consistent in 82.2% of all ad experiments. Focusing on experiments where ads are received by 10 or more users, the simulation accuracy is 86.1% where it was only 65% in the Facebook experiments. However, recreating the simulations in a way that mimics the bias in the set of monitored users brings the accuracy down to 73.8%. Not only does it explain the major factor behind the accuracy discrepancy, but it also shows that our simulation framework allows quantifying the effect on accuracy of bias in the set of monitored users. Finally, this indicates that the Bernoulli model of ad delivery, while simplifying, does not majorly affect the estimated accuracy in our simulation study.

(5) We perform extensive simulations (Section VI). We consider the case where the targeting formula consists of the conjunction of two attributes (but the framework extends to more complex formulas), and we estimate the inference accuracy in relation to the targeting formula, ad budget, and number of monitored users. Our results show that our estimators remain equally accurate even if we do not know the real ad budget. We observe an accuracy of over 75%, on average over a wide range of targeting formulas and ad budgets, as long as the ad campaign reaches 0.005 of the entire population while only monitoring 700 users.[1] We also see that if an ad is received by ten or more monitored users, we can expect an inference accuracy of over 90% irrespective of the number of monitored users. The number of monitored users plays an indirect role by impacting the likelihood for a given campaign that enough monitored users receive the ad. This is positive since (i) good accuracy may be achievable within reasonable circumstances and (ii) it allows us to evaluate the confidence in the inference from observed data. Finally, our results show that the number of monitored users required for good accuracy does not increase with the population size, which gives scalability.

Overall, the combination of our simulations and real-world experiments allows us to shed light on the conditions under which collaborative ad transparency may work and the challenges to get good accuracy. Our study is based on a number of simplifying assumptions that enable a rigorous analysis, in particular our Bernoulli ad delivery model. In practice, ad delivery is more complex. We provide a detailed discussion of the impact of these assumptions and of the limitations of collaborative ad transparency in Section VIII.

Our study is inspired by Facebook's ad platform, but it applies broadly as attribute-based targeting is common in online advertising. Understanding the promises and limits of collaborative ad transparency is necessary even if Facebook were to provide complete ad targeting explanations: (1) it applies to other less-cooperative ad platforms; (2) it is critical to continuously audit the platform-provided explanations.

All of our code is publicly available at https://gitlab.inria.fr/oagoga/collaborative-transparency-IEEE-SP2023/.

II. Model and Inference Method

We start by abstracting the complexity of targeted advertising through a simplified representation of the ad platform's population, and the ad targeting and ad delivery processes. Our representation models advertising on Facebook, but it generalizes to other ad platforms. We then describe our method to infer the targeting formula from data observed from a set of monitored users based on the maximum likelihood principle.

This section makes several assumptions and simplifications. We discuss throughout the paper their implications on results, and discuss how realistic they are and how to extend them in Sec. VIII. Table I (appendix) summarizes our notation.

A. Representation of Targeted Advertising

1) The platform's user population: Facebook collects data from its users while they are browsing the Internet (e.g., web pages visited) and when they use Facebook (e.g., user-filled data, likes, clicks, comments) and infers various *attributes*

[1] Ads that reach 0.005 of users in a country with 40 million Facebook users would cost about $1,330$€.

about them, which can be demographic (e.g., education, place of birth), behavioral (e.g., users of mobile devices), or interests (e.g., hobbies, food preferences). For simplicity in our investigation, we restrict to interest attributes, which are often deemed more sensitive [22]. The framework, however, remains general. We denote by $\mathcal{A}$ the set of all attributes, of size $A = |\mathcal{A}|$. We will denote by a_j ($j = 1, \cdots, A$) the attributes.

Let us denote by $\mathcal{N}$ the population of Facebook's users, and by $N = |\mathcal{N}|$ its size. For every user $i \in \mathcal{N}$, the platform infers a binary variable for each attribute $a_j \in \mathcal{A}$, which we denote by u_j^i: $u_j^i = 1$ if the user satisfies the attribute (i.e., the user is inferred to have the corresponding interest) and $u_j^i = 0$ otherwise. Hence, for the purpose of our study, each user $i \in \mathcal{N}$ simply corresponds to a binary (row) vector $u^i = [u_j^i]_{j\in\{1,\cdots,A\}}$ of size A that encodes which attributes the user satisfies; and the population $\mathcal{N}$ is described as an $N \times A$ binary matrix $U = [u_j^i]_{i\in\{1,\cdots,N\},j\in\{1,\cdots,A\}}$ where rows represent users and columns represent attributes.

2) The ad targeting process: We focus on attribute-based targeting where an advertiser selects a targeted audience by selecting the attributes it should satisfy (Facebook allows advertisers to choose from a predefined list of curated attributes). We restrict our investigation to cases where the targeting formula is a *combination of two attributes*: $a_j \wedge a_l$, $a_j, a_l \in \mathcal{A}$. Note, however, that the inference process we propose extends to other formulas without fundamental difficulty. We denote by $\mathcal{T}$ the set of all possible targeting formulas of this form.

When launching an ad campaign, besides the targeting formula, the advertiser specifies several additional parameters, including the campaign budget, bid cap, and campaign duration. We abstract these away in the simplest possible way: we assume that the advertiser defines an expected number K of users who will be shown the ad. We will usually express K as a percentage of the targeted audience (i.e., of N_θ below).

3) The ad delivery process: Let us denote by $\theta \in \mathcal{T}$ the targeting formula selected by the advertiser for the ad campaign, and let us denote by $\mathcal{N}_\theta \subset \mathcal{N}$ the set of users who satisfy the targeting formula θ. For each user $i \in \mathcal{N}$, we define a binary variable $y^i \in \{0, 1\}$ indicating whether or not the user is shown the ad, i.e., $y^i = 1$ if the user is shown the ad and $y^i = 0$ otherwise. We assume that $y^i = 0$ for all $i \in \mathcal{N} \setminus \mathcal{N}_\theta$, that is, a user that does not satisfy the targeting formula cannot be shown the ad. We denote by $\mathcal{K}$ the set of users that are shown the ad (i.e., for which $y^i = 1$).

With this formalism, our problem formulation is: Let $\mathcal{N}_m \subset \mathcal{N}$ be the set of monitored users. We assume that we can observe the attributes u_i of users $i \in \mathcal{N}_m$ as well as whether or not they receive an ad (i.e., the value of y^i). Then the problem is: *given an ad observed by a user in $\mathcal{N}_m$, infer the targeting formula θ that the advertiser used to target the ad.*

B. The Bernoulli Model and Estimator

To solve the above problem, we propose a maximum likelihood estimator. To that end, we need a model of ad delivery that specifies for a given formula θ the probability, for each user in $\mathcal{N}_\theta$ (i.e., satisfying the targeting formula), to receive the ad. We propose a simple Bernoulli model.

1) The Bernoulli assumption: The central assumption we make in this paper is that each user satisfying the targeting formula has an equal probability of being shown the ad independent of other users. Formally, we assume that y^i for all $i \in \mathcal{N}_\theta$ are mutually independent Bernoulli random variables $y^i \sim Ber(p_\theta)$ of probability $p_\theta = K/N_\theta$, where $N_\theta = |\mathcal{N}_\theta|$. Note that, then, $|\mathcal{K}|$ is K in expectation. It follows that, for every user in $\mathcal{N}_{m,\theta} = \mathcal{N}_m \cap \mathcal{N}_\theta$ (i.e., monitored user satisfying the formula θ), y^i is also a Bernoulli random variable with parameter p_θ, independent of other users. For all users $i \in \mathcal{N}_m \setminus \mathcal{N}_{m,\theta}$, we have $y^i = 0$ with probability one by our earlier assumption that $y^i = 0$ for all $i \in \mathcal{N} \setminus \mathcal{N}_\theta$.

2) The Maximum Likelihood Estimator: Consider a given ad for which we would like to infer the targeting formula. We assume first that parameter K is known for the inference and that we can access the distribution of attributes in the whole population (i.e., compute N_θ for any θ); then for any given θ, it is possible to compute p_θ. We explain below how to adapt the inference process when K and/or the N_θ's are unknown.

Let $\hat{y}^i$ for all $i \in \mathcal{N}_m$ be the values of the labels y^i observed. With the Bernoulli model detailed above, the likelihood of observing $(\hat{y}^i)_{i\in\mathcal{N}_m}$ if the targeting formula is $\theta \in \mathcal{T}$ is

$$\mathcal{L}(\theta) = \prod_{i\in\mathcal{N}_m} P(y^i = \hat{y}^i|\theta) = \left(\prod_{i\in\mathcal{N}_m\setminus\mathcal{N}_{m,\theta}} \mathbb{1}_{\hat{y}^i=0}\right) \cdot \left(\prod_{i\in\mathcal{N}_{m,\theta}} p_\theta^{\hat{y}^i}(1-p_\theta)^{1-\hat{y}^i}\right),$$

where $\mathbb{1}$ denotes the indicator function. Intuitively this formula computes, for a particular targeting formula, the likelihood that users receiving/not receiving the ad correspond to what is actually observed. The first part ensures that the likelihood is zero if a user not satisfying the formula receives the ad.

The likelihood above does not depend on which users receive the ad, but only on how many users in $\mathcal{N}_{m,\theta}$ are shown the ad. To simplify the writing, let $\mathcal{N}_{m,r}$ be the set of monitored users who are shown the ad (i.e., such that $\hat{y}^i = 1$). Then the likelihood is

$$\mathcal{L}(\theta) = \mathbb{1}_{\mathcal{N}_{m,r}\subseteq\mathcal{N}_{m,\theta}} \cdot p_\theta^{N_{m,r}} (1 - p_\theta)^{N_{m,\theta}-N_{m,r}}, \quad (1)$$

where, as per our usual notation convention, $N_{m,r} = |\mathcal{N}_{m,r}|$.

The maximum likelihood estimator of the targeting formula, which can be computed from the observation $\hat{y}^i$ for all $i \in \mathcal{N}_m$ (or from the observation of the subset $\mathcal{N}_{m,r}$), is defined as

$$\hat{\theta} = \arg\max_{\theta\in\mathcal{T}} \mathcal{L}(\theta). \quad (2)$$

We will call it the *B-ML* estimator (*B* stands for Bernoulli).

In practice, one does not need to check all possible targeting formulas while computing the $\arg\max$; all formulas θ such that one does not have $\mathcal{N}_{m,r} \subseteq \mathcal{N}_{m,\theta}$ lead to a likelihood zero, which cannot be the max. The set of all formulas that lead to a non-zero likelihood can be easily computed: we first compute the intersection of attributes shared by all users in $\mathcal{N}_{m,r}$, denoted $\mathcal{A}_r \subset \mathcal{A}$; and then we construct all combinations of two such attributes, denoted $\mathcal{T}_r \subset \mathcal{T}$. We denote by $A_r = |\mathcal{A}_r|$ and $T_r = |\mathcal{T}_r|$ the size of these sets. In practice, we expect that $\mathcal{T}_r$ will be significantly smaller than $\mathcal{T}$ as soon as a few monitored users received the ad, since $\mathcal{A}_r$ will constrain the

possible targeting formulas. As $\mathcal{T}_r$ is a discrete set, computing the $\arg\max$ is simply done by computing the likelihood for any $\theta \in \mathcal{T}_r$ and taking the largest. We break ties at random.

3) The Expectation-Maximization Estimator: In practice, the value of K may not be known at inference time. To handle this case, we apply Expectation-Maximization (EM) [23]. EM is a general method for maximum-likelihood estimation when dealing with missing values. Here, the missing values are K and the N_θ for the corresponding targeting formula, which would enable computing the parameter $p_\theta = K/N_\theta$.

The EM algorithm runs as follows: we start with some initial value of the parameter p_θ. Then we alternate between two steps. In the E-step, we compute the conditional expectation of the likelihood given the current value of p_θ, which is the expected number of users receiving an ad and the expected number of users satisfying the ad given the current p_θ. It is obtained by taking the inner product of the observed data $N_{m,r}$ and $N_{m,\theta}$ with the current p_θ. In the M-step, we maximize this conditional likelihood; for p_θ it is simply obtained by dividing the expected number of users receiving the ad by the expected number of satisfying users calculated in E-step. This M-step gives a new parameter estimate. The E-step and M-step of the algorithm are repeated iteratively until convergence occurs. The tolerance in our experiments is set to Tol=10^{-3} and the maximum number of iterations is 100. We will refer to the estimator that uses EM as *B-EM* estimator.

III. Evaluation over real-world experiments

This section proposes measurement methodologies, experimental designs, and software that enable us to evaluate the accuracy of our estimators in the wild on an actual ad platform. We first describe how we collected the necessary data and then how we instrumented the ad platform to perform controlled ad experiments. All our experiments are run on Facebook.

A. Data Collection

To test our inference, we need a set of monitored users $\mathcal{N}_m$, for which we need to collect (i) the ads they see (i.e., the value of y^i) and (ii) the attributes they have (i.e., the u_i for users $i \in \mathcal{N}_m$). The *B-ML* estimator also requires knowledge on the number of Facebook users satisfying a formula (N_θ) for all θ. Ads and attributes of monitored users: To collect this data, we use the AdAnalyst *monitoring tool* we developed for [24]—a Chrome extension that users can install on their computers, which is publicly available [14]. After installation, the tool collects all the ads that the user receives while browsing Facebook using the technique proposed by Andreou et al. [8].

The tool also silently collects the user's Ad Preferences page's content—a page where Facebook lists all attributes they inferred about the user—once every two days. Then, to construct the profiles u^i of those users (i.e., the list of attributes that the user has), we aggregate the attributes collected in the Ad Preference page during the data collection period.

One of the critical difficulties of this data collection is to recruit users to install our monitoring tool—the *volunteer monitored users*. We recruited users in Brazil in late 2018 (around the Presidential election) and managed to have 420 active users (in total) during a period ranging from 09/11/2018 to 29/03/2019 (where active means we collected at least one ad from them during the period). This recruitment campaign is the same as the (Brazilian) campaign of [24]. We discuss our recruiting strategy in more details in Appendix D-A. As mentioned in the previous section, we restrict our analysis to interest-based attributes; there are 322 curated interest attributes in this period (i.e., $A = 322$). The median number of attributes per user in our data is 125 (STD : 46.71). We call the dataset containing the 420 user profiles $\mathcal{D}_{\texttt{attributes}}$.

Population statistics: We collect the number of Facebook users satisfying a particular formula (N_θ) from the Facebook Ads Manager. When placing an ad, the platform provides the advertiser with an estimate of the audience (i.e., number of users N_θ) satisfying a given targeting formula θ (possibly in combination with other criteria such as gender or location). Hence, we queried the system with every possible combination of one or two attributes on 09/06/2019 and gathered its worldwide monthly active users estimates. We collected this data for the 322 curated interest attributes, hence in total, we collected $51,681$ N_θ values. We call this dataset $\mathcal{D}_p$.

Datasets for the simulation framework: We use $\mathcal{D}_p$ to generate synthetic populations for the simulations. The N_θ's for formulas with a single attribute and the matrix of N_θ's for all formulas of two attributes define the second-order attribute distribution (denoted p). To validate our population generation (Section IV-3) we also use $\mathcal{D}_{\texttt{attributes}}$ to validate the number of attributes per user and the third-order probabilities.

Ethics: We describe these aspects in detail in Appendix D-B.

B. Controlled Ad Experiments

The goal of the real-world ad experiments is to see whether, at least in a few cases, our estimators can infer the targeting formula correctly (and not to do a systematic and comprehensive study). In fact, there are many reasons to expect that the *B-ML* estimator and the *B-EM* estimator will not provide correct inferences in the wild: we make several simplifying assumptions about how ads are delivered (see Sec. II-B), we do not know how Facebook optimizes the ad delivery, we do not have control over what advertisers are competing and what users are active, and our data collection method is imperfect and depends on the correctness of the information provided by Facebook in the Ad Preference and Ad Manager pages.

1) General principles: We performed real-world ad experiments on Facebook from 09/11/18 to 29/03/19, where we took the role of an advertiser and targeted Brazilian users with ads to have ground-truth data on the targeting formula. In each experiment, we launch ad campaigns targeting users with various combinations of two attributes, i.e., $\theta = a_j \wedge a_l$. Then using data we collect from the 420 monitored users, we infer the targeting formula using the *B-ML* estimator and the *B-EM* estimator and compare it with the actual targeting formula we defined in our ad campaign (i.e., the ground truth).

2) Experimental design: To have successful ad experiments, we need our ad campaigns to reach several of our monitored users since we cannot make inferences otherwise. The simplest way to increase the likelihood of reaching our monitored users is to have large ad budgets that can be used to target a large fraction of the Facebook Brazilian users that satisfy our targeting formula. However, such a budget can quickly get prohibitive in an academic setting, especially if we need to perform multiple ad campaigns. To address this challenge, we devise *three strategies to increase the chances of reaching our monitored users on an acceptable ad budget*.

1. We launched ad campaigns using attributes picked from the most active users in the previous week and shared by most monitored users. The goal of these experiments is to see if it is possible to have correct inferences, at least in a few cases. Hence, hand-picking targeting formulas is not problematic.
2. We restrict the considered population in two ways:

(i) *Location-based experiments*: We target only users from Belo Horizonte (the place with most active monitored users).

(ii) *Custom-audiences experiments*: We target our exact monitored users by providing Facebook with the list of their emails. Both strategies restrict the considered population to a subset of the whole Facebook population (which is much smaller for custom-audiences experiments). Then, the targeting formula is applied only to this subpopulation which artificially boosts our budget per monitored user. From an inference perspective, restricting the targeting to a subpopulation does not affect the accuracy of the inference since the estimators only look at the attributes of the monitored users, as long as the subpopulation has similar statistical properties (i.e., second-order attribute distribution) as the whole population (which is the case).

3. We performed ad campaigns that span multiple days (between 3 and 6 days) to increase the chances more monitored users are active during the ad campaign, and we set high bid caps ranging from 10 to 40€ per 1000 impressions, that are higher than the average–7€ [25]–to outperform competitors.

3) Results: In total, 79 ad experiments reached at least one monitored user. Table II (appendix) provides detailed information about the 45 experiments (16 location-based and 29 custom audience-based) that reached three or more monitored users. We omit the rest because we do not expect accurate inferences when only one or two monitored users received an ad. Table II provides information about the parameters used in our ad experiments (θ), the parameters observed (N_m, $N_{m,\theta}$, $N_{m,r}$, A_r, T_r), and information on the accuracy of our estimators (whether the *B-ML* estimator and the *B-EM* estimator inferred correctly the targeting formula, the inferred targeting formula, the rank of the correct targeting formula). Overall, the number of active monitored users during the different ad campaigns ranged from 193 to 280 (not all the 420 monitored users were active during all campaigns).

We observed several cases where some monitored users received our ad even if their u^i did not contain all the targeting attributes we specified. This happened for 86 cases (out of a total of 1021 cases of users receiving an ad across the 45 ad campaigns). We believe this is likely a data collection problem rather than an ad delivery bug from Facebook. First, we do not continuously monitor the Ad Preference page (we collect it once every two days); hence, we might miss the period when the attribute was present on the page. Second, the information provided by Facebook on the Ad Preference page might be incomplete [8]. For these cases, we artificially input the missing targeting attributes in u^i. We acknowledge that this may artificially ease the inference and provide an optimistic estimate. However, we will see that these results are consistent with simulations; hence we believe the effect is minimal. We discuss further the practical problems in implementing collaborative ad transparency in Sec. VI.

For the *B-ML* estimator, we use $\mathcal{D}_p$ to fill N_θ. Since in the wild we do not have knowledge of K, we test the estimator with various $\hat{K}$ (we use a different notation to distinguish from the real K), where $\hat{K} \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1\}$. The *B-ML* estimator achieves the highest accuracy when using $\hat{K} = 25\%$.

In terms of accuracy, both the *B-ML* estimator and *B-EM* estimator inferred correctly 0 out of the 16 location-based ad experiments and 17 out of the 29 custom audiences experiments. While the aggregated accuracy overall might not seem high, these results are rather remarkable because they bring concrete proof that it is possible to infer the targeting formula in the real world despite the assumptions we make, and the complexity and the lack of control we have over the ad delivery process. In addition, we can see the inference difficulty as the values of T_r are large; that is, our estimator can pick the correct formula out of a large number of valid options in a non-negligible fraction of cases.

Furthermore, the *B-EM* estimator gives the same accuracy as the *B-ML* estimator (that has a hand-picked optimal parameter K)—although the experiments with correct inference are not necessarily the same. This is also remarkable as the *B-EM* estimator does not need K and N_θ as a parameter and can infer the targeting formula only from the observed data.

Table II suggest that one likely reason why we achieve higher accuracy in the custom audience experiments is differences in $N_{m,r}$ which is much higher for custom audience ($MED : 27$) than location-based experiments ($MED : 4$). In fact, if we only look at ad experiments that reached 10 or more users, we make correct inferences in 65% of cases.

While these results are encouraging, they are limited by the number of experiments that are feasible on the real platform. To understand precisely under which conditions our method can provide reasonable accuracy, we resort to simulations that allow us to vary parameters much more freely while ensuring that the results are consistent with our real-world experiments.

IV. Generation of a Representative Synthetic Population

In order to have realistic simulations, we need to base them on a synthetic population whose statistical characteristics mimic those of the true platform's population. In this section, we present and validate our method to solve this problem.

1) Problem formulation and approach: Given that each user i is defined by its binary vector of attributes $u^i = [u^i_j]_{j\in\{1,\cdots,A\}}$, the statistical properties of the population $\mathcal{N}$ are entirely defined by the probability of every combination of attributes. For any attribute $a_j \in \mathcal{A}$ we define $p_j = P(u_j = 1)$ as the probability that an arbitrary user satisfies attribute a_j (as this is independent of the user, we omit the exponent i). We denote by $q_j = (1-p_j) = P(u_j = 0)$ the complementary probability that a user does not satisfy a_j. Then, for any two attributes $a_j, a_l \in \mathcal{A}$, we define $p_{jl} = P(u_j = 1, u_l = 1)$ as the probability that a user satisfies both attributes a_j and a_l. Together with $(p_j)_{j\in\{1,\cdots,A\}}$, the matrix $(p_{jl})_{j,l\in\{1,\cdots,A\}}$ completely defines the second-order distribution. By abuse of notation, we refer to the first two orders together as simply p.

We need to generate a population of users such that their attributes distribution follows the one of actual users. *While specifying the marginal probability of each attribute is insufficient to describe the population, specifying the probability of any combination of arbitrarily many attributes is impossible due to the curse of dimensionality.* Hence, we focus on the first two orders (we will still discuss higher orders below).

Problem formulation: *Generate a sequence of binary vectors $u^i = [u^i_j]_{j\in\{1,\cdots,A\}}$, for i up to an arbitrary population size, such that the joint attribute probability matches up to the second order the distribution p from the true population.*

Naive (non satisfying) approaches: A naive solution would be to find an algorithm that deterministically fills in the attributes in a large $N \times A$ matrix such that for any pair of attributes a_j, a_l, the fraction having it is p_{jl}. However, this approach would create undesirable artificial deterministic patterns. Instead, we are looking for a method to generate each vector $u^i = [u^i_j]_{j\in\{1,\cdots,A\}}$ as an i.i.d. random variable such that the joint probability of attributes is p. A standard method for generating multivariate i.i.d. random variables is the rejection method [26]. However, it is *not suitable for high-dimensional data*. It would require generating uniform random variables over 2^A combinations, as well as making assumptions to compute whether to reject or not a given combination based only on the second-order distribution p.

Correlated binary vector approaches: We turn our focus on methods to generate sequences of correlated binary variables: [27] proposed a method that is able to accommodate efficiently a binary sequence that follows an arbitrary correlation structure; [28] proposed another efficient algorithm for generating binary data by dichotomizing a Poisson distribution. The algorithm of [27] allows for more general correlations (e.g., negative correlations that may occur in practice) and is faster when the dimension (i.e., number of attributes) is high. Hence, we adapt the algorithm based on [27] and the package released by [29] to our population generation problem.

2) Algorithm: As the correlated binary vector methods are largely unknown in our community, we describe how the algorithm of [27] works and how to adapt it to our population generation problem where the original method may not directly apply. We believe this approach is valuable to our community beyond this specific paper.

Original method: The general idea of the algorithm of [27] is to generate an A-dimensional random vector from a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ and to transform it to binary values by thresholding. Implementing such a method requires $\boldsymbol{\mu}$ and Σ adapted to the binary data that we want to generate—i.e., essentially to p—we explain next how this is done.

Let us consider a random vector $(X_1, \ldots, X_A)$, normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix Σ and let us fix the transformation to binary as $u_j = 1$ if and only if $X_j > 0$, for any $j \in \{1, \cdots, A\}$. Then $P(u_j = 1) = P(X_j > 0) = P((X_j - \mu_j) > -\mu_j) = \Phi(\mu_j)$, where $\Phi(\cdot)$ is the standard normal distribution. Hence, to ensure that $P(u_j = 1) = p_j$, we take $\mu_j = \Phi^{-1}(p_j)$—the p-th quantile—for all $j \in \{1, \cdots, A\}$. Similarly, we have $P(u_j = 1, u_l = 1) = \Phi(\mu_j, \mu_l; \rho_{jl})$, where $\Phi(\cdot, \cdot; \rho)$ is the CDF of a bivariate standard normal random variable with correlation coefficient ρ and ρ_{jl} is the correlation coefficient between X_j and X_l. To ensure that $P(u_j = 1, u_l = 1) = p_{jl}$, the equations $p_{jl} = \Phi(\mu_j, \mu_l; \rho_{jl})$ are solved for ρ_{jl}. In solving those equations, the values $\Phi(\mu_j, \mu_l; \rho)$ are obtained through Monte Carlo simulations.

Adjustments: A drawback of the algorithm by [27] is that, for an arbitrary p, there is no guarantee that the covariance matrix $\tilde{\Sigma}$ derived from the ρ_{jl} that we get from solving these equations is positive semi-definite—which we need to generate the normal random variables. In fact, in our experiments, it is not. To solve this issue, we take as the covariance matrix Σ the closest positive semi-definite matrix to $\tilde{\Sigma}$, approximated by applying the optimization algorithm [30]—this preserves the dependence structure of the distribution. Then, in summary, to generate a user, we sample a random variable $(X_1, \cdots, X_A)$ normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix Σ as computed above; and we transform it to binary by applying $u_j = 1$ if and only if $X_j > 0$. To generate a whole population, we generate each user independently with the same process. The generation process is summarized in Algorithm 1.

Code: We implemented Algorithm 1 in R, using the packages "bindata" [31] and "Matrix" [32]. The code for the generation process is public (see link in the introduction) and allows anyone to generate a synthetic population based on an arbitrary input of first and second-order probabilities.

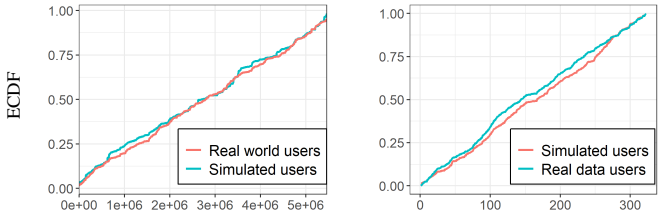
3) Validation: To validate our population generation method, we use the dataset $\mathcal{D}_{\texttt{attributes}}$ of 420 users (with $A = 322$ attributes) described in Section III-A. From this dataset, we estimate p, then we generate a population of $N = 10^6$ users using this p as an input.

Validation of statistical characteristics: To check that our generated population matches the statistical characteristics of the true population, we first compare the second-order distribution of attributes to the one that was imposed, i.e., to p. Specifically, for any $j, l \in \{1, \cdots, A\}$, let $\hat{p}_{jl}$ be the fraction, in $\mathcal{N}$, of users who satisfy both a_j and a_l, i.e., for which $u_j = 1$ and $u_l = 1$. The maximal absolute deviation $|p_{jl} - \hat{p}_{jl}|$ observed in

input : First and Second-order $(p_j)_{j\in\{1,\cdots,A\}}, (p_{jl})_{j,l\in\{1,\cdots,A\}}$
output: A matrix of users $\mathcal{N}$ of size $N \times A$

Set $\mu_j = \Phi^{-1}(p_j)$ for all $j \in \{1, \cdots, A\}$;
Compute all ρ_{jl} by solving $p_{jl} = \Phi(\mu_j, \mu_l; \rho_{jl})$ and fill covariance $\tilde{\Sigma}$;
if $\tilde{\Sigma}$ *is not positive semi-definite* **then**
  Set Σ as the nearest positive semi-definite matrix to $\tilde{\Sigma}$
else
  Set Σ as $\tilde{\Sigma}$
end
for *i = 1 to N* **do**
  Generate a normal vector x with mean $\boldsymbol{\mu}$ and covariance Σ ;
  Set $u_j^i = 1$ iif $x_j > 0$;
end

Algorithm 1: Generation of the synthetic population.



(a) Third-order probabilities (b) Nb of attributes per user

Fig. 1: Comparison of ECDFs for actual platform users in $\mathcal{D}_{\texttt{attributes}}$ and for the synthetic population generated.

our population was 0.003. Hence, as expected, the generated synthetic population matches the attribute distribution of the true population up to the second order.

Next, we compare the third-order probabilities, that is, the probabilities of the form p_{jlm} for combinations of three attributes a_j, a_l, a_m. Figure 1a shows the ECDFs of all combinations taken in the lexicographic order for the true and synthetic population (that is, the ECDF of a random variable corresponding to the index of a given combination of three attributes in the lexicographic order). It shows that the third-order distribution of the synthetic generated population matches that of the true one—even though it is not prescribed by the generation method, which is an interesting property.

Validation of the number of attributes per user: To further increase our confidence that the synthetic population resembles the true one, we compare the number of attributes per user in both. This is interesting because the equality of the distribution of number of attributes is not implied by the equality of the second-order distribution. Figure 1b shows the ECDF of the number of attributes per user for both the true population (of 420 users) and the generated population $\mathcal{N}$. We observe that they are very close to each other. We also performed a two-sample Kolmogorov-Smirnov, a Wilcoxon non-parametric test, and a permutation test on the two distributions; all show that we cannot reject the null hypothesis that the populations have identical distributions of the number of attributes. Hence, the synthetically generated population resembles the true one even beyond what is imposed by the generation method.

V. Simulation Framework

As experiments on Facebook are costly, we complement them with simulations to study the accuracy of collaborative transparency. In this section, we present the framework we use to perform our simulation study; and we show that it gives results largely consistent with our real-world experiments.

A. Key Pillars of the Simulation Framework

The goal of the simulation framework is to allow us to investigate the accuracy of our inference method in ad campaigns with various parameters (e.g., targeting formulas θ, budgets K) and where the inference is performed under various conditions (e.g., numbers of monitored users N_m); with the aim of understanding under which conditions collaborative ad transparency can provide satisfying accuracy. The simulation framework should then allow us to fix parameters arbitrarily, and then to simulate the ad delivery process in sufficient detail to test our estimator as in the controlled experiments.

Our simulation framework relies on two key pillars:

(i) *A synthetic population.* We will generate it with the method of Section IV based on dataset $\mathcal{D}_p$ so that its statistical characteristics resemble that of the real Facebook population.

(ii) *An ad delivery process* that specifies, given all the ad campaign parameters, which (synthetic) users receive the ad. For this, we will use the same Bernoulli model proposed in Sec. II-B1 for constructing the estimator. That is, we assume that each each user in $\mathcal{N}_\theta$ receives the ad with the same probability, independent of other users. Equivalently, if the number of users receiving the ad ($|\mathcal{K}|$) is fixed (i.e., we condition on a particular value of $|\mathcal{K}|$), then the $|\mathcal{K}|$ users receiving the ad are drawn randomly from $\mathcal{N}_\theta$.

Size of the synthetic population: In our simulation study, we use a synthetic population of size $N = 10^6$ users. We argue, however, that this size is irrelevant and does not affect our results for a fixed N_m so long as the budget is scaled with the population size. Indeed, our estimator only looks at the set $\mathcal{N}_m$ to do the inference. Hence, the size of the rest of the population is irrelevant as long as the probability for a user in $\mathcal{N}_m$ to receive the ad is constant. For this to happen, we just need that the budget K scales with N (if K was fixed in absolute value, the probability would decrease as N grows), or equivalently with N_θ (since N_θ is proportional to N). Since we express K in percentage of N_θ, we this is easily guaranteed as long as this percentage remains the same.

From the above argument, one can observe that it is in fact not necessary to simulate the ad delivery for the whole population $\mathcal{N}$; it is enough to simulate it for the set of monitored users. We still need a larger simulated population to be able to randomly draw multiple realizations of the monitored set (potentially under some constraints of $N_{m,\theta}$ in the consistency experiments, next subsection). However, there is no need to try and simulate a population of the size of the real Facebook population, a much smaller size is enough for that purpose (and $N = 10^6$ is more than enough), because our population generation method generates i.i.d. users.

B. Consistency Real-World vs. Simulated Experiments

Before running our simulation study, it is crucial to verify that our simulation framework (in particular the two key pillars

above) leads to results consistent with the real-world observations. We do this by creating 45 simulated ad experiments (using our simulation framework) that attempt to reproduce the precise conditions of the 45 real-world ad experiments that reached at least 3 users (Sec. III-B). Recall that we use a synthetic population $\mathcal{N}$ based on the dataset $\mathcal{D}_p$, which contains the same attributes as in the real-world experiments.

1) Simulations: For each of the real-world experiment in Table II, we create a simulation that preserves the same θ, N_m, $N_{m,\theta}$ and $N_{m,r}$ by doing the following:

a. We set θ to the targeting formula of the real-world experiment and get the corresponding set of users $\mathcal{N}_\theta$.
b. We construct the set $\mathcal{N}_{m,\theta}$ by randomly picking $N_{m,\theta}$ users from $\mathcal{N}_\theta$.
c. We construct the set $\mathcal{N}_{m,r}$ by randomly picking $N_{m,r}$ users from the previously constructed $\mathcal{N}_{m,\theta}$ set.
d. To complete $\mathcal{N}_m$, we randomly choose $N_m - N_{m,\theta}$ users out of the $\mathcal{N} \setminus \mathcal{N}_{m,\theta}$ set.

Note that step c. of this process is equivalent to the Bernoulli model described in the previous subsection (key pillar (ii)), when conditioning on the value of $N_{m,r}$ (which is crucial for a meaningful comparison to the real-world experiments).

We perform ten independent runs of the above process for each ad experiment. For each run, we compute the estimated $\hat{\theta}$ using the *B-ML* estimator (with $\hat{K}$=25% of N_θ as for the real-world experiments). This setup allows us to have in each run different users in $\mathcal{N}_m$ and $\mathcal{N}_{m,\theta}$, which leads to potentially different $\hat{\theta}$. We compute the accuracy over the ten iterations.

2) Results: Table II presents the accuracy of the *B-ML* estimator in the simulation framework corresponding to each real-world ad experiment (column Sim. Acc.). For each of the 45 ad experiments, we consider that our simulation result is consistent with the real-world result if the real-world inference is correct (resp., incorrect) and the simulation accuracy is over 50% (resp., under 50%). With this definition, 72.4% of the custom-audience ad experiments and 100% of the location-based ad experiments are consistent (82.2% over all experiments), which is remarkable. In particular, we observe a qualitative consistency in the results; for instance, real-world experiments where few users receive the ad give poor accuracy and the same happens with simulations. If we restrict to the 26 (custom-audience) experiments for which $N_{m,r} \geq 10$, however, the average accuracy of simulation is 86.1%, while the real-world accuracy is only 65%. In the next subsection, we discuss in details the reasons for this discrepancy.

3) Discrepancy between real-world experiments and simulations: Recall that we restrict for this discussion to the 26 (custom-audience) experiments for which $N_{m,r} \geq 10$. First, the observed discrepancy in accuracy may be due to randomness. To test that, we run a one-sided statistical test where the null hypothesis is that the 26 experiments results are independent Bernoulli success/failure variables with parameter 0.861 (the alternative hypothesis is that the parameter is less than 0.861). The p-value is 0.0064, hence we reject this null hypothesis with a standard threshold of 5% (even 1%). This means that not all the discrepancy can be attributed to randomness.

To understand why real-world experiments give lower accuracy, it is useful to look at the quantity A_r (number of attributes common to all users who received the ad). Intuitively, the larger A_r, the more difficult to disambiguate the targeting attributes for inference. Table III (in appendix) shows the values of A_r for the real-world experiments and corresponding simulations (averaged over ten runs). We see that A_r is much smaller in the simulations, which explains that the accuracy is higher. There are two main factors to explain this:

(i) *The bias in the set of monitored users.* The set of monitored users $\mathcal{N}_m$ can be biased, that is that its distribution of attributes is not representative of the global population. There can be multiple reasons for that. The recruitment strategy may lead to recruiting biased users, e.g., more active users with more attributes. A bias can also originate from the measurement methodology (and potential measurement errors). For instance, as the measured attributes profiles change over time, we take the union of all snapshots; this could lead to users having more attributes. Table III indicates that, indeed, our set of monitored users is biased towards having more attributes. This can be seen from the fact that the median number of attributes in $\mathcal{N}_m$ is higher in the real-world experiments than in the simulations based on the synthetic population $\mathcal{N}$ generated from $\mathcal{D}_p$ (obtained from the global Facebook population estimates).

To test (and quantify) the effect of the bias in the set of monitored users on accuracy, we proceed as follows: for each of the 26 experiments, we generate a new synthetic population, still using the method of Sec. IV, but using as an input the empirical estimate of the second-order attribute probabilities observed in the set of monitored users active for that particular experiment (instead of the global Facebook estimates as in the population $\mathcal{N}$). This process attempts to mimic in the simulation the bias in the set of monitored users as closely as possible for each experiment. Then, we redo the 10 runs exactly as before. We refer to these simulations as the *biased simulations*; the results are displayed in Table III (right columns). We observe that, indeed, the median number of attributes per users in the monitored set (134.9 on average) is now close to that of the real-world simulations (136.7 on average), consistently with our results of Sec. IV. The values of A_r also became closer. As a result the average accuracy over the 26 experiment is now 73.8%. If we make the same one-sided statistical test as above with parameter 0.738, the p-value is now 0.22, hence we cannot reject anymore the null hypothesis that the discrepancy is purely due to randomness. Nevertheless, the accuracy is still higher than 65% (and the values of A_r still smaller than in the real-world experiments), which is possibly also explained by the second reason below.

(ii) *The ad delivery optimization mechanisms.* Our simulations use a Bernoulli model of ad delivery: every user satisfying the targeting formula has an equal probability of receiving the ad.[2] In practice, the probability might vary due to ad delivery

[2]We are only referring to the Bernoulli assumption used in the simulation here. We also use it in the inference, but that is equally done for simulations and real-world experiments so it cannot create a discrepancy between the two.

optimizations performed by platforms and might be affected by bidding strategies of other competing advertisers. For example, previous work showed that ad delivery was skewed towards men when the image of an ad was showing a gym [33]. More simply, ad delivery may be skewed towards users that are more active and therefore have more attributes. To verify that, Table III displays the median number of attributes per user in both $\mathcal{N}_m$ and $\mathcal{N}_{m,r}$ (columns $\bar{A}_m$ and $\bar{A}_{m,r}$). We observe that while this median is essentially the same between the real-world experiments and the biased simulations for $\mathcal{N}_m$, it is larger in the real-world experiments for $\mathcal{N}_{m,r}$.[3] This means that, in the real-world experiments, users receiving the ads tend to have more attributes than in our (biased) simulations.

This effect typically makes inference harder because it may lead to the set of monitored receiving the ad sharing more common attributes. This can be seen from Table III, by observing that the values of A_r in the real-world experiments are higher than those for the biased simulations that account for the bias in the monitored set. This effect may explain the remaining 9% discrepancy in accuracy (from 73.8% to 65%). Recall from above, however, that it is no longer statistically significant after taking into account the bias in the set of monitored users. This indicates that skews coming from ad optimizations and auctions only have a limited impact on the inference (at least in the setting of our real-world experiments). Note that skews towards a given attribute has an effect on the inference only if the attribute is shared by all (and not just some) of the users receiving the ad—which is not likely as long as sufficiently many monitored users receive a given ad.

4) Take-aways on consistency between simulations and real-world experiments: Our results show that the accuracy we obtain in our simulation framework is globally consistent with the accuracy in the real-world experiments, despite our simplifying assumptions. This gives us confidence that the accuracy computed in the simulation framework is realistic and therefore that the key high-level take-aways of our simulation study are transferable to the real-world—although we do not claim that the numbers from our simulations are exact predictions. The discrepancy between simulations and real-world experiments is well explained: the primary factor is the bias of the monitored users. It is important to take it into account when evaluating whether collaborative transparency can work well since it may decrease the accuracy. It is, however, possible to observe it and to compute its effect as we did above for any given set of monitored users (e.g., for ours, it decreased the accuracy by around 12%). A secondary factor is ad delivery optimizations. This is not under our control and may decrease the accuracy. Our results above, however, suggest that the effect is limited, so these optimizations are not a fundamental limitation of collaborative ad transparency, even though they should be kept in mind.

[3]The difference between the median number of attributes in $\mathcal{N}_m$ and $\mathcal{N}_{m,r}$ in the biased simulations is explained by the conditioning on the targeting formulas. We verify that by checking that the median number of attributes is the same in $\mathcal{N}_{m,r}$ and $\mathcal{N}_{m,\theta}$ (columns $\bar{A}_{m,r}$ and $\bar{A}_{m,\theta}$ in Table III).

C. Setup of Simulated Ad Experiments

In the rest of the paper, we perform our simulation study according to the two key pillars described in Section V-A and validated in Section V-B. We detail here the precise simulation protocol, summarized in Algorithm 2. We use a synthetic population of $N = 10^6$ users created from dataset $\mathcal{D}_p$ (Sec. III-A) over $A = 322$ attributes. Recall that, as discussed above, the actual size of the synthetic population is irrelevant, only values of N_m and K will matter. From this population, we precompute N_θ for all $\theta \in \mathcal{T}$ (i.e., all combination of two attributes); this will be used for inference.

In each ad experiment, we can set *three main parameters*: (1) the targeting formula θ (which determines the number N_θ of users who satisfy it); (2) a number K linked to the ad campaign budget (expressed in percentage of N_θ); (3) the number of monitored users N_m. Then we perform $n_{\text{runs}} = 10$ runs as follows: *First*, we uniformly randomly sample a set of monitored users $\mathcal{N}_m$ with predefined size N_m. *Second*, we randomly draw, for users in $\mathcal{N}_{m,\theta}$, i.i.d. Bernoulli random variables with probability K. That defines the labels $\hat{y}^i$ corresponding to which users are shown the ad in $\mathcal{N}_m$ (i.e., this defines $\mathcal{N}_{m,r}$)—recall that users in $\mathcal{N}_m \setminus \mathcal{N}_{m,\theta}$ have $\hat{y}^i = 0$. If no user has $\hat{y}^i = 1$, i.e., $\mathcal{N}_{m,r}$ is empty, then we resample the experiment. This is consistent with the fact that we want in practice to infer parameters of the ad campaigns only for ads that at least one monitored user receives. Note that we draw labels only for users in $\mathcal{N}_m$. Whether or not non-monitored users are shown the ad is irrelevant for our inference as discussed earlier, and this approach is computationally faster. Note also that by drawing labels as Bernoulli random variables according to our second key pillar, we no longer fix the value of $N_{m,r}$, which allows exploring its impact on accuracy.

Finally, we compute the estimate of the targeting formula $\hat{\theta}$ using the *B-ML* estimator as defined in (2) (as well as the *B-EM* estimator) for each run. As each run has a different (randomly sampled) $\mathcal{N}_m$ and $\mathcal{N}_{m,r}$, this gives ten different estimates for a given set of parameters (θ, K, N_m). Note that computing the *B-ML* estimator requires a value of K. As the true value is often unknown at inference time, we use a value $\hat{K}$ that is passed as an input of Algorithm 2.

VI. Simulation Study: Results

In the real-world ad experiments, we tried limited values of θ (chosen to increase the likelihood that our monitored users receive our ads), K (we were only able to use small ad budgets), and N_m (we were constrained by the number of users who installed our monitoring tool). This left unanswered the question of *how the accuracy would change for other θ's, other ad budgets, or other numbers of monitored users*. In this section, we exploit the simulation framework to look at the variations of accuracy for a wide range of θ, K, and N_m in order to answer this question and gain a broader view on the potential and limits of collaborative ad transparency.

input : Population parameters: $\mathcal{N}$, $\mathcal{A}$, N_θ for all $\theta \in \mathcal{T}$
Ad campaign parameters: θ, K
Collaborative transparency parameters: N_m
Inference parameters: $\hat{K}$
Simulation parameters: number of runs n_{runs}
output: Collection of n_{runs} inferred targeting formulas $\hat{\theta}$

while *Number of estimates* $\hat{\theta}$ *obtained* $< n_{runs}$ **do**
  Uniformly sample a set $\mathcal{N}_m \subset \mathcal{N}$ of N_m monitored users;
  Compute the corresponding set $\mathcal{N}_{m,\theta}$;
  if $\mathcal{N}_{m,\theta} \neq \emptyset$ **then**
    repeat
      For every $i \in \mathcal{N}_{m,\theta}$, draw an iid Bernoulli random variable $\hat{y}^i$ with parameter K;
      $\mathcal{N}_{m,r} \leftarrow \{i \in \mathcal{N}_m : \hat{y}^i = 1\}$;
    until $\mathcal{N}_{m,r} \neq \emptyset$;
    Calculate $\mathcal{A}_r$ from $\mathcal{N}_{m,r}$ and $\mathcal{T}_r$ from $\mathcal{A}_r$;
    Compute $N_{m,\theta}$ and then $\mathcal{L}(\mathcal{T}_r)$ for all $\theta \in \mathcal{T}_r$;
    Compute maximum likelihood estimate $\hat{\theta}$ from (2) with $\hat{K}$;
  end
end

Algorithm 2: Experiments description.

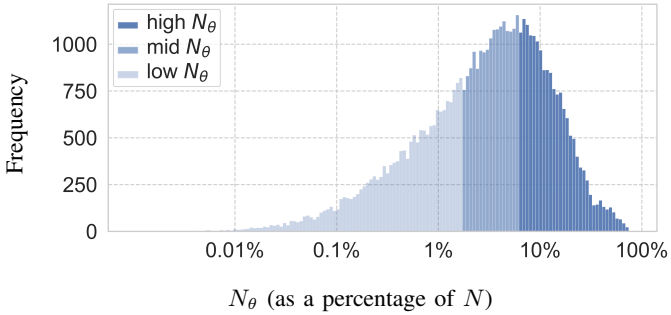


Fig. 2: Histogram of N_θ in the population $\mathcal{N}$.

A. Simulation Dataset

Each simulation with a fixed tuple of parameters $(\theta, K, N_m, \hat{K})$ is done according to Algorithm 2, with 10 runs. We vary these parameters as follows:

θ: In total, there are $51,681$ possible values of θ in $\mathcal{T}$ (i.e., combinations of two attributes from the 322 we consider). Figure 2 shows the histogram of N_θ (the x-axis is in log-scale). We see a wide range of N_θ going from 10 to $801,438$. In the analysis, we split $\mathcal{T}$ (the set of all θ) in three categories: low θ (with N_θ in the bottom 33.3% percentile, $N_\theta < 1.7\%N$); medium θ (with N_θ between the 33.3% and 66.6% percentile, $N_\theta \in [1.7\%N, 6.2\%N)$), and high θ (with N_θ in the highest 33% percentile, $N_\theta \geq 6.2\%N$). To keep simulations computationally feasible, we pick $1,000$ random θ to produce simulations results in this section.

K: Recall that we express K as a fraction of N_θ (instead of the absolute value of reach estimate). That is, we investigate the expected accuracy for ad campaigns where the advertiser sets a budget to reach a given fraction (say 0.2, or 20%) of the users that satisfy a particular targeting formula. This way, our results are easily transferable to real-world settings with various populations sizes N (e.g., different countries). We consider a wide range of ad campaign budgets by varying K in $\{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

N_m: Finally, we consider different numbers of monitored users N_m: 150, 200, 250, 300, 700, 1000, 2000, 3000.

$\hat{K}$ and estimates: As we perform 10 runs for each unique combination of (θ, K, N_m), we have a total of $1,000 \cdot 11 \cdot 8 \cdot 10 = 880,000$ runs. In each run, Algorithm 2 outputs the estimated targeting formula $\hat{\theta}$ using the *B-ML* estimator and the *B-EM* estimator. As K is usually unknown in the wild (we do not know budget used by the advertiser in the ad campaign), the *B-ML* estimator uses a value $\hat{K}$ as an input. We compute four estimates per run with: $\hat{K} = K$ (we assume we know the real K), $\hat{K} = 0.1N_\theta$ (we assume K is small), $\hat{K} = 0.25N_\theta$ (the value we used for real-world experiments), $\hat{K} = 0.9N_\theta$ (we assume K is high). Note that the *B-EM* estimator does not need an input $\hat{K}$ (which is what makes it appealing).

Accuracy: For a precise set of parameters (θ, K, N_m, $\hat{K}$), we compute accuracy over the 10 runs as the fraction of times where $\hat{\theta} = \theta$. Wherever relevant, we also aggregate runs with multiple values of a parameter to compute the accuracy.

Note on results reporting: Unless otherwise specified, we report the results for N_m = 700 (Sec. VI-C2 discusses the impact of N_m) and for the *B-ML* estimator with $\hat{K} = K$ (Sec. VI-B1 presents the analysis for other $\hat{K}$ and the *B-EM* estimator).

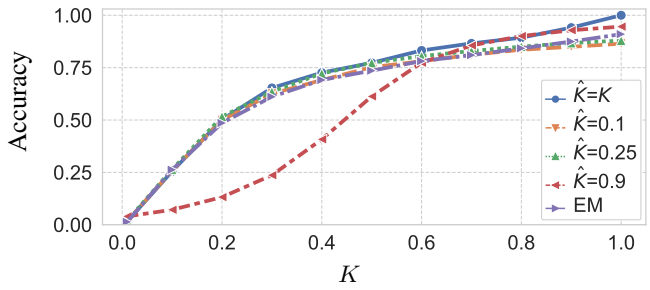
Note on results transferability to the real-world: We recall that we use a synthetic population of size $N = 10^6$, but that as discussed earlier, this does not impact transferability of our results to the real-world for a fixed N_m as long as K remains constant expressed as a fraction of N_θ. We refer the reader to our more detailed discussion in Section V-A on that aspect.

B. Accuracy and Ad Targeting Parameters

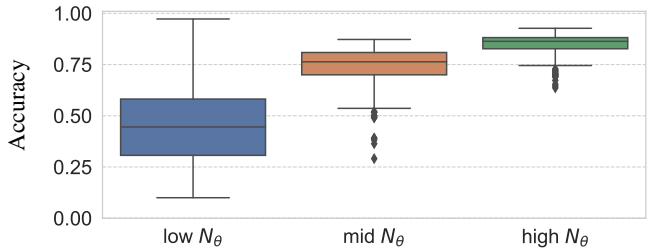
1) Accuracy as a function of K and $\hat{K}$: The ad budget K is an essential parameter in the targeting as it controls the probability that a user of $\mathcal{N}_\theta$ receives the ad and, hence, could impact in a major way the accuracy of the inference. Figure 3a shows how the accuracy varies with K. The plot compares the accuracy for the *B-ML* estimator with $\hat{K} = K$ (the ideal case) with the accuracy for the *B-ML* estimator with other values of $\hat{K}$ and for the *B-EM* estimator. The accuracy is computed by aggregating over the $1,000$ random θ we picked.

The first and most important observation is that the *B-EM* estimator as well as the *B-ML* estimator with $\hat{K} = 0.1$ or $\hat{K} = 0.25$ have an accuracy close to the ideal case $\hat{K} = K$ (although slightly smaller for large K). This is very positive as it means that not knowing the budget used by the advertiser does not pose any major issue for our inference accuracy (as long as we do not consider large values such as $\hat{K} = 0.9$).

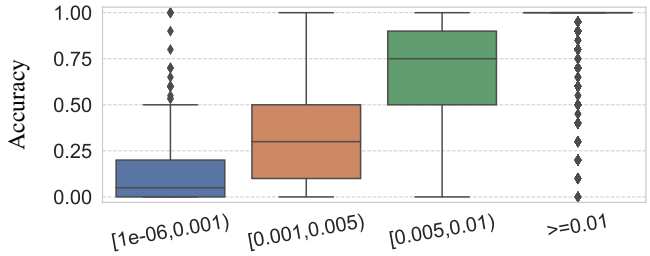
Figure 3a also shows that the accuracy increases with K. That is expected since when K is high, more of the monitored users will receive the ad, which gives more information to compute the inference. We also see that if an advertiser uses an ad budget that is high enough to target 50% of the users satisfying θ, we observe an inference accuracy of 75% on average across θ (since the results are aggregated over $1,000$ random θ they represent an average and not an idealized easy case). This is very encouraging as we only use $N_m = 700$ (recall again that it holds irrespective of the population size).



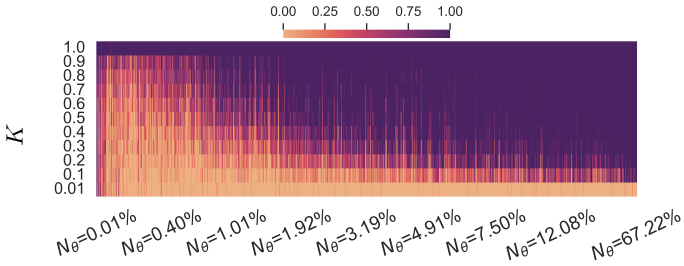
(a) Accuracy for different values of K (aggregated across θ).



(b) Accuracy for different values of N_θ (aggregated across K).



(c) Accuracy vs. estimated reach ($K \times N_\theta$).



(d) Accuracy vs. K and θ.

Fig. 3: Accuracy vs. targeting parameters. Statistics computed for $N_m = 700$, over all 1000 random θ and for $K \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

2) Accuracy as a function of θ: The targeting formula θ is clearly an important parameter as it affects the number of users that can receive the ad N_θ, which in turn impacts the number of monitored users receiving the ad. Figure 3b shows how the accuracy of the inference varies with N_θ. The accuracy is aggregated across the different values of K to provide a fairer and broader picture. As expected, the accuracy generally increases as N_θ increases. More interestingly, the plot shows an accuracy of over 75% for θ with N_θ as small as $1.7\%N$, on average over campaign budgets.

3) Accuracy as a function of both K and θ: To investigate the effect of both K and θ on the accuracy, Figure 3c plots the accuracy as a function of the estimated reach, i.e., the total number $K \times N_\theta$ of users an ad campaign reaches. The plot shows an average accuracy of over 75%, as long as the estimated reach is over 0.005 (i.e., 0.5%) of N. To give an idea, let us assume that we use collaborative ad transparency in France, hence we have an entire Facebook population of 38 million users. Let us also consider that the average cost for 1000 impressions is 7€ [25]. An ad that reaches 0.005 of the population (i.e., 190,000 users) will cost $0.005 \cdot 38,000,000 \cdot 7/1,000 = 1,330$€. Hence, we can see that collaborative ad transparency can achieve good accuracy for campaigns with relatively small budgets while only monitoring 700 users. If we consider countries with larger populations, collaborative ad transparency will also achieve a 75% accuracy if the ad reaches 0.005 of the population, but this will translate into more expensive ad campaigns. The next section discusses how the number of monitored users impacts the ad campaigns for which we can expect good accuracy.

To disentangle better the effect of θ and K on the inference accuracy, Figure 3d presents a heatmap of the accuracy for all the combinations of (K, θ) we consider. On the x-axis, we ordered the 1000 θ according to their N_θ, and on the y-axis, we present K. For each combination of θ and K, we represent the accuracy over the ten runs as colored rectangle (where darker shades represent higher accuracy). The heatmap provides a complete picture of the cases where we can achieve high accuracy (and those where we cannot). For example, we can achieve high accuracy for formulas with high N_θ ($\geq 6.2\%N$) even when K is as small as 0.2, but we need a K of over 0.9 for formulas with small N_θ ($< 1.7\%N$). This constraint bodes well with what we expect to happen in real-life: advertisers either set a broad nest of users they want to reach (N_θ high), but only target a fraction of them as their overall ad budget is limited, or focus their whole budget on a small set of users with precise characteristics they think will lead to high conversion rate (N_θ low), and target most of them. The ad campaigns for which the inference does not work well are the ones that reach only a small fraction of a small set of users with precise characteristics. We expect such ad campaigns to be small test drives and not generally represent advertisers' behavior in real life. Finally, the heatmap of Figure 3d does not appear as a smooth color gradient, that is, we observe large color variations for points next to each other. This means that, for a fixed K, ads with very similar N_θ but with different θ may lead to different accuracies. This may be due to the fact that two formulas with the same N_θ can have attributes with different marginal probabilities, although the precise relationship to accuracy is more intricate. We observe, however, that this second-order effect vanishes in regions of very high (or very low) accuracy—hence our results based only on N_θ in these regions are robust.

4) Takeaways: There are two key results in this section. First, the *B-ML* estimator and *B-EM* estimator remain almost equally accurate even if we do not know the real K. Second, we observe an accuracy of over 75% on average over a wide range of targeting formulas θ and ad budgets K as long as

the ad campaign reaches 0.005 of the entire population while only monitoring 700 users in the location we study.

C. Accuracy and Collaborative Ad transparency Parameters

The parameter directly controlled in collaborative ad transparency is the number of monitored users N_m. Naturally we expect that a larger number of monitored users will lead to a higher accuracy *on average* as we can observe more information. However, we saw in the real-world experiments (Sec. III-B) that for a similar number of monitored users, the inference accuracy varies significantly for different ad campaigns. Beyond the variation in the ad parameters, a key parameter seemed to be the number of users who received the ad in the monitored set $N_{m,r}$ (recall that when less than ten monitored users received our ad, we never inferred the targeting formula correctly). Therefore, in this section, we investigate separately the effect of $N_{m,r}$ and then of N_m.

1) Accuracy as a function of $N_{m,r}$: $N_{m,r}$ intuitively captures the "amount of information" available to make an inference. Figure 4 shows the inference accuracy as a function of $N_{m,r}$. In this plot, we aggregate all simulations for the $1,000$ random θ, $K \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, but we fix $N_m = 700$. To compute the accuracy for a particular $N_{m,r}$, we take all the simulations that reached that $N_{m,r}$, irrespective of θ or K, and compute the fraction for which inferred correctly $\hat{\theta} = \theta$. For a more detailed view, we compute accuracy across all 1000 θ together, as well as for low, mid, and high N_θ separately. Figure 4 shows that the accuracy is close to zero if only a few monitored users received the ad, but that it rapidly increases with $N_{m,r}$ until $N_{m,r}$ is around eight, after which the increase is much slower. More importantly, the plot shows that *we get an accuracy of over 90% on average for ads that reach ten or more monitored users*. This is one of the central results of this paper. It is a promising result not only because ten is a relatively small number, but also because it provides predictability and confidence in the inference: since we can directly measure $N_{m,r}$, we know when we can have confidence in our inference and when we cannot.

Figure 4 also shows that there are minor differences in accuracy for different values of N_θ: the higher N_θ, the lower the average accuracy. This is normal since we condition on $N_{m,r}$: for a fixed value of $N_{m,r}$, if N_θ gets larger it means that we are in a particularly skewed realization of the random process where particularly few monitored users received the ad compared to what was expected. Nevertheless, for $N_{m,r} \geq 10$, all values of N_θ give high accuracy—in fact, out of all simulations such that $N_{m,r} \geq 10$, 97% have correct inference. Hence, $N_{m,r} \geq 10$ is a remarkably robust signal of high accuracy. At this point, it is useful to recall once again that our results are transferable to the real world irrespective of its population size—hence this cutoff number of $N_{m,r} \geq 10$ to obtain reliably good accuracy would remain the same.

A key reason why $N_{m,r}$ has such an important effect on accuracy is that it directly impacts the number of attributes that are common between the users that received the ad A_r,

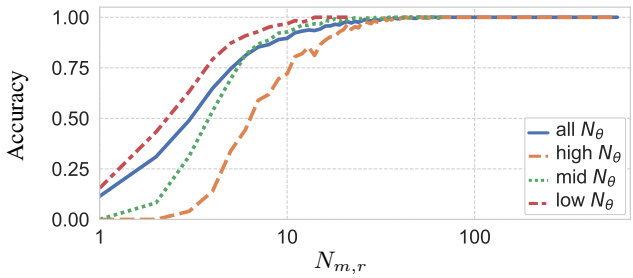


Fig. 4: Accuracy for different values of $N_{m,r}$ (for all $1,000$ θ as well as for θ with low, mid, and high N_θ). Statistics computed for $N_m = 700$, and aggregated over $K \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

which in turn has an impact on the size of the set of targeting formulas T_r that the maximum likelihood estimator considers. We show this in more detail in App. C-A.

2) Accuracy as a function of N_m: To explore the effect of N_m on accuracy we use the simulations with various $N_m \in \{150, 200, 250, 300, 700, 1000, 2000, 3000\}$, still with $1,000$ random θ, and $K \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Figure 5 displays the accuracy (computed by aggregating across all θ and K) as a function of N_m. The plot shows that the accuracy rapidly increases with N_m until N_m is around 1000, then the growth is much slower. Note that the accuracy does not reach 100% because we aggregate it over multiple K, including $K = 0.01$, which always has bad accuracy. These results suggest that we can expect collaborative ad transparency to provide accurate inferences if we can monitor 1000 users or more. To get a deeper understanding, we compute accuracy separately for simulations for which less than ten users received the ad and simulations for which ten or more users received the ad. We see that if $N_{m,r} \geq 10$, the accuracy is over 90% for any number of monitored users (e.g., as low as 300), while if $N_{m,r} < 10$, the accuracy is poor even if N_m as high as 3000. These results confirm our previous findings that the most important parameter to determine the expected accuracy for a particular ad is $N_{m,r}$.

We also plot the accuracy as a function of N_m conditioned on N_θ and K (the figures, Figures 8a and 8b, can be found in App. C-B). The plots show that higher values of N_m are beneficial in terms of accuracy for small N_θ or K. The reason for this is that while $N_{m,r}$ determines whether we can make an accurate inference or not for a particular ad, N_m plays an essential role in increasing the number of ads for which we can make an accurate inference because it increases the number of ads for which $N_{m,r}$ is ten or more. To visualize the impact of N_m on $N_{m,r}$, Figure 6 presents a contour plot that shows for which combinations of (K, N_θ) we can expect to have a certain value of $N_{m,r}$ $(10, 20, \cdots)$, when $N_m = 300$ and when $N_m = 3000$. We observe that the benefit of large values of N_m is that it gives a much bigger range of values of N_θ and/or K (in particular small values) for which $N_{m,r}$ is ten or more, hence we can have an accurate inference. (Note that the contours in this plot have a shape similar to the heatmap of Figure 3d, which is normal since there is almost an equivalence

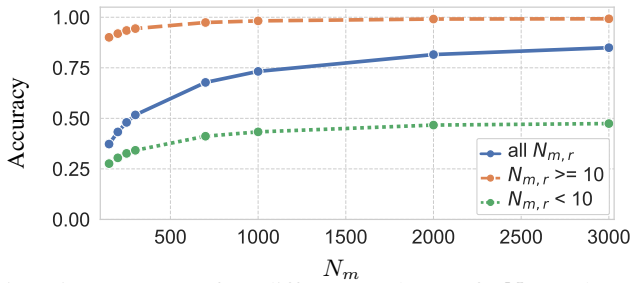


Fig. 5: Accuracy for different values of N_m. The accuracy is aggregated over $1,000$ random θ, and $K \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

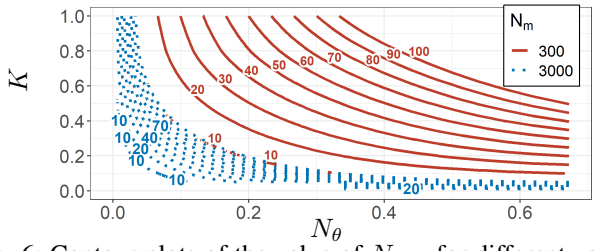


Fig. 6: Contour plots of the value of $N_{m,r}$ for different values of K and N_θ, for $N_m = 3000$ and $N_m = 300$. For instance, the red line with 20 corresponds to all the couples (K, N_θ) for which $N_{m,r} = 20$ in expectation.

between high accuracy and $N_{m,r} \geq 10$).

3) Takeaways: The results show that when an ad is received by ten or more monitored users, we observe an inference accuracy of over 90% irrespective of the number of monitored users N_m. N_m plays an indirect role in the accuracy by increasing or decreasing the likelihood for a particular ad campaign that enough monitored users will receive the ad. Increasing N_m plays an essential role in getting high accuracy for ad campaigns that have either a θ with a small N_θ or a small budget. Choosing the right number of monitored users depends, of course, on whether we want to use collaborative ad transparency in a best-effort way or we want to be comprehensive. One thousand monitored users seems to be a good enough number to achieve good accuracy for most ad campaigns, but we need $1,000$ *active* monitored users—which probably requires recruiting about twice as many.

VII. Related Work

Limits of transparency mechanisms provided by platforms. To address the demand for more transparency, advertising platforms have developed three main transparency mechanisms. *First*, platforms started to provide ad explanations to users about why they received a specific ad [34]–[36]. However, in Facebook's case, Andreou et al. [8], showed in 2018 that these explanations are incomplete and only show one (out of many) micro-targeting parameters used by the advertiser to target an ad. *Secondly*, ad platforms started to offer online Ad Libraries [37]–[39] that can be used by the public to investigate—mostly—political ads. Several reports pointed out limitations and vulnerabilities with such Ad Libraries [17], [18], and many criticized them for not showing information about the precise targeting parameters used by the advertiser. *Finally*, ad platforms provide users with Ad Preference Managers [40]–[42] that show what attributes the platform has inferred about them. Several works showed they might not include all information that can be used to target users both in the case of Google [43], [44] and Facebook [8], [45]. In addition, several studies have investigated the inferred attributes and found that they deal with sensitive topics [22], and their accuracy in reflecting user's interest is questionable [46]–[48]. Irrespective of the quality of the inference process [8], [49], [50], these are the attributes used by the advertiser to reach users. While ad platforms have since revamped Ad Preference Managers and ad explanations and now provide more comprehensive information, the vulnerabilities mentioned above show the importance of external auditing systems and the need for third-party efforts to bring more transparency.

Third-party efforts to infer ad targeting. To address this need, several early studies developed methodologies to infer whether an ad is contextual, re-targeted, or behavioral [43], [51]–[54] and see how the activities of a user influence the ads s/he receives [55]–[57]. Our scope is complementary and proposes to infer the actual targeting parameters used by the advertiser. These works are based on creating fake personas (e.g., creating a browsing profile with a clean slate browser) and monitoring the ads that these personas receive. While these techniques can pinpoint problems with the ad ecosystem, they rely on methods that cannot be applied at scale to infer why users receive some particular ads because they require the creation of multiple fake accounts, which is very costly in time and resources. One method that does not use fake personas is by [57], who developed a model using hypothesis testing to check if an ad is interest-based by monitoring the user's browsing behavior. Closest to our work is the study by Iordanou et al. [58] that used ads donated by users to detect targeted ads by counting how often an ad appears across different users while keeping the ads and browsing history of users private. This study focuses on designing a privacy-preserving protocol that only infers if an ad was targeted and not the specific targeting attributes used by the advertiser.

Measurements of ad targeting. Several studies have looked at how advertisers use the system to target users and to which extent they use micro-targeting [24], [59]. Such studies shed light on users' targeting but rely on the platform's transparency mechanisms for the data they utilize. Finally, in a different direction, Ali et al. [33] showed how Facebook's ad delivery process affects who receives an ad and observed that the delivery of ads can be skewed across gender or racial lines, even when the advertisers did not intend it. A follow-up work focused on political ads [10] also observed that the price of reaching a user differs according to their political alliance.

VIII. Discussion and Limitations

In this paper, we investigate the accuracy of collaborative ad transparency in inferring the targeting formula specified by an advertiser, using a combination of real-world experiments and simulations based on a simple abstraction of the ad

delivery process. Our study aims at laying the foundations of collaborative ad transparency by understanding in which conditions it can or cannot work. The precise results of our real-world experiments are likely biased by the specific population from which our monitored users are drawn (which also explains the discrepancy with our simulation results). More generally though, a key insight of our study is that bias in the set of monitored users may negatively affect accuracy—this should be kept in mind for potential deployment—but it is possible to quantify this effect with our simulation framework.

Limitations related to the targeting formulas we considered: Our study focuses on targeting formulas that consists of a combination of two attributes. This was intended to be rich enough compared to a single attribute but to remain simple to set clear foundations. Our framework, however, is generic and can handle more complex formulas. As we have seen, a key parameter to ensure the inference quality is T_r, the number of targeting formulas compatible with the set of monitored users who received the ad. As T_r is constrained by the intersection of attributes in this set, we expect that it would not increase too much when considering combinations of more attributes, hence that the accuracy would not drop drastically.

We also only considered curated attributes, which are only around 300. In principle, considering free-text attributes (with thousands of them) would be possible. Even though simulations might be too costly (i.e., the generation of a synthetic population), performing the real-world inference would presumably be possible since we only consider the attributes shared by all users that received an ad. However, the real challenge for inferring targeting formulas with free-text attributes is to have enough monitored users that received the ad. As we discussed, inferring targeting formulas with a small reach requires more users to monitor. Our framework can be used to calculate the number of monitored users needed for a desired accuracy and for a particular targeting formula.

Finally, ad platforms optimize ad delivery based on (potentially thousands) of internal signals and advanced data mining algorithms. While auditing how ad platforms deliver ads is important, the focus of this study is to infer the targeting formula specified by the advertiser. Having this information is an essential step in detecting ill-intentioned advertisers.

Assumptions on the ad delivery: Our main simplifying assumption resides in the Bernoulli model of ad delivery—we discussed it in Section V-B. Another key limitation of our model is that it assumes that users who do not satisfy the targeting formula cannot receive the ad. If this assumption is not satisfied, our inference will be wrong. Nevertheless, it is unlikely that ad platforms would deliberately deliver ads to users that do not satisfy the criteria asked by paying advertisers. In the real-world experiments, several of our ads were received by users that did not have all the targeted attributes in their user vectors. As discussed, we believe this is a data collection problem rather than an ad delivery problem. Hence, our approach requires frequent monitoring of users' attributes to limit such cases. Alternatively, it would be possible to extend our model to allow a small probability of receiving the ad when not satisfying the targeting formula.

Feasibility and access to the necessary data: Collaborative ad transparency requires gathering, for a set of monitored users, the ads they see and their attributes. While Facebook has attempted on numerous occasions to disrupt the data collection of similar tools [60], [61], the platforms are bound by law to mark posts paid by advertisers [62]. Hence, it will always be theoretically possible (even if technically challenging) to collect this information. Our approach requires frequent monitoring of users' attributes. This information can be collected from the Ad Preference pages of users (other platforms such as Google and Twitter have similar pages) but may vary over time (in our study the first snapshot had an average of 78 attributes per user and during the 20 weeks we observed an average of 61 new attributes per user added and 64 attributes deleted). Finally, the *B-ML* estimator requires knowledge on the number of users satisfying all the considered targeting formulas. Currently, this information is available in the Ads Manager. Nevertheless, even if Facebook stopped making such information available, we can use the *B-EM* estimator which has similar accuracy and does not require this information.

One limitation of collaborative ad transparency is that we rely on data provided by the ad platform on the user attributes and reach estimates. If this information is wrong, the inferences will also be wrong. If the Ad Preference page is wrong but the ad delivery process uses the same wrong information, then our method would correctly infer the targeting formula set by the advertiser. However, if the Ad Preference page shows purposely incorrect information different from that used to deliver the ads, then our method would make an incorrect inference. It may be possible to ask users to describe their interests (e.g., via a survey) to make an inference based on those—this would not recover the advertiser's formula but rather the main characteristics of users receiving an ad.

Finally, advertisers can use dynamic content in ads, that makes ads appear different to different recipients. Linking ads that are part of the same ad campaign but have different creatives is not trivial but could be doable depending on the information provided by ad platforms. For example, the ad explanation ("Why am I seeing this") provided by Facebook has an id. From a few anecdotal experiments that we did, we observed that if two ads are part of the same campaign, they lead to ad explanations with the same id, hence allowing the linkage. A more comprehensive study would, however, be required to confirm that and to assess the prevalence of dynamic contents. Nevertheless, even if this information is not available, we can still infer the targeting formulas by groups of identical ads. Of course, it reduces the available data, making it less likely that ten or more monitored users receive the ad.

ACKNOWLEDGMENT

Part of this work was done when the authors were with Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG. It was supported by ANR (ANR-17-CE23-0014, ANR-21-CE23-0031-02, ANR-20-CE23-0007); by MIAI@Grenoble Alpes (ANR-19-P3IA-0003); and by the EU (101021377 and 952215).

REFERENCES

[1] "Facebook ads manager," https://www.facebook.com/business/ads.

[2] K. Steinmetz, "Lawmakers Hint at Regulating Social Media During Hearing With Facebook and Twitter Execs," https://time.com/5387560/senate-intelligence-hearing-facebook-twitter/, 2018.

[3] M. de La Baume, M. Scott, and L. Kayali, "Facebook to cave to EU pressure after row over political ad rules," https://www.politico.eu/article/facebook-european-elections-advertising-political-social-media-europe/, 2019.

[4] M. Wall, "Facebook, Google and Twitter in data regulators' sights," https://www.bbc.com/news/business-48357772, 2019.

[5] "Facebook's UK political ad rules kick in," https://www.bbc.com/news/technology-46385050, 2018.

[6] K. Bell, "Facebook pushes new rules for political advertising worldwide," https://mashable.com/article/facebook-political-advertising-rules-worldwide/, 2019.

[7] D. Ingram, "Foreign governments are fed up with social media–and threatening prison for tech employees," https://nbcnews.to/3seP8RL, 2019.

[8] A. Andreou, G. Venkatadri, O. Goga, K. Gummadi, P. Loiseau, and A. Mislove, "Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations," in *NDSS*, 2018.

[9] "Facebook and Google: This is What an Effective Ad Archive API Looks Like," https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/, 2019.

[10] M. Ali, P. Sapiezynski, A. Korolova, A. Mislove, and A. Rieke, "Ad delivery algorithms: The hidden arbiters of political messaging," in *WSDM*, 2021.

[11] "Leveling the platform: Real transparency for paid messages on facebook," https://www.teamupturn.org/reports/2018/facebook-ads/.

[12] C. Wardle, "Enlisting the public to build a healthier web information commons," in *WWW*, 2019.

[13] "Vestager moves closer to the data heart of digital giants," https://politi.co/3iKNhRg.

[14] "AdAnalyst: A tool to help you make sense of the ads you receive on Facebook," https://adanalyst.mpi-sws.org/.

[15] "Ad Observer," https://adobserver.org/.

[16] "Who Targets Me," https://whotargets.me/.

[17] M. Silva, L. Santos de Oliveira, A. Andreou, P. O. Vaz de Melo, O. Goga, and F. Benevenuto, "Facebook ads monitor: An independent auditing system for political ads on facebook," in *TheWebConf*, 2020.

[18] L. Edelson, T. Lauinger, and D. McCoy, "A security analysis of the facebook ad library," in *S&P*, 2020.

[19] J. Merrill, "Does facebook still sell discriminatory ads?" *The Markup*, 2020. [Online]. Available: https://bit.ly/3CMVfBn

[20] J. Merrill and H. Kozlowska, "How facebook fueled a precious-metal scheme targeting older conservatives," *QUARTZ*, 2019. [Online]. Available: https://qz.com/1751030/facebook-ads-lured-seniors-into-giving-savings-to-metals-com/

[21] J. Merrill, "How facebook's ad system lets companies talk out of both sides of their mouths," *The Markup*, 2021. [Online]. Available: https://bit.ly/3AE4XnZ

[22] J. G. Cabañas, Á. Cuevas, and R. Cuevas, "Unveiling and quantifying facebook exploitation of sensitive personal data for advertising purposes," in *USENIX Security*, 2018.

[23] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[24] A. Andreou, M. Silva, F. Benevenuto, O. Goga, P. Loiseau, and A. Mislove, "Measuring the Facebook advertising ecosystem," in *NDSS*, 2019.

[25] "How Much Does Facebook Advertising Cost in 2021?" https://www.webfx.com/social-media/how-much-does-facebook-advertising-cost.html.

[26] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004.

[27] L. Emrich and M. Piedmonte, "A method for generating high-dimensional multivariate binary variates," *The American Statistician*, vol. 45, no. 4, pp. 302–304, 1991.

[28] C. Park, T. Park, and D. Shin, "A simple method for generating correlated binary variates," *The American Statistician*, vol. 50, no. 4, pp. 306–310, 1996.

[29] F. Leisch, A. Weingessel, and K. Hornik, "On the generation of correlated artificial binary data," WU Vienna University of Economics and Business, Working Paper, 1998.

[30] N. Higham, "Computing the nearest correlation matrix—a problem from finance," *IMA journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

[31] F. Leisch and A. Weingessel, "The bindata package," 2006.

[32] D. Bates and M. Maechler, *Matrix: Sparse and Dense Matrix Classes and Methods*, 2019, R package version 1.2-18. [Online]. Available: https://cran.r-project.org/package=Matrix

[33] M. Ali, P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, and A. Rieke, "Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes," *CSCW*, 2019.

[34] "How does facebook decide which ads to show me?" https://www.facebook.com/help/562973647153813.

[35] "Why you're seeing an ad – google account help," https://support.google.com/accounts/answer/1634057?

[36] "Why you're seeing an ad," https://help.twitter.com/en/why-are-you-seeing-this-ad.

[37] "Ad library," https://www.facebook.com/ads/archive/.

[38] "Political advertising on google – google transparency report," https://transparencyreport.google.com/political-ads/library.

[39] "Ad transparency center," https://ads.twitter.com/transparency.

[40] "Ad personalization," https://adssettings.google.com/.

[41] "Ad preferences," https://www.facebook.com/adpreferences.

[42] "Want to see your data?" http://bluekai.com/registry.

[43] C. Wills and C. Tatar, "Understanding what they do with what they know," in *WPES*, 2012.

[44] A. Datta, M. C. Tschantz, and A. Datta, "Automated experiments on ad privacy settings," *PETS*, 2015.

[45] G. Venkatadri, E. Lucherini, P. Sapiezyński, and A. Mislove, "Investigating sources of PII used in Facebook's targeted advertising," in *PETS*, 2019.

[46] A. Galán, J. Cabañas, Á. Cuevas, M. Calderón, and R. Cuevas, "Large-scale analysis of user exposure to online advertising on facebook," *IEEE Access*, vol. 7, pp. 11 959–11 971, 2019.

[47] P. Hitlin and L. Rainie, "Facebook algorithms and personal data," https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/, 2019.

[48] G. Venkatadri, P. Sapiezynski, E. Redmiles, A. Mislove, O. Goga, M. Mazurek, and K. Gummadi, "Auditing offline data brokers via Facebook's advertising platform," in *WWW*, 2019.

[49] M. Degeling and J. Nierhoff, "Tracking and tricking a profiler: Automated measuring and influencing of bluekai's interest profiling," in *WPES*, 2018.

[50] M. Bashir, U. Farooq, M. Shahid, M. Zaffar, and C. Wilson, "Quantity vs. quality: Evaluating user interest profiles using ad preference managers," in *NDSS*, 2019.

[51] B. Liu, A. Sheth, U. Weinsberg, J. Chandrashekar, and R. Govindan, "Adreveal: improving transparency into online targeted advertising," in *HotNets*, 2013.

[52] J. Carrascosa, J. Mikians, R. Cuevas, V. Erramilli, and N. Laoutaris, "I always feel like somebody's watching me: measuring online behavioural advertising," in *CoNEXT*, 2015.

[53] S. Guha, B. Cheng, and P. Francis, "Challenges in measuring online advertising systems," in *IMC*, 2010.

[54] P. Barford, I. Canadi, D. Krushevskaja, Q. Ma, and S. Muthukrishnan, "Adscape: Harvesting and analyzing online display ads," in *WWW*, 2014.

[55] M. Lecuyer, R. Spahn, Y. Spiliopolous, A. Chaintreau, R. Geambasu, and D. Hsu, "Sunlight: Fine-grained targeting detection at scale with statistical confidence," in *CCS*, 2015.

[56] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu, "Xray: Enhancing the web's transparency with differential correlation." in *USENIX Security*, 2014.

[57] J. Parra-Arnau, J. Achara, and C. Castelluccia, "Myadchoices: Bringing transparency and control to online advertising," *ACM Trans. Web*, vol. 11, no. 1, 2017.

[58] C. Iordanou, N. Kourtellis, J. M. Carrascosa, C. Soriente, R. Cuevas, and N. Laoutaris, "Beyond content analysis: Detecting targeted ads via distributed counting," in *CoNEXT*, 2019.

[59] A. Ghosh, G. Venkatadri, and A. Mislove, "Analyzing Facebook Political Advertisers' Targeting," in *ConPro*, 2019.

[60] "Facebook Moves to Block Ad Transparency Tools — Including Ours," https://www.propublica.org/article/facebook-blocks-ad-transparency-tools.
[61] "Facebook bans academics who researched ad transparency and misinformation on Facebook," https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-obser
[62] ".com Disclosures: How to Make Effective Disclosures in Digital Advertising," https://www.ftc.gov/sites/default/files/attachments/press-releases/ftc-staff-revises-online-advertising-disclosure-guidelines/130312dotcomdisclosures.pdf.

APPENDIX A
SUMMARY OF NOTATIONS

See Table I.

APPENDIX B
REAL-WORLD EXPERIMENTS AND SIMULATIONS DETAILS

See Table II and Table III.

APPENDIX C
ADDITIONAL RESULTS FROM THE SIMULATION STUDY

A. *Effect of $N_{m,r}$ on T_r*

Figure 7 presents the effect of $N_{m,r}$ on T_r. T_r, the number of targeting formulas compatible with the set $\mathcal{N}_{m,r}$ that the maximum likelihood needs to distinguish is another important quantity that affects the inference accuracy. Intuitively, the smaller T_r, the easier the inference; in the extreme when $T_r = 1$ then there is only one formula and the maximization is guaranteed to find the true formula. Of course, quantities T_r and $N_{m,r}$ are highly correlated and a higher $N_{m,r}$ will naturally lead to a smaller T_r. Figure 7 shows how the values of T_r are distributed according to $N_{m,r}$. The box plots are obtained as follows: for each possible value of $N_{m,r}$, we take separately for each targeting formula all experiments that have this value of T_r. From this plots we observe that as long as $N_{m,r} \geq 10$ then T_r is smaller than 100. This gives an extra confidence signal: for $T_r \leq 100$, we are almost guaranteed to get high accuracy.

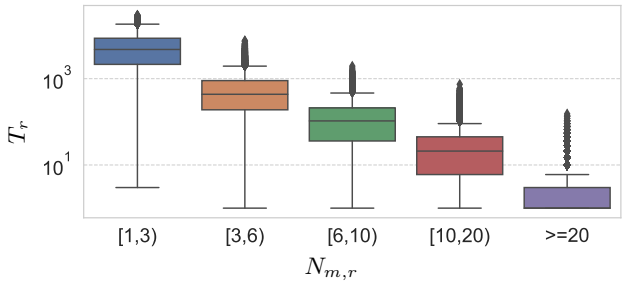


Fig. 7: Boxplot of T_r for different values of $N_{m,r}$.

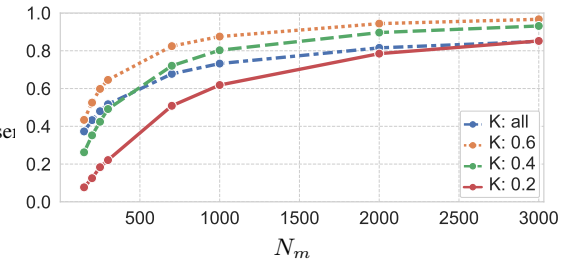
B. *Additional plots of accuracy as a function of N_m*

See Figure 8.

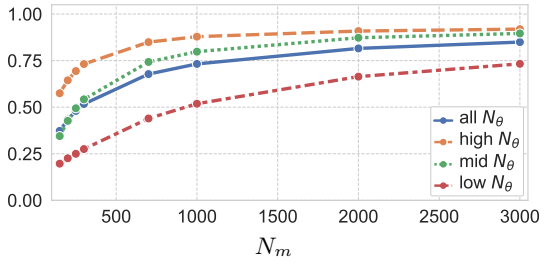
APPENDIX D
DETAILS ON THE RECRUITMENT STRATEGY AND ETHICS

A. *Additional details on the recruitment strategy*

Our main strategy to recruit participants was to publish articles in mainstream media, encouraging users to install the monitoring tool. This was done during the presidential election in Brazil. Participants were not paid for the study, and they were volunteers. Their main motivation was to help understand ad targeting during the election. Our tool provided users with a graphical interface showing them statistics about the ads they have received. We managed to attract 420 users that were active from 09/11/2018 to 29/03/2019 (where active means we collected at least one ad from them during the study period).



(a) Conditioned on K.



(b) Conditioned on N_θ.

Fig. 8: Accuracy for different values of N_m. The accuracy is aggregated over $1,000$ random θ, and $K \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

With this recruiting strategy, most of our participants are users from Brazil. Focusing on a single country reduces the size of the population while having a reasonable amount of monitored users. This is useful for our experiments because it reduces the budget needed to reach users with our ad campaigns (for a fixed N_m, the budget needs to scale with the population size, see Section V-A).[4] In addition, recruiting users is a very country-specific effort, and this effort must be replicated for each country with local specifics (such as language), which would be arduous.

Our recruitment strategy leads to a biased set of monitored users. First, Brazilian users may differ from the worldwide population. Second, and more importantly, the set of users who installed our tool may not be representative of the Brazil population, for instance because they will likely be active and more educated users. The bias in the set of monitored users may have impact at two levels:

(i) *On the inference.* The bias in the set of monitored users makes inference harder. We discuss that in details in Section V-B3. In short, our set of monitored users is biased towards users having more attributes and, through simulations recreating the bias, we can quantify that (in our experiments) it decreases the accuracy by about 12% compared to a monitored

[4]Note that we employ other strategies to further increase the chances of success of our experiments with a limited budget as described in Section III-B2.

TABLE I: Summary of the notation. As a general convention, we use calligraphic letters for sets (e.g., $\mathcal{N}$) and the corresponding regular letter to denote its size (e.g., $N = |\mathcal{N}|$).

Notation	Description	Additional mentions and related notation
Advertising platform parameters		
$\mathcal{A}$	The set of available targeting attributes	$A = \|\mathcal{A}\|$; a_j ($j = 1, \cdots, A$) denotes the attributes
$\mathcal{N}$	The population of users	$N = \|\mathcal{N}\|$; $U = [u_j^i]_{i\in\{1,\cdots,N\},j\in\{1,\cdots,A\}}$ where $u^i = [u_j^i]_{j\in\{1,\cdots,A\}}$; where $u_j^i = 1$ if the user satisfies the attribute a_j and 0 otherwise
p	The distribution on attributes up to second order	By abuse, the notation p covers both the first order probabilities $p_j = P(u_j = 1)$ and second-order probabilities $p_{jl} = P(u_j = 1, u_l = 1)$
$\mathcal{T}$	All targeting formulas of the form $a_j \wedge a_l$	$\theta \in \mathcal{T}$ the targeting formula selected by the advertiser
$\mathcal{N}_\theta \subset \mathcal{N}$	The set of users who satisfy the targeting formula θ	$N_\theta = \|\mathcal{N}_\theta\|$
$\mathcal{K} \subset \mathcal{N}_\theta$	The set of users that are shown the ad	$\|\mathcal{K}\|$ is K in expectation where K encodes the ad budget; for each user $i \in \mathcal{N}$, $y^i \in \{0, 1\}$ indicates whether or not the user is shown the ad ($y^i = 1$ iif $i \in \mathcal{K}$)
Collaborative ad transparency parameters (we can set)		
$\mathcal{N}_m \subset \mathcal{N}$	The set of monitored users	This is a key design parameter and depends on the available resources and reach of collaborative transparency method.
$\mathcal{N}_{m,\theta} \subseteq \mathcal{N}_m$	The set of monitored users that satisfy θ	$\mathcal{N}_{m,\theta} = \mathcal{N}_m \cap \mathcal{N}_\theta$; $N_{m,\theta} = \|\mathcal{N}_{m,\theta}\|$
$p_\theta = K/N_\theta$	The probability for a user in $\mathcal{N}_\theta$ to be shown the ad, identical for every user in $\mathcal{N}_\theta$	Whether or not a user is shown the ad is assumed independent of other users and/or of other ads
$\mathcal{N}_{m,r} \subseteq \mathcal{N}_{m,\theta}$	The set of monitored users who are shown the ad	$\mathcal{N}_{m,r} = \mathcal{N}_{m,\theta} \cap \mathcal{K}$; $N_{m,r} = \|\mathcal{N}_{m,r}\|$
Collaborative ad transparency inferred variables		
$\mathcal{A}_r \subset \mathcal{A}$	Attributes shared by all users in $\mathcal{N}_{m,r}$	
$\mathcal{T}_r \subset \mathcal{T}$	All the possible targeting formulas derived from the pairwise combinations of $\mathcal{A}_r$	The inference is choosing a targeting formula from the formulas available in this set
$\hat{\theta}$	The inferred targeting formula	$\hat{\theta} \in \mathcal{T}_r$

set totally unbiased. Of course, this number would change with a different set of monitored users.

(ii) *On the generation.* In Section IV-3, we validate that our population generation method produces a synthetic population that has the same statistical characteristics as the characteristics we input (i.e., p), for a p corresponding to our biased 420 users. We do not see any reason to believe, however, that the validation is specific to that p. In fact, we also did it with other populations (some of up to 700 users, coming from a different set of experiments campaigns with different users not in Brazil) and had very similar results.

B. Ethics considerations

To perform the study, we sought an IRB/ERB approval in 2017. The participants were made aware of the data being collected and had to give their consent for the collection, storing, and processing of their ad preferences, the ads they see, and consent to being targeted by our ads. The consent was initially written in English and translated to Portuguese to facilitate understanding for Brazilian users. It was done using 2017 standards in terms of ethics and data collection, which is less strict than post-GDPR standards. The study was initiated (and the ERB and consent form were made) pre-GDPR but not all data collection was done pre-GDPR. In 2019, we applied for a new ERB for the same type of data collection (with the major difference that PIIs were pseudonymized), with an updated GDPR-compliant consent, which was also approved. This approval also went through the Data Protection Officer (a new role created by the GDPR). The version of the monitoring tool active in 2018 (that collected all the data used in this paper) no longer exists as it was shut down in 2019.

The privacy concerns were handled primarily by having the data stored on a secure server behind a firewall, using the highest university security standards, and having a restricted set of people who can see/access the data.

Finally, in our study, we target users with ads. To minimize the impact such ads could have on participants, our ads were generic with neutral content. They made use of landscape stock photos provided by Facebook, and the accompanying text suggested users spend their vacation in Saarbrücken. We did not include any links or track conversions for any ad.

The 2017 ERB did not provide any specific recommendation. The 2019 ERB recommended to instruct new students working on the project about the sensitivity of the data, and to remove access to members that no longer work on the project.

Next, we describe some key differences between the ERB form and the user consent from 2017 (pre-GDPR) and 2019 (post-GDPR) versions. Both are accessible on our public repository (see link in the introduction).

1. *Sensitive nature of data*: Some of the data we collect, particularly the Facebook ad preferences, are of a potentially sensitive nature [22]. Our 2017 ERB form and consent did not discuss this (as we were not fully aware of it at the time). This was one of the main points discussed in the 2019 version.
2. *PIIs*: For this study, we collected users' PIIs (email addresses and Facebook IDs). This was used to authenticate users to show statistics about the ads they received but was mostly needed to use the custom audience feature to target the participants with our ads. This was made clear to our 2017 ERB and approved, although one would today need to be more explicit and exhaustive on which PIIs are exactly collected. We only collected hashed versions of emails and Facebook IDs in subsequent data collections (in accordance with our 2019 ERB/consent form). These are sufficient to show ads statistics to users (but not to target users via custom

TABLE II: Parameters used and observed in our real-world ad experiments. N_m–number of active monitored users during the ad campaign; $N_{m,\theta}$–number of monitored users that satisfy θ; $N_{m,r}$–number of monitored users that received our ad; A_r–number of common attributes between users in $\mathcal{N}_{m,r}$; T_r–number of θ considered by the estimators; ***B-ML*** estimator output–1 for correct inference and 0 for incorrect inference; Rank–rank of the real θ with the ***B-ML*** estimator; ***B-EM*** estimator output–1 for correct inference and 0 for incorrect inference; Sim. Acc.–the accuracy obtained in the simulation study (based on the population $\mathcal{N}$ generated from $\mathcal{D}_p$) for inferring θ to check consistency (Sec. V-B).

Attributes targeted	N_m	$N_{m,\theta}$	$N_{m,r}$	A_r	T_r	# θ inputed	*B-ML* est.	Rank (*B-ML* est.)	Attributes predicted (*B-ML* estimator)	*B-EM* est.	Sim. Acc.
Location based ad experiments											
Hip hop music & Fishing	280	37	3	138	9428	0	0	273	Discount stores & Fishing	0	0.0
Environmentalism & Tourism	280	85	4	100	4949	0	0	4841	Horses & American football	0	0.0
Music & Classical music	218	101	5	77	2926	0	0	2031	Interior design & Fashion accessories	0	0.0
Politics & Online shopping	218	173	4	89	3871	0	0	3436	Interior design & Fashion accessories	0	0.0
Music & TV	218	204	3	72	2134	0	0	441	Entertainment & Hobbies and activities	0	0.0
Magazines & Dance	225	140	5	39	741	1	0	710	Entertainment & Hobbies and activities	0	0.0
Entrepreneurship & Beverages	225	125	4	71	2478	0	0	1970	Games & Gambling	0	0.0
Interior design & Online	204	72	4	110	5652	0	0	203	Shoes & Swimming	0	0.0
Fast food & Travel	204	83	3	77	2718	0	0	606	Games & Online games	0	0.0
Design & Online shopping	204	174	3	72	2272	0	0	710	Entertainment & Hobbies and activities	0	0.0
Tea & Game consoles	214	99	6	45	990	2	0	984	Thriller movies & Hip hop music	0	0.0
Nature & Association football (Soccer)	214	179	3	101	4627	0	0	1839	Entertainment & Hobbies and activities	0	0.0
Classical music & Family	196	87	4	96	4553	0	0	2975	Furniture & Beaches	0	0.0
Construction & Sports	196	102	4	68	2278	0	0	1236	Entertainment & Hobbies and activities	0	0.0
Interior design & Food	204	70	5	86	3601	0	0	12	Design & Interior design	0	0.0
Online & Reading	204	196	3	47	1026	2	0	301	Entertainment & Hobbies and activities	0	0.0
Custom audiences ad experiments											
Cats & Tattoos	248	72	17	42	861	0	1	1	Cats & Tattoos	1	1.0
Blues music & Restaurants	248	121	37	25	300	2	1	1	Blues music & Restaurants	1	1.0
Economics & Software	248	148	50	26	325	2	1	1	Economics & Software	1	1.0
Reading & Law	248	214	73	14	91	13	0	5	Law & Televisions	1	1.0
Bars & Environmentalism	240	40	3	123	7503	0	0	4854	Insurance & Dating	0	0.0
Veganism & Software	240	90	29	33	528	2	1	1	Veganism & Software	1	1.0
Action movies & Photography	240	143	48	24	276	1	0	5	Live events & Action movies	0	1.0
Science & Beverages	240	199	74	21	210	2	1	1	Science & Beverages	1	1.0
Food & Association football (Soccer)	240	229	112	13	78	1	1	1	Food & Association football (Soccer)	1	1.0
Vacations & Clothing	198	69	18	43	903	1	1	1	Vacations & Clothing	1	0.3
Motherhood & Painting	198	75	16	37	666	0	1	1	Motherhood & Painting	1	0.7
Video games & Community issues	198	111	43	31	465	1	0	5	Rock music & Community issues	0	1.0
Association football (Soccer) & Mobile phones	198	167	56	20	190	1	1	1	Association football (Soccer) & Mobile phones	0	1.0
Graphic design & Cooking	227	88	22	33	528	0	1	1	Graphic design & Cooking	1	1.0
Coffeehouses & Law	227	100	27	37	666	0	1	1	Coffeehouses & Law	0	1.0
Higher education & Coffee	227	132	52	21	210	13	1	1	Higher education & Coffee	1	1.0
Graphic design & Drama movies	225	58	8	91	4095	0	0	81	Drama movies & Acting	0	0.0
Small business & Tea	225	24	8	105	5460	0	0	12	Small business & Drama movies	0	0.7
Family & Acting	202	59	16	48	1128	0	0	5	Acting & Dance	0	0.4
Higher education & Italian cuisine	202	44	13	52	1326	4	0	5	Documentary movies & Italian cuisine	0	1.0
Personal finance & Dogs	202	89	13	51	1275	6	0	14	Restaurants & Dogs	1	0.0
Jazz music & Painting	202	102	26	38	703	0	1	1	Jazz music & Painting	1	1.0
Electronic music & Home and garden	214	146	22	33	528	21	0	37	Family & Home and garden	0	0.1
Theatre & Cuisine	214	113	29	36	630	4	1	1	Theatre & Cuisine	1	1.0
Comics & Volleyball	195	46	15	54	1431	0	1	1	Comics & Volleyball	1	1.0
Personal finance & TV	195	150	35	26	325	10	0	5	Personal finance & Software	0	0.9
Retail & Photography	195	123	36	38	703	3	0	14	Retail & Pop music	0	1.0
Parties & Cooking	204	81	20	39	741	1	1	1	Parties & Cooking	1	1.0
Rhythm and blues music & Software	204	129	40	30	435	2	1	1	Rhythm and blues music & Software	1	1.0

TABLE III: Parameters used and observed in our real-world ad experiments and in the simulations used to check consistency (Sec. V-B), restricted to the 26 experiments for which $N_{m,r} \geq 10$. N_m–number of active monitored users during the ad campaign; $N_{m,r}$–number of monitored users that received our ad; $\bar{A}_m$–median number of attributes per monitored user; $\bar{A}_{m,\theta}$–median number of attributes per monitored user that satisfy θ; $\bar{A}_{m,r}$–median number of attributes per monitored user that received our ad; A_r–number of common attributes between users in $\mathcal{N}_{m,r}$; ***B-ML*** estimator output–1 for correct inference and 0 for incorrect inference; Sim. Acc.–accuracy obtained in the simulation study for inferring θ over 10 runs. Simulations based on $\mathcal{N}$ use a single population generation from $\mathcal{D}_p$; biased simulations use one population per experiment tailored to reproduce the bias in the set of monitored users—see Sec. V-B. In the simulations, $\bar{A}_m$, $\bar{A}_{m,\theta}$, $\bar{A}_{m,r}$, and A_r are averaged over 10 runs.

			real-world experiments					simulations on $\mathcal{N}$					biased simulations				
Attributes targeted	N_m	$N_{m,r}$	$\bar{A}_m$	$\bar{A}_{m,\theta}$	$\bar{A}_{m,r}$	A_r	BML est.	Sim. Acc.	A_r	$\bar{A}_m$	$\bar{A}_{m,\theta}$	$\bar{A}_{m,r}$	Sim. Acc.	A_r	$\bar{A}_m$	$\bar{A}_{m,\theta}$	$\bar{A}_{m,r}$
Cats & Tattoos	248	17	133.5	164.5	165	42	1	1	8	78.6	116.3	115.9	1	28	133.4	152.8	154.3
Blues music & Restaurants	248	37	133.5	160	159	25	1	1	2.7	85.8	110.7	113.7	1	16.7	134.3	148.7	148.2
Economics & Software	248	50	133.5	151	158	26	1	1	2.2	94.8	112.6	110	1	12.5	135	146.2	145.4
Reading & Law	248	73	133.5	142	145	14	0	1	2	96.5	103.5	102.4	0.3	8	133.4	137.6	138.8
Veganism & Software	240	29	135.5	160.5	173	33	1	1	3.3	79.7	110.8	111.6	0.9	20.3	134.7	153.1	152.3
Action movies & Photography	240	48	135.5	152	158	24	0	1	2.1	85.4	102.5	105.5	0.7	14.9	134.2	145.4	144.7
Science & Beverages	240	74	135.5	143	161	21	1	1	2.1	100.3	106.6	104.7	1	12.2	135.1	139.2	140.1
Food & Association football (Soccer)	240	112	135.5	137.5	134.5	13	1	1	2	88.7	90.2	91.8	0.9	7.7	134.9	135.8	136.4
Vacations & Clothing	198	18	139.5	162	173.5	43	1	0.3	5.5	77.4	106.2	105.3	0.9	26.5	134.5	145.3	143.2
Motherhood & Painting	198	16	139.5	167	159	37	1	0.7	6.4	82.7	116	114.3	0.8	31.4	134.3	152.9	155.3
Video games & Community issues	198	43	139.5	157	159	31	0	1	2.3	91.1	107.1	108.1	0.3	14.4	133.9	144.3	145.3
Association football (Soccer) & Mobile phones	198	56	139.5	142	151	20	1	1	2.1	88.5	95.1	98.7	1	14.2	134.1	136.7	137.4
Graphic design & Cooking	227	22	134	168.5	174.5	33	1	1	2.8	84.9	119.5	120.3	0.9	22.5	134.2	156.5	155.8
Coffeehouses & Law	227	27	134	167	175	37	1	1	3.2	87.2	118.4	119.2	1	22.2	135	155.2	156.7
Higher education & Coffee	227	52	134	158	152	21	1	1	2.3	94.8	115.5	117.7	1	13.1	136	148.2	146.7
Family & Acting	202	16	135.5	173	169	48	0	0.4	4.7	78.4	107.1	103.7	0.4	24.3	131.6	153.3	150.4
Higher education & Italian cuisine	202	13	135.5	182.5	164	52	0	1	8.7	78.2	121	124.9	0.6	30.6	132.3	158.7	156.0
Personal finance & Dogs	202	13	135.5	162	173	51	0	0	12.2	85.5	115.5	120.9	0	27.3	132.6	146.7	145.2
Jazz music & Painting	202	26	135.5	164	172.5	38	1	1	4	91.5	118.4	119.9	1	19.9	132.4	149.8	147.9
Electronic music & Home and garden	214	22	141	156	153.5	33	0	0.1	5.8	95.3	110.8	110.7	0	22.3	136.6	145.2	143.9
Theatre & Cuisine	214	29	141	163	172	36	1	1	3.3	88.8	114.1	115.1	1	21.3	136.8	150.7	150.6
Comics & Volleyball	195	15	139	176.5	180	54	1	1	7.7	77.3	117.2	115.7	0.9	38.1	136.6	162.1	163.9
Personal finance & TV	195	35	139	150.5	168	26	0	0.9	3.2	97.2	106.9	103.8	0.4	18.6	137.9	144.2	143.5
Retail & Photography	195	36	139	158	158	38	0	1	2.7	88.5	105	101.8	0.2	17.9	138.5	148.5	148.6
Parties & Cooking	204	20	138.5	167	158	39	1	1	4.5	82.8	117.6	117.8	1	27	137.4	157.9	162.9
Rhythm and blues music & Software	204	40	138.5	158	159	30	1	1	2.6	88.8	106	102.3	1	17.2	137.4	148.7	151.6
Average over the experiments	217.5	36.1	136.7	159.3	162.5	33.3	65.4%	86.1%	4.2	87.3	110.4	110.6	73.8%	20.4	134.9	148.6	148.7

audiences). Recommendations today ask to collect PII only if strictly needed and not to store emails (and other PIIs) and research data in the same place.

3. *Duration of the data storage*: Initially, the participants were not made aware of how long their data will be stored. The consent only informed users that the data would be anonymized after the project was completed (but without a specific timeframe). We have now indeed deleted the PIIs from the data collection of this study. In the 2019 ERB, we explain that data will be kept for six years and then anonymized before archiving (but, again, it is only pseudonymized PIIs in that case). The duration of six years was chosen according to a justified long-term scientific goal and was updated after iterations with our ERB from our initial consent form mentioning ten years (the maximum allowed by the GDPR).

4. *User's fundamental rights*: The 2019 user consent clearly states the fundamental rights provided by GDPR, e.g., why the data is collected, that the data collection is voluntary, who has access to the data, how long the data will be stored, users' rights to remove or update their data, and the right to send a complaint to local authorities. This information is legally required to appear post-GDPR.

5. *Age*: Legislation and ethics requirements when performing data collections from children and teenagers are understandably stricter; it is preferable to exclude such population from studies. In the 2019 user consent, the DPO instructed us to ask users to confirm they are 18 or older to participate.

6. *Strengthened security*: In 2019, to evaluate the security and privacy of the data collection, we performed an in-depth security homologation with the university engineers to ensure we handle potential attacks at every level of the data flow (from the front-end to the security and access of the database backups). We found a few weak points, notably at the backup level, and reinforced the security for this data collection.