Making Sense of Private Advertising: A Principled Approach to a Complex Ecosystem

Kyle Hogan
Massachusetts Institute of Technology
klhogan@csail.mit.edu

Alishah Chator*
Baruch College
alishah.chator@baruch.cuny.edu

Gabriel Kaptchuk*
University of Maryland
kaptchuk@umd.edu

Mayank Varia
Boston University
varia@bu.edu

Srinivas Devadas
Massachusetts Institute of Technology
devadas@mit.edu

Abstract

In this work, we model the end-to-end pipeline of the advertising ecosystem, allowing us to identify two main issues with the current trajectory of private advertising proposals. First, prior work has largely considered ad targeting and engagement metrics individually rather than in composition. This has resulted in privacy notions that, while reasonable for each protocol in isolation, fail to compose to a natural notion of privacy for the ecosystem as a whole, permitting advertisers to extract new information about the audience of their advertisements. The second issue serves to explain the first: we prove that *perfect* privacy is impossible for any, even minimally, useful advertising ecosystem, due to the advertisers' expectation of conducting market research on the results.

Having demonstrated that leakage is inherent in advertising, we re-examine what privacy could realistically mean in advertising, building on the well-established notion of *sensitive* data in a specific context. We identify that fundamentally new approaches are needed when designing privacy-preserving advertising subsystems in order to ensure that the privacy properties of the end-to-end advertising system are well aligned with people's privacy desires.

Keywords

advertising, privacy norms, universal composability, information leakage, attribute privacy

1 Introduction

Behavioral advertising, in which people are preferentially shown advertisements that align with their interests and demographics, has become the financial backbone of the internet and the default business model for large swaths of the technology sector. This advertising ecosystem—and the user-tracking infrastructure that powers it—are widely known to be privacy invasive [50], not only due to the collection of sensitive, personal data, but also because of the ways in which that data is *used* [17, 36, 95, 96].

A great deal of research has focused on the harms caused to people through the use of their data for the targeting of advertisements. Targeting algorithms are designed to maximize profit for the constituents of the advertising industry, not to serve the best interests of the viewer, and prior studies have shown that behavioral targeting is used to perpetuate harmful biases [3, 35, 54, 75, 88], facilitate the spread of disinformation [15, 34, 53, 102], and exploit sensitive information such as mental health data to target vulnerable populations [6, 17, 24, 98].

While targeting has been the most widely-researched component of behavioral advertising, matching ads to users is not the only functionality of the ecosystem. Advertisers expect to conduct market research using the relative success of their different advertisements. To do so, they demand *metrics* on how users responded to ads—did they buy something after viewing the ad? Subscribe to the advertiser's mailing list? If so, which ad drove this engagement? Ad networks keep records of this information and submit it back to advertisers, allowing them to improve their understanding about the preferences of their consumers and refine future ad campaigns.

Targeted advertising is not unique to digital advertising: print advertisements have long been targeting their ads towards specific groups of people using both the context in which the ad will be displayed (e.g., the magazine or billboard on which to advertise) as well as cues within the ad material itself (e.g., visuals, audio, etc.). A famous example is the 90's Subaru campaign that was targeted towards the LGBT community by including, e.g., queer-coded license plates on the depicted cars [39, 65]. The shift to digital advertising is, therefore, not a change in *type*, but a change in *magnitude* and *precision*. Digital advertising, with its specific target audiences and accurate attribution of user behavior to associated ad views, allows advertisers to conduct market research that is far more invasive than was possible with print media, calling into question the ethics of campaigns that focus on sensitive audiences. This is the case not only due to the well-understood harms of collecting sensitive data for targeting ads [85], but also because it is currently unclear what or how much information advertisers are able to learn about these sensitive audiences from the metrics released on their ad campaigns.

In response, researchers and technology companies have proposed a shift towards "privacy-preserving advertising systems," a collection of proposals [8, 13, 42, 46, 48, 51, 52, 56, 73, 74, 83, 84, 90, 92, 100, 106, 107] that aim to maintain the existing advertising business model while making the process of (1) targeting advertisements and (2) reporting metrics on advertisement efficacy in a *privacy-preserving* manner. Proponents of privacy-preserving advertising systems have made significant progress in refining the

*Work done while at Boston University

Proceedings on Privacy Enhancing Technologies 2026(1), 450–469

https://doi.org/10.56553/popets-2026-0023

design of these systems, some of which have even been deployed in popular browsers [42, 69, 100].

Unpacking "privacy" within advertising. In this work, we take a step back to re-examine "privacy-preserving advertising systems" from first principles. This choice is motivated by the desire to better understand what is possible to achieve when adding privacy to advertising when, largely, the concept feels like an oxymoron.

Ultimately, we find that this initial reaction is not far off: any *useful* behavioral advertising ecosystem must necessarily permit advertisers to extract information about end users, regardless of what privacy protections are in place. Leveraging the language of ideal functionalities, we give an implementation-independent modeling of privacy-preserving advertising that focuses on the *minimal* functionality required by the ecosystem.

Notably, our model departs from prior work in that it focuses on the entire *end-to-end* advertising pipeline, with a particular emphasis on the ways that privacy-preserving targeting and privacy-preserving metrics interact with one another and form a feedback loop. By seeing privacy-preserving advertising in this way, we are able to identify real-world advertising use cases in which common notions of privacy for targeting and metrics fail to compose satisfyingly, which undermines natural privacy guarantees for the end-to-end system despite targeting and metrics protocols *independently* achieving reasonable standards for privacy. We emphasize that this composition failure is not merely the result of specific, ill-designed protocols from early research; instead, it is fundamental to the nature of targeted advertising itself.

Looking further: we also examine how to consider privacy formally in the context of advertising. While *perfect* privacy may not be possible for advertising, we observe that it is also not necessarily required or even desirable. For instance, while contextual advertising (where ads are targeted only to the context in which they will be displayed, i.e., the website or article) can suffer from the same problems as behavioral advertising in the worst case, it is still strongly preferred as an alternative to behavioral advertising. While the privacy provided by contextual advertising may be imperfect, it is likely "good enough" [63, 99]. Hence, using the language of information leakage alone makes it difficult to distinguish between advertising systems that are widely considered "invasive" and those that are not. We additionally observe that a narrow focus on building mechanisms to regulate—without eliminating—information leakage, like differential privacy [37], risk treating all types of information leakage identically and missing the ways in which *people* feel differently about some sensitive information categories. As a result, we adopt the framework of *attribute privacy* [105] to evaluate privacy in the context of advertising.

Relatedly, not all advertising campaigns have the same potential for privacy harm, even if they do leak the same amount of information. For this reason, we also employ the framework of contextual integrity [71] to reason about the *sensitivity* of the data involved. This sentiment is captured, if not well-enforced, by current tech policy and legal regulations for ad targeting, but absent from consideration in private metrics. This motivates a more holistic approach toward the design of privacy-preserving advertising systems that reason carefully about the amount of information revealed by metrics *and* the sensitivity of that leakage.

1.1 Our Contributions

In this work, we provide a careful accounting of the structure underpinning the advertising ecosystem. In doing so, we make the following contributions:

- **A clean, formal, and flexible abstraction of the end-to-end advertising process.** Our model, detailed in Section 4, makes extensive use of parameterizing functions to ensure that our abstraction is flexible enough to describe the behavior of the real-world advertising pipeline as well as ongoing proposals to make it more private. We choose Canetti's Universally Composable (UC) security framework[1] [21] as the runtime for our modeling in order to give structure to our analysis.
- **A formal illustration of the inherent tension between privacy and utility.** In Section 3, we highlight a gap between the individual focus of current private advertising protocols and the structure of advertising itself which considers *audiences*, or groups of people. Leveraging our model, we concretize this gap in Section 5 by providing lightweight minimum utility notions required of each component of an advertising ecosystem, which we use to prove an inherent incompatibility with the group privacy notion, *attribute privacy* [105]. Specifically, we prove that any *useful* advertising ecosystem must necessarily leak some information about its users—even when it employs strong, individual privacy protections, such as differential privacy. We characterize this leakage both theoretically and empirically in terms of the difference in *sample complexity* [22]—the campaign size required for a private advertising ecosystem to leak the same information as its non-private counterpart.
- **A refocus on normative privacy notations.** Having shown that some leakage is inherent in advertising, in Section 6 we advocate for rooting future discussions of privacy in data *sensitivity* as understood by end users, ad tech platforms, and regulatory bodies. Data sensitivity is well-studied when it comes to private ad targeting, but is largely ignored by private metrics protocols, which focus exclusively on the *quantity* of leakage. We propose that by making metrics *targeting-aware*, protocols could incorporate the idea of sensitivity and serve as a second layer of enforcement–and accountability–for private advertising, refocusing on *what* information is revealed about users.

2 Background on (Private) Advertising

2.1 The Advertising Ecosystem

The language used within the advertising literature can be difficult to parse for the unfamiliar reader. As such, we provide a brief overview of digital advertising, focusing on the creation of metrics data (we direct the reader to other surveys for more detail [77]).

We illustrate the life cycle of an advertising campaign in Figure 1. A *campaign*, or collection of ads directed at a target audience, begins at step ⓪ when the advertiser registers it with the ad network. In this case, the advertiser is using their campaign to conduct an A/B test, a common practice that we discuss in depth in Section 3, by registering two ads: one ad (A) depicting a family with children and

[1] As we discuss in Section 4, the way in which we use this model is non-standard, as we use it to prove *a lack of security*.

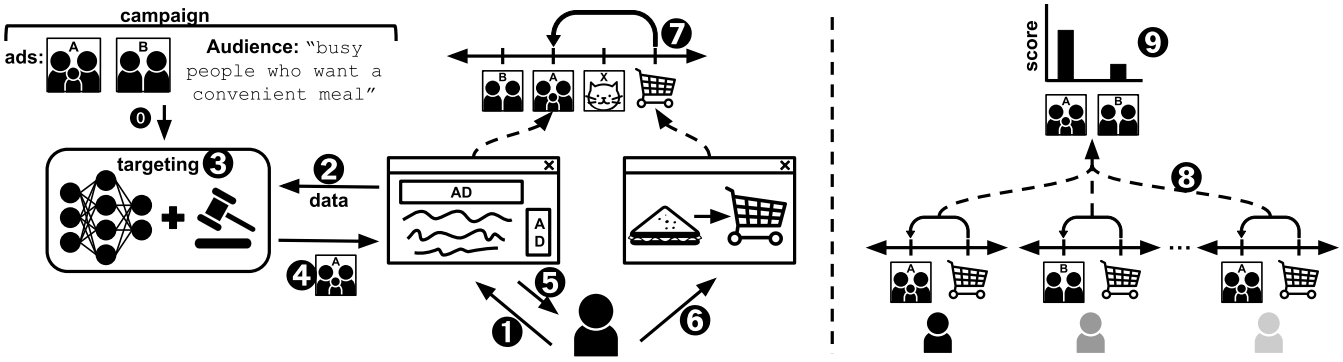

Figure 1: An illustration of the advertising ecosystem, depicting the process of generating metrics on a behavioral advertising campaign. See Section 2.1 for details.

another ad (B) without children. In both cases the ads are directed at an audience of "busy people looking for a convenient meal".

Later, ① when a user browses to a publisher website displaying ads, that website will send a request ② to the ad network containing data on the site itself as well as an identifier for the user. The ad network then runs a targeting model and a real-time bidding auction ③, incorporating data from the advertiser, in order to select which ad to deliver to the user ④. The user views the ad ⑤ and may, at some point in the future ⑥ subscribe to a meal kit service (likely on a different site) as a result of seeing this ad.

Engagement with an advertiser, such as making a purchase or even adding items to a cart, is known as a *conversion*. The ad network attributes ⑦ the user's conversion to the *impression*, or ad view, it believes was responsible. In this case, that was the most recent impression, ignoring ads from unrelated campaigns. This process is known as *attribution*, and attributing a conversion to the most recent impression is a strategy called "last touch."

Eventually, the ad network collects attribution data ⑧ from **all** users who viewed ads in this campaign and uses it to compute metrics ⑨ that the advertiser can use to determine which ad from the campaign was more successful in driving purchases.

In more detail, these metrics largely correspond to a count of how many conversions were attributed to each ad, and they are what allow the advertiser to run its A/B test and refine their strategy for future campaigns to focus more effort on ad content that drove higher engagement [14, 55]. For this example, ad A outperformed ad B, so the advertiser will likely focus future advertising spend towards parents. We go into more detail on the leakage from advertising metrics and its potential for harm in Section 3.

2.2 Related Work

While there is a rich history of academic research on privatizing advertising [8, 13, 46, 48, 51, 52, 56, 62, 74, 83, 84, 90, 92, 106, 107], the majority of this work has not considered the potential impact of releasing metrics data (beyond the possibility of linking an individual conversion report to the specific user who generated it). Exceptions to this include AdVeil [83], Themis [74], CookieMonster [90], and various industry proposals for privatizing metrics (Apple's PCM/PAM [5, 100], Google's ARA [42], Meta/Mozilla's IPA [73], and a W3C standardization effort PPA [47]) that we discuss next.

Adveil [83] presents a protocol for the full advertising pipeline, and it considers the fact that metrics reports can be revealing of personal data even if they are not directly linkable to an individual user. Themis [74] is an early industry proposal that uses a consortium blockchain to provide transparency and accountability for metrics data. It, again, provides unlinkability between users and their reports, but it does not have further privacy protections for metrics data. CookieMonster [90] is a recent work out of the W3C Private Advertising Technology Community Group (PATCG), which is working to standardize a private metrics protocol. CookieMonster provides a full model and security analysis for differentially-private ad metrics with emphasis on handling complexities in privacy loss budgets. It grew out of earlier work on Interoperable Private Attribution (IPA) [73] and is part of the work on Privacy Preserving Attribution (PPA) [47]. Apple's Private Attribution Measurement (PAM) [100] and Google's Attribution Reporting API (ARA) [42] are both alternative, differential-privacy-based proposals—though PAM was recently superseded by PPA.

To the best of our knowledge, ours is the first work to formally model and prove the *presence* of leakage for *all* advertising systems, rather than the absence of leakage for a specific system.

3 Defining Privacy for Advertising

Privacy is a multifaceted [86] and contextually embedded [70] concept that does not permit a unified definition, so we first concretize what we mean by privacy within advertising. We begin with a *leakage*-based notion; ideally, advertising systems that aim to preserve privacy should prevent *any* information from leaking about users. Emerging proposals for private advertising are rapidly moving towards this "no-leakage" world by pushing more of the targeting logic to clients' own devices (e.g., FLEDGE [78]). Such a shift is a step in the right direction—away from the mass surveillance [16, 29, 82, 94, 103] that currently supplies personal data for ad targeting. However, delivering relevant ads is only one step in the advertising pipeline.

Advertisers also want metrics on how these ads perform. Ads can be expensive, and performance metrics allow advertisers to direct their spending to campaigns that drive a better return on ad spend (ROAS) [77]. Yet, even very basic metrics, such as which ads were delivered, violate our zero-leakage goals as, due to the nature

of targeting, the ads themselves are revealing of their audience [25, 66]. Recognizing that metrics leakage could cause significant harm, there has been a widespread effort [47, 90] to make the metrics computed on advertisement performance differentially private, limiting the amount of information contained about individuals. However, as we demonstrate in this work, advertising requires more than individual privacy in order to adequately protect the information that users consider to be important.

We use this section to introduce metrics as a critical component of the advertising ecosystem, outline why it renders perfect privacy impossible for advertising, and argue that seemingly-natural fixes—such as differentially-private metrics—fail to adequately mitigate the privacy harms that can arise from advertising ecosystems.

3.1 Market Research as Information Leakage

The insights that advertisers derive from metrics go far beyond simple counts of how often advertisements are shown, and they are used to conduct market research on how users engage with the ads they are shown. By collecting metrics on the relative successes of their current advertising campaigns, advertisers can refine the content and target audience of future campaigns to focus their ad spending on serving appealing content to the people who are most likely to engage with it [41].

The most clear example of this type of market research is the A/B test, a practice where advertisers can create two versions of an ad and test which is preferred by their target audience or, conversely, test which of two possible target audiences gets better results for a given ad campaign. A/B tests are so commonplace that major advertising platforms have built-in tools for advertisers to set up their experiments.[2]

To illustrate how such tests are conducted, we revisit Figure 1 and consider an instant meal-kit company that wants to decide whether to focus its ad spend toward parents with young children. Such a company could set up its campaign in two main ways, testing on the target audience or on the ad content:

(1) Create two different ads, both depicting someone in a rush using the meal-kit to prepare a quick meal, but ad_A features a toddler and ad_B does not.
(2) Show the generic meal-kit ad (i.e., one without any particular features that suggest its relevance to parents) to two different audiences, audience A being, e.g., "busy parents looking for a convenient meal option"[3] while audience B removes the parent feature, e.g., "people looking for a convenient meal option."

No matter which A/B testing approach is leveraged, the result is fundamentally the same: advertisers learn whether members of their audience are more likely to be parents with young children based on the relative performance of A and B. In case (1), this follows from one of the core axioms of advertising: people are more likely to engage with ads that are more relevant to them [57, 64]. By contrast, in case (2) the targeting algorithm will preferentially show the ad it believes to be more relevant to the audience, i.e., if many of the audience members are parents, then ad A will be shown more frequently.

Market research is an iterative process. Consider our earlier campaign example of marketing ready-made meal kits: the first iteration could test whether the audience of "busy people who don't cook" also tends to have the "new parent" feature. Supposing that this turns out to be the case, the advertiser can then test whether "busy new parents who don't cook" tend to prefer "health-focused" meals and so forth. Thus, this practice does not only reveal a "little bit more" information about audiences, but can be used (over time) to extract tremendous amounts of information about audiences.

Much market research is, like this example, relatively innocuous. However, this same infrastructure can be (and is) used to learn about *arbitrary* topics. We see examples of this with researchers leveraging these platforms to carry out their own research studies [81]—studies that are suspiciously close to "human subject research" that is generally expected to be under the close supervision of institutional review boards. For instance, consider the study by Chan et al. that uses advertising to assess whether conservatives are likely to have stronger brand attachment [26]. Another study used advertising to assess public perceptions of refugees, though it did acknowledge the ethical considerations of the research [1].

The upshot is that market research—an inherently desired component within any advertising system—enables a level of data-mining that goes far beyond improving the quality of advertising. However, the existing discussion of privacy in advertising is centered on *individual* privacy, whereas the leakage we describe here is a *group* privacy harm.

3.2 Distributional Privacy for Advertising

At first glance, this may seem like a natural place to utilize differential privacy (DP) [37], and indeed most proposals for private advertising metrics systems use DP. However, simply privatizing the aggregated metrics in this way is insufficient. Market research involves inference over a target audience or *group* of people, not an individual user, and DP is intentionally designed to enable this type of inference [38].

Unlike a typical research study, the selection process for advertising audiences is designed to ensure that their members are *not* representative of the general population [28, 59]. Instead, audiences will often overwhelmingly represent small, arcane minority groups whose members may not even realize that such a grouping exists [67]: examples of audience profiles include "receptive to emotional messaging," "rollercoaster romantics," "heavy buyers of pregnancy tests," and "strugglers and strivers – credit reliant" [6].

Many of these audiences represent vulnerable populations and allowing advertisers to extract arbitrary information about them can be harmful even when it doesn't permit linking this information to individuals. Manipulative advertising practices use these inferences to tailor their messaging to the viewer, increasing its effectiveness [11, 89, 104]. A common example of this practice is in political advertising where ads are typically *microtargeted* to specific populations with the intent to influence their vote [60, 79, 108].

An additional challenge is that DP doesn't provide protection against an advertiser applying the group-level inferences made over the audience to its individual members [81]. A common example

[2] See, for example, Facebook A/B Testing and Google A/B Testing.
[3] While this may seem quite abstract, in practice, targeting on these types of audiences is enabled with a combination of machine learning models, externally gathered data, and identification of lookalike audiences. See Criteo's Audience Overview for more about audience generation practices.

for DP is that learning the group trend that "smoking causes cancer" would also imply that any specific smoker is at risk of cancer. But advertising takes this a step further due to "custom audiences" that can be composed of specific, identifiable individuals such as those on the advertiser's mailing list. For these audiences, a better analogy might be revealing that members of a specific sci-fi book club have an unusually high rate of cancer. Unlike in the case of smoking, this is not a global inference implying that sci-fi books cause cancer, but is instead reflective of the health status of these specific people. Revealing this type of inference—even with DP guarantees—is likely counter to peoples' expectation of privacy, especially given the information asymmetry in advertising where inferences are revealed to the advertisers, but not their audience.

For this reason, we instead employ *attribute privacy* [105] to capture the potential privacy harms from advertising market research.

3.3 Attribute Privacy in the Advertising Context

Attribute privacy, proposed by Zhang et al. [105], describes the ability of an adversary to learn information about specific, sensitive attributes of a population given summary statistics about that population. It defines sensitivity around the maximum contribution that the *distribution* of some sensitive feature in a population may have on the output of statistics computed over that population.

In this section, we provide the formal definition of attribute privacy, which is built on the pufferfish privacy framework [58].[4] Later, in Section 6, we provide some guidance on how attribute privacy could be integrated into the advertising ecosystem as a potential enforcement mechanism for user-focused privacy policies.

Definition 3.1 (Dataset Attribute Privacy, Definition 3 from Zhang et al. [105])**.** Let $(X_1^j, X_2^j, \ldots, X_m^j)$ be a record with m attributes that is sampled from an unknown distribution $\mathcal{D}$, and let $X = [X_1, \ldots, X_m]$ be a dataset of n records sampled i.i.d. from $\mathcal{D}$ where X_i denotes the (column) vector containing values of the ith attribute of every record. Let $C \subseteq [m]$ be the set of indices of sensitive attributes, and for each $i \in C$, let $g_i(X_i)$ be a function with codomain $\mathcal{U}^i$.

A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-*dataset attribute privacy* if it is (ϵ, δ)-Pufferfish private for the following framework $(S, \mathcal{Q}, \Theta)$:

Set of secrets: $S = \{s_a^i := \mathbb{1}[g_i(X_i) \in \mathcal{U}_a^i] : \mathcal{U}_a^i \subseteq \mathcal{U}^i, i \in C\}$.
Set of secret pairs: $\mathcal{Q} = \{(s_a^i, s_b^i) \in S \times S, i \in C\}$.
Distribution: Θ is a set of possible distributions θ over the dataset X. For each possible distribution $\mathcal{D}$ over records, there exists a $\theta_{\mathcal{D}} \in \Theta$ that corresponds to the distribution over n i.i.d. samples from $\mathcal{D}$.

To contextualize this definition in the advertising setting, consider the dataset X to be an advertising audience with n members, each represented by a feature vector $(X_1^j, X_2^j, \ldots, X_m^j)$ of length m indicating the attributes of the j^{th} user. Some attributes will be considered *sensitive* and represented in $C \subseteq [m]$.[5] Then:

- The secret pairs (s_a^i, s_b^i) for a sensitive attribute i are possible realizations of some function $g_i(X_i)$ over that sensitive attribute. For advertising metrics, we can think of $g_i()$ as computing the fraction of the audience who possess the sensitive attribute.
- Θ is the set of possible distributions that could have generated the audience shown in X. Each θ is intended to capture possible correlations across attributes.

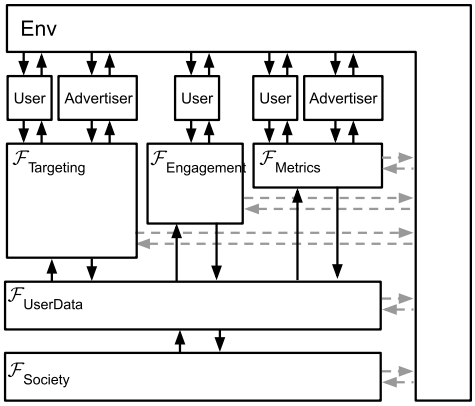

Figure 2: UC functionalities for advertising ecosystem.

Formally, the sensitivity of an output statistic $F(X)$ over the dataset X is computed as follows:

$$\Delta_i F = \max_{\theta \in \Theta} \max_{(s_a^i, s_b^i) \in \mathcal{Q}} \left| \mathbb{E}[F(X) | s_a^i, \theta] - \mathbb{E}[F(X) | s_b^i, \theta] \right|. \quad (1)$$

For advertising, we consider $F(X)$ to be the metrics for an advertising campaign (e.g., a count of ad clicks, conversions, or purchases). Sensitivity captures the maximum impact on $F(X)$, which occurs for the pair of potential secrets (s_a^i, s_b^i) in which all or none of the audience (respectively) have the sensitive attribute and for the θ with the tightest correlation between this attribute and the conversion rate. In words: if possessing the sensitive attribute makes a user significantly more (or less) likely to engage with an ad, then varying the prevalence of this feature within the audience will have a strong impact on reported number of conversions. For instance, in our example A/B test from the previous section with an audience of "busy people who don't cook," a toddler-focused ad is much more strongly correlated with the sensitive attribute (parental status) than an ad focused on the types of food contained in the meal kit.

We demonstrate later in Section 5 that a lack of attribute privacy is inherent to advertising; i.e., any minimally useful ads ecosystem will reveal some new information about its audiences. However, current instantiations and associated privacy definitions give very little control over *what* information leaks. In Section 6, we discuss the concept of sensitivity in depth and argue that advertising requires more than individual privacy in order to meet users' expectations.

4 Modeling the Advertising Ecosystem

In this section, we present our minimalist modeling of the advertising ecosystem. Our modeling captures the *minimum* information leakage present in the advertising ecosystem, and it represents an ecosystem that has been designed to eliminate all unintended information flows back to advertisers. We perform this modeling from the perspective of advertiser by having the advertiser set target audiences and receive summary reports on ad display and

[4] For some background on Pufferfish privacy, see Section E.
[5] In Section 6, we employ the contextual integrity framework [71] to provide guidance on how to decide which attributes might be sensitive in the context of advertising.

Ideal Functionality $\mathcal{F}^{n,\mathcal{D}}_{Society}$

$\mathcal{F}_{Society}$ is parameterized with a number of individuals n and a distribution over the user feature space $\mathcal{D}$.

Initialize: Upon receiving an `init` message from $\mathcal{F}_{UserData}$ for $User_i$:

(1) if $User_i$ does not already have recorded features, sample features for $User_i$ from $\mathcal{D}$ and record.
(2) Send an `init` message to $\mathcal{F}_{UserData}$ with $User_i$ and its recorded features.

Figure 3: $\mathcal{F}^{n,\mathcal{D}}_{Society}$, our way of modeling the features provided to people in society.

engagement. While other actors in the advertising ecosystem (e.g., publisher websites and ad networks) certainly have their own functionality goals, advertisers are the driving force behind the feedback loop on user data through the advertising ecosystem. In Section 5 we will use this model to "prove *insecurity*," i.e., show that useful advertising ecosystems will necessarily leak information about their users. Thus, if some future system provably instantiates our ideal functionalities, that should not be misconstrued as a demonstration that it is privacy-preserving in a normative sense. Instead, such a system would have *at least* the leakage we demonstrate here, and may have substantially more.

4.1 Parameterizing Functions

Our model makes heavy use of parameterizing functions when specifying ideal functionalities. These parameterizing functions mean that our model is *flexible* enough to capture a wide variety of potential advertising systems, including those that attempt to preserve privacy, those that are widely understood to be privacy invasive, or even systems that would make little sense to deploy in practice. We note that it might be best to think of some of these functions as being *stateful* (e.g., if a privacy budget must be managed over many queries); for simplicity, we do not explicitly manage state for these functions, but observe that it is trivial to modify our modeling to make them stateful. We briefly introduce these parameterizing functions before presenting our formal model.

The Targeting Filter ρ and Targeting Function f_t. We model the decision to select which ad for a user $user_i$ when they visit a website as a two phase process: (1) from the set of $user_i$'s `features`, a subset `features'` is extracted using a (*deterministic*) filter function ρ and is then (2) fed into an arbitrary (*randomized*) selection function f_t, the output of which is an advertisement. The filter function ρ can be thought of as a policy that limits the type of information that the targeting logic f_t is allowed to access. Real-world instantiations of ρ could model: (1) intentional restrictions such as Google's Topics API [43] and (2) unintentional inaccuracies in targeting profiles. Then, f_t could embed any "secret sauce" used by the advertising network to select the most effective advertisement to match to a user, including an opaque machine learning model or even one of the academic proposals for private ad targeting and auctions [106, 107].

The Browsing Function f_b and Engagement Function f_e. We make use of two (randomized) parameterizing functions, f_b and f_e in order to capture *human behavior* in our model. Specifically, f_b decides which website a user $User_i$ will visit, and f_e decides how a user $User_i$ will interact when presented with a particular advertisement (e.g., will they generate a "conversion," by purchasing the advertised product). These are best thought of as "black boxes" that need not be opened in order to understand the end-to-end functioning of the system.

The Attribution Function f_a. Within the advertising ecosystem, each *conversion* event must be *attributed* to an advertisement impression (see Section 2). The attribution function f_a performs the logic of this attribution. For example, a common attribution function is "last-touch attribution," where the most recent impression prior to a conversion receives all the "credit" for the conversion. In the name of generality, the parameterizing attribution function f_a takes in a set of impressions (along with the context in which the impression occurred) and allocates scores to each of these impressions according to arbitrary logic. In practice, f_a could be instantiated by the Privacy Preserving Attribution protocol [47].

The Reporting Function f_r. Advertisers learn about the performance of various advertisements within a campaign by generating a report. The exact nature of how this report is compiled from the attribution scores is system specific, but we encourage readers who want a concrete example to think about the report as simply a histogram of advertisement performance (i.e., a measure of how efficiently advertisement impressions became conversions). In many of the emerging private advertising system proposals [5, 42, 47, 73], report generation is done with differential privacy. We capture this process generically with the f_r parameterizing function, which could be instantiated by any of these proposals or by some future protocol with an alternative privacy mechanism.

4.2 Ideal Advertising Functionalities

Before providing an overview of our model, we first introduce definitions for an ad and an audience (we represent the latter using *feature vectors* in this work).

Definition 4.1 (Ad)**.** An advertisement $\texttt{ad} = \{x_1, ..., x_\ell\}, x_i \in \{0, 1\}$ is a binary vector of length ℓ. Each index in the vector represents a particular (implicit) quality the media for that advertisement could encode. When a particular index is 1, that means the feature is present in the media. Importantly, we use this formalism to describe a piece of media directly, rather than allowing an advertiser to present media and then choose a binary vector associated with that media; in this way, we assume that it is *impossible* for an advertiser to lie about the features of an ad.

Definition 4.2 (Audience)**.** An $\texttt{audience} = \{x_1, ..., x_\ell\}, x_i \in \{0, 1\}$ is also a binary vector of length ℓ. Each index in this vector encodes an attribute that members of the audiences should have. We assume that the meanings of indices for advertisements and audiences are consistent with one another—that is, the i^{th} element of each encodes the same feature.

Ideal Functionality $\mathcal{F}_{UserData}$

$\mathcal{F}_{UserData}$ has a set of users $\{User_0, User_1, ..., User_n\}$ each with a set of `features` (specified by $\mathcal{F}^{n,\mathcal{D}}_{Society}$) and a `browsing-history`. Additionally, for each $User_i$, $\mathcal{F}_{UserData}$ also maintains three indices into their `browsing-history`: $\texttt{targeting}_i$, $\texttt{engagement}_i$, and $\texttt{attribution}_i$.

Browsing: Upon receiving a `browsing` message from $\mathcal{F}^{f_b,f_e}_{Engagement}$ for $User_i$:

(1) If $User_i$ has no `features`, send an `init` message to $\mathcal{F}^{n,\mathcal{D}}_{Society}$ and record the response. Set $\texttt{targeting}_i$, $\texttt{engagement}_i$, and $\texttt{attribution}_i$ to `0`.
(2) Send $User_i$'s identity, `features`, and full `browsing-history` to $\mathcal{F}^{f_b,f_e}_{Engagement}$.
(3) Upon receiving a response from $\mathcal{F}^{f_b,f_e}_{Engagement}$ with `site`, append `site` to the `browsing-history` for $User_i$. Then, send `ok` message to $\mathcal{F}^{f_b,f_e}_{Engagement}$.

Ad Targeting: Upon receiving a `target` message from $\mathcal{F}^{f_t,\rho}_{Targeting}$ for $User_i$:

(1) If $\texttt{targeting}_i = \texttt{null}$ (i.e., `browsing` has not been called for this user) respond `fail` to $\mathcal{F}^{f_t,\rho}_{Targeting}$.
(2) Send $User_i$'s identity, `features`, and $\texttt{browsing-history}[0 : \texttt{targeting}_i]$ to $\mathcal{F}^{f_t,\rho}_{Targeting}$.
(3) Upon receiving a response from $\mathcal{F}^{f_t,\rho}_{Targeting}$ for $User_i$ of the form (`site`, `ad`), if $\texttt{browsing-history}[\texttt{targeting}_i]$ contains an matching entry `site`, then overwrite the entry with (`site`, `ad`) and increment $\texttt{targeting}_i$. Then, send `ok` to $\mathcal{F}^{f_t,\rho}_{Targeting}$.

Ad engagement: Upon receiving an `engagement` message from $\mathcal{F}^{f_b,f_e}_{Engagement}$ for $User_i$:

(1) If $\texttt{engagement}_i = \texttt{targeting}_i$ (i.e., engagement decisions were made for all impressions), respond `fail` to $\mathcal{F}^{f_b,f_e}_{Engagement}$.
(2) Respond to $\mathcal{F}^{f_b,f_e}_{Engagement}$ with $User_i$'s `features` and the tuple $\texttt{browsing-history}[\texttt{engagement}_i] = (\texttt{site}, \texttt{ad})$.
(3) Upon receiving a response from $\mathcal{F}^{f_b,f_e}_{Engagement}$ with `ad` and `conversion` for $User_i$, overwrite $\texttt{browsing-history}[\texttt{engagement}_i]$ to be (`site`, `ad`, `conversion`) and increment $\texttt{engagement}_i$. Then, send `ok` message to $\mathcal{F}^{f_b,f_e}_{Engagement}$.

Attribution: Upon receiving a `attribute` message from $\mathcal{F}^{f_a,f_r}_{Metrics}$ for $User_i$:

(1) Let $j = \texttt{attribution}_i$. Starting with j, find the first entry of `browsing-history` with a tuple (`site`, `ad`, `conversion`) such that `conversion` is non-`None`. Set $\texttt{attribution}_i$ to be the index of this entry. If no such index exists, respond `fail` to $\mathcal{F}^{f_a,f_r}_{Metrics}$ instead.
(2) Send an `attribution` message to $\mathcal{F}_{Metrics}$ for $User_i$ with $\texttt{browsing-history}[j : \texttt{attribution}_i]$.

Figure 4: $\mathcal{F}_{UserData}$ holds data about each user and is responsible to shuttling information between $\mathcal{F}_{Targeting}$, $\mathcal{F}_{Engagement}$, and $\mathcal{F}_{Metrics}$. Handles user features along with all data related to ad impressions and conversions.

Model overview. We give a high-level depiction of our model in Figure 2. Namely, our model consists of five main ideal functionalities: $\mathcal{F}^{n,\mathcal{D}}_{Society}$, $\mathcal{F}_{UserData}$, $\mathcal{F}^{f_t,\rho}_{Targeting}$, $\mathcal{F}^{f_b,f_e}_{Engagement}$, and $\mathcal{F}^{f_a,f_r}_{Metrics}$. In Figure 3, $\mathcal{F}^{n,\mathcal{D}}_{Society}$ is responsible for sampling the features for each of the n users in the system from some distribution $\mathcal{D}$. Importantly, this means that the exact features for each user is hidden from the environment—although the distribution $\mathcal{D}$ may be known to the environment. In Figure 4, $\mathcal{F}_{UserData}$ is a subroutine that serves as the shared data infrastructure of the entire system, including holding each user's features and information about their interactions with the advertising system. We note that $\mathcal{F}_{UserData}$ might be implemented in a distributed manner, such that different elements of the data may be held by different real-world computational parties. In Figures 5 to 7, $\mathcal{F}^{f_t,\rho}_{Targeting}$, $\mathcal{F}^{f_b,f_e}_{Engagement}$, and $\mathcal{F}^{f_a,f_r}_{Metrics}$ make up the core of the advertising ecosystem. Concretely, $\mathcal{F}^{f_t,\rho}_{Targeting}$ is responsible for choosing advertisements to deliver to users, $\mathcal{F}^{f_b,f_e}_{Engagement}$ is responsible for determining the websites that a user visits and how a user will interact with advertisements on those websites, and $\mathcal{F}^{f_a,f_r}_{Metrics}$ is responsible for attributing conversion events and reporting on the performance of advertisements.

Model flow details. Next, we illustrate how these functionalities work and provide a detailed description of the way data flows through the system. We note that some of our functionalities also allow for interactions to occur in a different order.

(1) Populating user features: The features associated with each user are set up on demand. Specifically, $\mathcal{F}^{n,\mathcal{D}}_{Society}$ is set up with a total number of individuals n that it will create and a distribution from which each individual's features will be sampled. The environment does not need to explicitly initiate this sampling process, as $\mathcal{F}^{n,\mathcal{D}}_{Society}$ will perform this "just in time" whenever $\mathcal{F}_{UserData}$ encounters a user with no recorded features.

(2) Registering Ad Campaigns: When an advertiser wants to send an advertisement, they begin by sending a **Register Campaign** message to $\mathcal{F}^{f_t,\rho}_{Targeting}$ (Figure 5) that specifies the explicit target `audience` to which they want to advertisements to be shown as well as the features of the advertisements $\{\texttt{ad}_1, \ldots, \texttt{ad}_k\}$ in the campaign (e.g., embedded within the visual media). We emphasize that the modeling is done such that the advertiser cannot "lie" about the semantic content of the advertisement—the feature vector *is* the advertisement.

Ideal Functionality $\mathcal{F}_{Targeting}^{f_t,\rho}$

$\mathcal{F}_{Targeting}$ is parameterized by a *stateful*, randomized function f_t and a filtering function ρ. It also maintains a set `active-campaigns`.

Register Campaign: Upon receiving a `campaign` : (`audience`, $\{\mathsf{ad}_1, \ldots, \mathsf{ad}_k\}$) message from the *Env*:

(1) Add `campaign` to `active-campaigns`.
(2) Send an `ok` message to the *Env*.

Target Ad: Upon receiving an `ad` message from *Env* for $User_i$:

(1) Send a `target` message to $\mathcal{F}_{UserData}$ for $User_i$.
(2) Upon receiving a response from $\mathcal{F}_{UserData}$ with `features` and `browsing-history` for $User_i$:
   (a) run `features'` $\leftarrow \rho(User_i, \texttt{features})$ to obtain `features'` $\subseteq$ `features`
   (b) Extract `site` from the final element of `browsing-history`.
(3) Compute `ad` $\leftarrow f_t$(`active-campaigns`, `features'`, `site`).
(4) Send $\mathcal{F}_{UserData}$ a message with the identifier $User_i$ and a tuple of the form (`site`, `ad`).
(5) Upon receiving `ok` or `fail` from $\mathcal{F}_{UserData}$, send `ok` to the *Env*.

Figure 5: Targeting functionality $\mathcal{F}_{Targeting}$

Ideal Functionality $\mathcal{F}_{Engagement}^{f_b,f_e}$

$\mathcal{F}_{Engagement}$ is parameterized by f_b that selects a site for a user to visit and f_e that determines how a user will interact with an advertisement.

Browsing: Upon receiving an `browsing` message for $User_i$ from *Env*:

(1) Send a `browsing` message to $\mathcal{F}_{UserData}$ for $User_i$.
(2) Upon receiving a response from $\mathcal{F}_{UserData}$ for $User_i$ with `features`, and full `browsing-history`, generate `site` $\leftarrow f_b$(`features`, `browsing-history`).
(3) Send *site* to $\mathcal{F}_{UserData}$ for $User_i$.
(4) Upon receiving `ok` from $\mathcal{F}_{UserData}$, send `ok` message to *Env*.

Ad Engagement: Upon receiving an `engagement` message for $User_i$ from *Env*:

(1) Send an `engagement` message to $\mathcal{F}_{UserData}$ for $User_i$.
(2) Upon receiving a response from $\mathcal{F}_{UserData}$ with $User_i$'s `features` and a tuple (`site`, `ad`), generate `conversion` $\leftarrow f_e$(`features`, `site`, `ad`). Note that `conversion` may be `None`.
(3) Respond to $\mathcal{F}_{UserData}$ with $User_i$, `ad`, and `conversion`.
(4) Upon receiving `ok` or `fail` from $\mathcal{F}_{UserData}$, send `ok` message to *Env*.

Figure 6: Engagement functionality $\mathcal{F}_{Engagement}$

(3) Initiating browsing: The environment prompts the user to browse a website by calling the **Browsing** interface of $\mathcal{F}_{Engagement}^{f_b,f_e}$. Note that the environment does not know the specific features of any given user, so we don't have the environment specify the website directly. Rather, we use f_b to choose the website that the user visits, possibly based on the user's features. Specifically, $\mathcal{F}_{Engagement}^{f_b,f_e}$ requests $User_i$'s features from $\mathcal{F}_{UserData}$ and obtains the `site` using f_b, which is defined over the features of the user and their previous browsing history. Then, $\mathcal{F}_{Engagement}^{f_b,f_e}$ informs $\mathcal{F}_{UserData}$ that $User_i$ has visited `site`.

(4) Advertisement Targeting and Delivery: To model the delivery of an advertisement to a user that has been prompted to visit a website, the environment uses the **Target Ad** interface of $\mathcal{F}_{Targeting}^{f_t,\rho}$. This triggers a `target` message to $\mathcal{F}_{UserData}$ in order to obtain the necessary information about the user and the context (i.e., `site`) in which the ad will be displayed. Next, $\mathcal{F}_{Targeting}^{f_t,\rho}$ uses ρ and f_t to select the ad that will be shown to the user. Note that the input to f_t should operate over the the `audience` associated with the advertisements. The chosen advertisement is then sent to $\mathcal{F}_{UserData}$ to be stored.

(5) User Engagement: After the user has viewed the advertisement (i.e., $\mathcal{F}_{UserData}$ holds a tuple containing a `site` and an `ad`), the environment triggers possible user engagement using the **Ad engagement** interface of $\mathcal{F}_{Engagement}^{f_b,f_e}$. In response, $\mathcal{F}_{Engagement}^{f_b,f_e}$ retrieves the necessary information from $\mathcal{F}_{UserData}$ and, using the parameterizing function f_e, determines if the user turns generates a conversion on that impression. Note that the input to f_e is the features associated with the advertising media `ad` (as the user actually sees the *media*, not the target `audience`). The results of this determination are then stored back in $\mathcal{F}_{UserData}$.

(6) Attribution: Before any metrics information can be provided, $\mathcal{F}_{Metrics}^{f_a,f_r}$ must first attribute each conversion event to at least one impression. The environment prompts this through the **Attribute** interface of $\mathcal{F}_{Metrics}^{f_a,f_r}$. When invoked in this way, $\mathcal{F}_{Metrics}^{f_a,f_r}$ calls to $\mathcal{F}_{UserData}$ and retrieves the user's conversion history. Then, $\mathcal{F}_{Metrics}^{f_a,f_r}$ updates the "scores" of each ad based on the output f_a. It is easiest to think of this step as attributing the full "credit" for the conversion to the last impression.

(7) Report Creation: Finally, the environment (as the advertiser) requests a report on the performance of its campaign. To do this, the environment invokes the **Generate Report** interface of $\mathcal{F}_{Metrics}^{f_a,f_r}$, specifying a campaign (i.e., a set of advertisements). These are then transformed into a report by the f_r function, which is also responsible for adding noise or any other privacy protection mechanism.

Ideal Functionality $\mathcal{F}_{Metrics}^{f_a,f_r}$

$\mathcal{F}_{Metrics}$ is parameterized by an attribution function f_a and report generation function f_r. Additionally, $\mathcal{F}_{Metrics}$ maintains an updatable map `ad-scores`.

Attribute: Upon receiving an `attribute` message from *Env* for $User_i$:

(1) Send a `attribute` message to $\mathcal{F}_{UserData}$ for $User_i$.
(2) Upon receiving a response from $\mathcal{F}_{UserData}$ for $User_i$ with list `browsing-history`, run $[(\mathsf{ad}_1, \mathsf{score}_1), (\mathsf{ad}_2, \mathsf{score}_2), ...] \leftarrow f_a(\texttt{browsing-history})$.
(3) For each ad_i in the output, add score_i to the entry for ad_i in `ad-scores`.
(4) Send an `ok` message to *Env*.

Generate Report: Upon receiving a `Report` message from *Env* for a $\mathsf{campaign} : (\mathsf{audience}, \{\mathsf{ad}_1, \ldots, \mathsf{ad}_k\})$,

(1) Let $\texttt{ad-scores}|_{\mathsf{campaign}}$ be a map that is a subset of `ad-scores` such that

$$\texttt{ad-scores}|_{\mathsf{campaign}} = \{(\mathsf{ad}_i, \mathsf{score}_i) \in \texttt{ad-scores} \mid \mathsf{ad}_i \in \{\mathsf{ad}_1, \ldots, \mathsf{ad}_k\}\}.$$

(2) Generate $\texttt{report} \leftarrow f_r(\texttt{ad-scores}|_{\mathsf{campaign}})$.
(3) Respond to *Env* with a message containing `report`.

Figure 7: Metrics functionality $\mathcal{F}_{Metrics}$

5 Inherent Tension Between Privacy and Usefulness

Now that we have introduced our abstract modeling of the advertising ecosystem, in this section we formalize two key concepts: (i) what does it mean for this ecosystem to be "useful" (what is the minimal functionality we need from our parameterizing functions) and (ii) what does it mean to add privacy to this ecosystem. We use this formalism to show, analytically and empirically, that privacy and utility are inherently in tension.

5.1 Defining Utility

`active-ads`. In order to make the notation more direct, we define a set `active-ads` that represents the set of advertisements from which targeting may choose. Specifically, for any `active-campaigns`, let `active-ads` be defined as follows:

$$\texttt{active-ads} = \{(\mathsf{audience}, \mathsf{ad}) \mid \{\mathsf{audience}, \texttt{ad-set}\} \in \texttt{active-campaigns}, \mathsf{ad} \in \texttt{ad-set}\}.$$

Measuring closeness. We also require a concept of *relevancy* to capture the idea that behavioral advertising is intended to show users ads that are relevant, or closely matched, to their interests and demographics. We represent this with a *close metric* that takes as input two (binary) feature vectors and outputs a score that increases as the distance between the inputs shrinks.

Targeting. We begin with our utility function for targeting. Specifically, a useful targeting system should be one that delivers ads that are more relevant to people with higher probability. We formalize this notion by saying that the probability that an one advertisement is chosen over another is proportional to the difference in *close* between the targeting audience and the user's features.[6]

Definition 5.1 (Targeting Utility)**.** A targeting function f_t is α-*useful* with respect to a distance measurement *close* and filter function ρ if, given inputs `active-campaigns`, `features`, and `site`, for all $(\mathsf{audience}_1, \mathsf{ad}_1), (\mathsf{audience}_2, \mathsf{ad}_2) \in \texttt{active-ads}$, if

$$close(\mathsf{audience}_1, \texttt{features}) - close(\mathsf{audience}_2, \texttt{features}) = \Delta, \text{ then,}$$
$$\Pr[\mathsf{ad}_1 \leftarrow f_t(\texttt{active-campaigns}, \rho(\texttt{features}), \texttt{site})] - \Pr[\mathsf{ad}_2 \leftarrow f_t(\texttt{active-campaigns}, \rho(\texttt{features}), \texttt{site})] \geq \alpha \cdot \Delta.$$

Engagement. A foundational assumption of advertising is that individuals are more likely to engage with advertisements that are more "like them." We formalize this idea using a closeness metric, similar to the one in Definition 5.1 for targeting utility.

Definition 5.2 (Engagement Utility)**.** We say that an engagement function f_e is α-*useful* with respect to a distance measurement *close* if for any set of user features `features`, website `site`, pair of advertisements $(\mathsf{ad}_1, \mathsf{ad}_2)$, and non-None conversion event `conversion`:

$$\text{if } close(\mathsf{ad}_1, \texttt{features}) - close(\mathsf{ad}_2, \texttt{features}) = \Delta,$$
$$\text{then } \Pr[\texttt{conversion} \leftarrow f_e(\texttt{features}, \texttt{site}, \mathsf{ad}_1)] - \Pr[\texttt{conversion} \leftarrow f_e(\texttt{features}, \texttt{site}, \mathsf{ad}_2)] \geq \alpha \cdot \Delta.$$

Attribution. Attribution is considered useful if it is more likely to attribute a conversion to the impression that generated it than an unrelated impression. For the purposes of our analysis, we model utility of $\mathcal{F}_{Metrics}^{f_a,f_r}$ relative to the ground truth as generated by $\mathcal{F}_{Engagement}^{f_b,f_e}$. One of the limitations of our model is that $\mathcal{F}_{Engagement}^{f_b,f_e}$ does not model cases where many impressions contribute to a single conversion event. So, while our attribution functionality is generic and can handle multi-touch attribution, the ground truth for this analysis is that a *single* ad is responsible for each conversion.

Definition 5.3 (Attribution Utility)**.** An attribution function f_a is α-*useful* with respect to an engagement function f_e if for any feature vector `features`, any pair of advertisements $(\mathsf{ad}_1, \mathsf{ad}_2)$ and associated websites `site`, and any non-None conversion event `conversion`:

$$\text{if } \Pr[\texttt{conversion} \leftarrow f_e(\texttt{features}, \texttt{site}, \mathsf{ad}_1)] - \Pr[\texttt{conversion} \leftarrow f_e(\texttt{features}, \texttt{site}, \mathsf{ad}_2)] = \Delta,$$
$$\text{then } \Pr[scores[\mathsf{ad}_1] > scores[\mathsf{ad}_2] \mid scores \leftarrow f_a(\texttt{conversion}, \texttt{browsing-history})] \geq \alpha \cdot \Delta$$

if `browsing-history` contains both ad_1 and ad_2.

Metrics. Metrics is considered useful if it permits statistical tests to be conducted on the results. That is, if some test, such as an

[6] In practice, *close* should also take in `site` as an input. However, since this context is, in theory, just a coarse-grained view into a user's features, we ignore it in order to simplify our analysis.

A/B test, could be conducted on the raw attribution data, it should still be possible to conduct this test on the aggregated and possibly noisy version of this data output by metrics. That is, the utility of metrics is defined based on what the advertiser intended to do with the attribution data.

Definition 5.4 (Metrics Utility Preserving)**.** For all h, we say that a randomized metrics report generation function $f_r : \mathcal{D}^h \to \mathcal{D}^h$ is α-*utility-preserving* with respect to a (possibly randomized) processing function $f_s : \mathcal{D}^h \to \{0, 1\}$ if for all $\hat{d} = \{d_1, ..., d_h\} \in \mathcal{D}^h$,

$$|\Pr[f_s(\hat{d}) = 1] - \Pr[f_s(\{f_r(\hat{d})\}) = 1]| < \alpha,$$

where the probabilities are over the randomness of f_r and f_s.[7]

5.2 Formal Statement

In this section, we state and prove our formal result: there is an innate tension between preventing leakage and preserving utility in an advertising ecosystem.

Theorem 1. Any ads ecosystem composed of instantiations of $\mathcal{F}_{Targeting}^{f_t,\rho}$, $\mathcal{F}_{Engagement}^{f_b,f_e}$, and $\mathcal{F}_{Metrics}^{f_a,f_r}$ that are *useful* (as defined by Definitions 5.1 to 5.4, with the additional restriction that any non-trivial implementation of $\mathcal{F}_{Metrics}^{f_a,f_r}$ must use differential privacy) for a given ad campaign will not satisfy attribute privacy for some attribute of that campaign's audience.[8]

We prove this theorem by showing that anything that could be learned by an advertiser in a non-private advertising system could similarly be learned by advertiser in a private advertising system. Proving this statement formally requires defining a game-based privacy definition *on top* of our UC modeling. To that end, we define the following random variable.

Definition 5.5. Let $\mathsf{EXEC}_{Env}^{f_t,\rho,f_b,f_e,f_a,f_r,n,\mathcal{D}}$ be a random variable denoting the output distribution of an environment *Env* when interacting with the ideal functionalities $\mathcal{F}_{Targeting}^{f_t,\rho}$, $\mathcal{F}_{UserData}$, $\mathcal{F}_{Engagement}^{f_b,f_e}$, $\mathcal{F}_{Metrics}^{f_a,f_r}$ and $\mathcal{F}_{Society}^{n,\mathcal{D}}$ (connected as shown in Figure 2) in an instance of the UC experiment.

Typically in a game-based definition, we have a *challenger* that sets up the parameters of the game (sampling randomness as needed) and an *adversary* that is required to guess some function of the *challenger*'s randomness.

Definition 5.6 (Distinguishing)**.** We say that an adversary $\mathcal{A} = (\mathcal{A}_0, Env, \mathcal{A}_1)$ succeeds in distinguishing with probability p with respect to a distribution $\mathcal{D}_0$, processing function $f_s\mathcal{D}^h \to \{0, 1\}$, and an advertising system defined by the parameters $(f_t, \rho, f_b, f_e, f_a, f_r, n)$ if:

$$p = 2 \cdot \Pr[\mathcal{A}_1(f_s(\mathsf{EXEC}_{Env(\mathcal{D}_1,aux)}^{f_t,\rho,f_b,f_e,f_a,f_r,n,\mathcal{D}_b}), aux) = b] - 1,$$

where $(\mathcal{D}_1, aux) \leftarrow \mathcal{A}_0(1^\lambda, \mathcal{D}_0)$ and $b \xleftarrow{\$} \{0, 1\}$, and the probability is taken over the random choices of $\mathcal{A}$, b, and the execution. We define the adversary's advantage $\mathsf{Adv}_{\mathcal{A},\mathcal{D}_0,f_s}^{f_t,\rho,f_b,f_e,f_a,f_r,n} \coloneqq p$.

In this definition, $\mathcal{D}_0$ should be thought of as the *ground truth* distribution of features across people in society, whereas $\mathcal{D}_1$ represents the advertiser's *prior knowledge* about how people's features are distributed. Hence, the distance between these distributions corresponds to the precision of information gained by distinguishing. When $\mathcal{D}_1$ is close to $\mathcal{D}_0$, distinguishing will be more challenging, but the advertiser can learn finer-grained information [22].

With this notion in hand, we can now state that whenever there is an adversary that can succeed at distinguishing within non-private advertising systems, then there exists an adversary that can succeed in any private version of that system that preserves utility. More precisely, a useful $\mathcal{F}_{Metrics}^{f_a,f_r}$ requires that the campaign size n was large enough to still obtain a useful result from the output of f_r^ϵ. Here, n plays the role of *sample complexity* from distribution testing, which is the number of samples necessary to distinguish between two distributions. Thus, our approach here is to show that with a sufficiently-sized campaign in the private setting, an adversary can learn the same information as in the non-private setting. Borrowing from the conventions used in distribution testing, we focus on the goal of distinguishing with advantage $\frac{2}{3}$ in the non-private setting.

Specifically, we prove the following two lemmas.

Lemma 2. *Let $\mathcal{A} = (\mathcal{A}_0, Env, \mathcal{A}_1)$ be an adversary. Consider any two ad ecosystems:*

- *A non-private ecosystem with a targeting function f_t that is α_t-useful with respect to a close metric and the identity function I as the lens, an engagement function f_e that is α_e-useful with respect to close, an attribution function f_a that is α_a-useful with respect to f_e, and that uses the identity function I for reporting.*
- *A private ecosystem with a (possibly different) targeting function f_t' that is α_t'-useful with respect to the same close metric and filtering lens ρ', and with a reporting function f_r^ϵ, where $\epsilon < 1$ that is α_r-utility-preserving with respect to a processing function $f_s : \mathcal{D}^h \to \{0, 1\}$.*

For any distribution $\mathcal{D}_0$ over $\{0, 1\}^\ell$, for $(\mathcal{D}_1, aux) \leftarrow \mathcal{A}_0(1^\lambda, \mathcal{D}_0)$, where $\mathcal{D}_1$ has the same support as $\mathcal{D}_0$ and both are over the domain $\mathcal{X}$, and for any collection of `active-ads`*: if $\mathsf{Adv}_{\mathcal{A},\mathcal{D}_0,f_s}^{f_t,I,f_b,f_e,f_a,I,n} \geq \frac{2}{3}$, then $\mathsf{Adv}_{\mathcal{A},\mathcal{D}_0,f_s}^{f_t',\rho',f_b,f_e,f_a,f_r^\epsilon,n'} \geq \frac{8}{15}$, where $n' < \frac{100n}{\epsilon} \cdot \frac{1+\alpha_t K}{1+\alpha_t' K}$. Here, K is a computable term that depends only on $\mathcal{X}$,* `active-ads`*, $\mathcal{D}_0$, $\mathcal{D}_1$ and $\frac{100n}{\epsilon} \cdot \frac{1+\alpha_t K}{1+\alpha_t' K} < \frac{100n}{\epsilon} \cdot \frac{\alpha_t}{\alpha_t'}$.*

Lemma 3. *For any ads ecosystem where there exists an adversary with distinguishing advantage $\mathsf{Adv}_{\mathcal{A},\mathcal{D}_0,f_s}^{f_t,\rho,f_b,f_e,f_a,f_r,n} > 0$, this ads ecosystem will not satisfy attribute privacy.*

Proof Sketch. We formally prove Theorem 1 in Section C. It follows immediately from Theorems 2 and 3.

To prove Theorem 2, we show that utility implies the ability to distinguish the underlying distribution used by the ads ecosystem. Namely, any successful distinguisher for a given advertising campaign in a non-private ads ecosystem can be used to construct a distinguisher for the same campaign run in a useful, private ads

[7] In practice, this definition requires that a statistical test applied to the output of the report generation function will still provide the same result as the test on the raw data, albeit with an error rate of α. This can also be thought of as requiring the same result of a t-test with a worse p-value.

[8] This is not to say that the theorem cannot hold for other instantiations of $\mathcal{F}_{Metrics}^{f_a,f_r}$, however, all existing private metrics proposals make use of differential privacy so we make our proof in this setting as well. We additionally require $\epsilon < 1$.

ecosystem, albeit with a larger campaign size. The main idea behind our proof of Theorem 2 is that, while the private version of an ads ecosystem may have less utility than a non-private version, as long as it preserves *some* utility, then it is possible to amplify this signal to match the utility of the non-private version. The "cost" of amplification is in increasing the size of the campaign n, which provides the adversary with more samples to use in its distinguishing. We can leverage distribution testing techniques [9, 18, 23] to find a bound on this new campaign size n'.[9]

The proof of Theorem 3 follows from the definition of attribute privacy (Definition 3.1) and the ability of our distinguisher to identify the underlying distribution used by the private ads ecosystem. Specifically, if attribute privacy were achieved for all parameters governing the distribution of users in $\mathcal{F}^{n,\mathcal{D}}_{Society}$, then by definition the summary statistic output by $\mathcal{F}^{f_a,f_r}_{Metrics}$ should be independent of changes to this distribution. However, were this the case, then the output of $\mathcal{F}^{f_a,f_r}_{Metrics}$ would be independent of the choice of D_0 or D_1 and no successful distinguisher could exist. Since Theorem 3 assumes the opposite, there must exist *some* parameter of the underlying distribution for which attribute privacy is *not* preserved. We explore the idea that not every attribute may require such protection, i.e., that some inferences may be acceptable, in Section 6.

5.3 Empirical Sample Complexity

To provide some intuition and empirical data for the concrete sample complexity increase that we showed theoretically in Theorem 2, we implemented our ideal advertising functionalities in Python[10] and ran the distinguishing game from Section 5.2 for concrete realizations of our parameterizing functions and α-utility parameters.

Recall that α-utility for targeting is an indication of how tightly the parameterizing function respects its definition of $close()$, or how *accurate* targeting is able to be. The expectation is that private advertising ecosystems will likely have less of an ability to find the true closest ad for a user—whether through using less user data or less precise data—and we handle this challenge by decreasing the α parameter for targeting from the non-private version. Relatedly, α-utility for engagement represents how likely a user is to click on an ad at all; this is entirely about user behavior, so it does not vary between the private and non-private ad ecosystems. Our metrics parameterizing function is instantiated using differential privacy, as this is the most common method currently being proposed.

To run our empirical distinguishing game, we fix a campaign with two ads (ad_A and ad_B) differing in a single bit b_{test}, and a distribution $\mathcal{D}_0$ representing the "ground truth" distribution of users. We then create an alternate distribution $\mathcal{D}_1$ based off the same covariance matrix as $\mathcal{D}_0$, varying the marginal probability of b_{test} in $\mathcal{D}_1$ to gradually increase the total variation distance between $\mathcal{D}_0$ and $\mathcal{D}_1$. We plot the sample complexity required to distinguish $\mathcal{D}_1$ from $\mathcal{D}_0$ at a p-value of 0.05 using the uniformly most powerful tests from Awan and Slavkovic [7] for a private ads ecosystem in Figure 8. Then, using a standard binomial test to distinguish, we plot a non-private version of the same ecosystem as well as a baseline to demonstrate the increase in sample complexity.

[9] In practice, this demonstrates the disparate impact of differential privacy between large advertisers with huge campaigns versus smaller independent advertisers.
[10] https://github.com/kylehogan/idealAdsFunctionalities

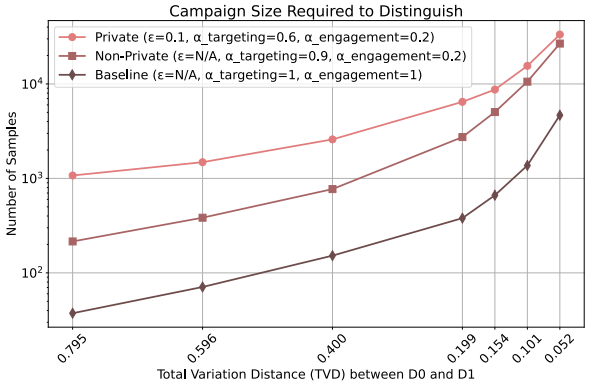

Figure 8: Impact of privatization on sample complexity.

The 'non-private' line still shows a substantial increase over the baseline due to the loss of accuracy from targeting and the drop in user engagement, while the 'private' line indicates the impact of further reducing targeting accuracy and, more importantly, introducing noise from differential privacy. The private and non-private lines begin to converge at higher sample complexity due to the lower relative impact of differential privacy for larger sample set sizes. We provide plots on the individual impacts of α-targeting, α-engagement, and ϵ-differential privacy in Section D.

6 Redefining Privacy for Advertising

Our results are a clear indication that the path forward for private advertising requires a careful re-imagining of what privacy *should* mean. That is, if we accept that some leakage is a necessary part of the advertising ecosystem, then how should advertising systems reason about the risks posed by this leakage? To start this process, we begin by introducing the perspectives of three stakeholder groups that inform this future: the people receiving targeted advertisements, advertising networks, and regulatory bodies. By identifying commonalities across these viewpoints—and gaps between existing policies and end-users needs—we hope a path forward can emerge.

In Section 6.1, we employ the contextual integrity framework [71] to help delimit when leakage may be (in)appropriate. We contrast under what conditions users, ad tech, and regulatory bodies consider different categories of data to be sensitive[11], and thus deserving of stronger protections. We then outline how existing enforcement mechanisms tailored to data sensitivity fail, despite laws and well-aligned ad tech policies. The principles underlying existing private metrics proposals offer technical solutions to some of these gaps, but they ultimately fall short of meeting user's privacy expectations. We argue that future proposals for privacy-preserving metrics must be *targeting-aware* in order to (1) discern between the implications of different information leakages and (2) understand that risks associated with leakage are context dependent. In doing so, private metrics systems of the future can apply alternate notions of privacy, such as attribute privacy [105], that can protect sensitive features about advertising audiences as a whole.

[11] In an advertising context specifically, as opposed to *generally* sensitive.

Appropriateness of Targeting on Feature			
Sensitive Data	As perceived by...		
	People [27, 49, 61, 80, 98]	Ad Tech [44]	Law EU[33],US[31]
health	⊘	⊘	⊘
relationship	⚠	⚠	✓
political beliefs	⊘	⊘	⊘
sexuality	⊘	⊘	⊘
gender	✓	⚖	⚖
location	⚠	⚖	⚖
age	✓	⚖	⚖

Figure 9: ✓ is used to indicate that targeting on this feature is acceptable/permitted, ⊘ indicates that targeting is *always* considered unacceptable/prohibited, ⚖ indicates that targeting is permitted unless illegal (in the case of tech policy) or discriminatory (in the case of regulations), and ⚠ indicates that targeting is *conditionally* acceptable.

6.1 *Sensitivity* in Advertising

Users, advertising technology companies, and regulators all agree that some types of information about individuals are *sensitive* and inappropriate to use in an advertising context. In this section, we use the framework of contextual integrity [71] to interrogate the conditions under which transmitting metrics from the ad network to the advertiser are appropriate. Explicitly, contextual integrity considers the flow of specific information types about a subject between a sender and recipient via a transmission principle and determines whether this flow is *appropriate*. In the setting of this paper, ad networks (source) transmit various types of data (ad targeting features) about the users (subject) in ad audiences to advertisers (recipient) via metrics reports (transmission principle) about ad delivery and engagement. Whether this process is perceived as appropriate largely depends on what type of information was used in targeting, which is the focus of Figure 9.

Contextual integrity allows us to bring nuance to our data categorizations. It isn't that advertisements can *never* feature health or sexuality information—in fact, it is often actively beneficial to promote awareness of mental health support options or to advertise events at LGBT organizations. Thus, our focus is specifically on data sensitivity in the context using these data types (implicitly or explicitly) to *target advertisements* and *conduct market research*, not simply displaying ads.[12]

The strongest ad tech policies [44] and modern, advertising-specific regulations like the Digital Services Act (DSA) [33] both align quite closely with user preferences. Unfortunately, it has proved difficult to enforceably put these policies into practice.

Existing policies are sophisticated and nuanced. While some types of personal data (like health data) are *always* considered sensitive in the context of ad targeting [33, 44, 49], other features can be more subtle. For example, while relationship status is used to target both ads for dating services and those for divorce lawyers, people are far less comfortable with the latter than the former, despite the same data being used in both cases [80]. This sentiment is captured by ad tech policy, which prohibits targeting based on "personal hardship"—such as divorce—or advertisements that "impose negativity" (e.g., body shaming). Similarly, targeting on the basis of features that people generally find acceptable, like age or gender, is illegal when the impact of that advertising results in *discriminatory* systems. As a result, regulations such as the Fair Housing Act (FHA) [31] have been used to prohibit use of characteristics like age, race, and gender for all housing or employment advertisements in the United States [72] and to enact changes to the targeting algorithm in the same vein [91].

Enforcement problems stymie policies' promise. Advertisement targeting is extremely opaque and largely built on machine learning models, making both technological and legal enforcement of these policies challenging [2, 19, 76]. The behavior of machine learning models is prohibitively difficult to interrogate, making it challenging to prove discrimination [3, 35, 54, 75, 88, 97]. Despite some efforts on the part of ad networks to mitigate bias [91], it has been found that these targeting systems distribute advertisements in a discriminatory way *even when it is not the intention of advertisers* [3]. It is also possible to intentionally circumvent protections using proxy features or "lookalike audiences" [4, 45, 98]. Circumventing protections this way is, of course, against policy, but even the ad networks themselves have been caught using an opaquely-defined audience to illegally target ads to children [68].

Often the opaque nature of targeting allows companies to avoid accountability with initial lawsuits struggling to prove discrimination [40, 93]. Moreover, it took until late 2023 for courts to recognize that ad networks, not only advertisers, are liable for the discriminatory targeting of ads [20]. Successful litigation has often had to circumvent the root problem of the *use* of sensitive data in targeted advertising to instead focus on how that data was *collected*, relying on regulations for deceptive business practices [30] or even wiretapping [10, 101]. This makes it burdensome for users to enforce their rights. Finally, while advertising is global, regulations decidedly are not and this ultimately limits the ability of even the strongest regulations to protect the privacy of all users.

6.2 Metrics is *Sensitivity* Agnostic

If there existed meaningful enforcement of existing laws and ad targeting policies—and confidence that these strong laws and policies applied across the full advertising ecosystem—then perhaps we would not need to be as concerned about information leakage. But without such enforcement, there is a real risk that the information leaking from the system will directly concern sensitive data. In this section, we turn our attention to metrics in the hope that it can make up for the identified failures of targeting.

Metrics is well-positioned to facilitate policy enforcement. Metrics does not face the same structural challenges that make aligning targeting systems and people's privacy preferences so difficult. First, the fraught (and legally tricky) decisions on which advertisements should be shown to which users have already been made. Second, the systems that collect and compute metrics are

[12] We again note that contextual advertising is a type of targeting and it is potentially still inappropriate to conduct market research over ads that were, for example, shown only on LGBT-focused webpages.

dramatically simpler and more transparent than those used to target advertisements. Thus, the metrics infrastructure could aid in identifying and documenting policy violations.

Users also have significant agency that they can exert when it comes to metrics. While users have no choice in the advertising networks to which they are subjected while browsing the internet, users' choice of browsers and devices are directly tied to the way their data is collected and processed within metrics. In principle, this creates an opportunity for organizations to compete in order to make their metrics systems as well-aligned with user's privacy preferences as possible. Indeed, different groups of ad tech companies are currently working on competing proposals for privatizing metrics [42, 47]. Importantly, these proposals are designed to be *interoperable*, meaning that no matter which system was used to target an advertisement, a variety of organizations, each offering a different suite of privacy protections, are capable of producing equivalent metrics output.[13]

Current proposals fall short. We identify three reasons why current proposals for privacy-preserving metrics do not adequately enforce policy. First, their technical underpinnings rely on aggregation [32] and the injection of statistically-calibrated noise (i.e., differential privacy [37]). The result is an implicit understanding that privacy in advertising is about preserving the *confidentiality* of individuals' features. As we observe in this work, however, some amount of leakage is inherent, and the leakage these systems permit is fundamentally de-contextualized, i.e., it is at odds with the understanding that *not all types of data should be treated the same*, as demonstrated by Figure 9.

Second, current metrics proposals are unaware of the content and target audience of the advertisements whose performance they measure. Thus, an ad campaign promoting clothing is treated identically to an ad campaign promoting therapy, despite the difference in the sensitivity of the data likely used to target these advertisements. Similarly, an ad campaign that uses gender to target clothing advertisements is indistinguishable from one that uses gender to target employment advertisements, despite the difference in how the law sees these campaigns. This is intentional; existing metrics systems embraced data minimization within their design, and differentiating between advertisements or audience would require collating this data across multiple systems (i.e., from targeting systems to metrics systems). While data minimization is generally the right approach for system design, in this case it has rendered metrics incapable of discerning between information leakages that people might consider harmful and innocuous.

Third, existing metrics proposals treat differential privacy as a privacy panacea, when, in fact, there are cases in which inference itself can be harmful (see Section 3.2). Specifically, there are audience types, so called "custom audiences," defined using personally identifiable information. In these cases, the inference facilitated by differential privacy has qualitatively different risk as learning a feature of their audience also gives them confidence that *individual members* of the audience possess this feature, contrary to the likely expectations of those audience members.

[13]User choice alone is likely insufficient to ensure that people's privacy is protected according to their preferences—default settings and other dark patterns are often successful in preventing users from exercising their ability to choose effectively [12].

6.3 Closing the Gap

While data sensitivity provides clear intuition for managing the risks of information leakage, the existing paradigms within which advertising systems are designed are insufficient to actualize this approach. Within targeting, there has been significant policy work to set standards for the treatment of sensitive data, but there are structural barriers to enforcing these policies. On the other hand, emerging metrics proposals are technologically sophisticated and relatively transparent, but are *incapable* of enforcing normative privacy policies because they lack context on how ads were targeted.

In order to close this gap, we advocate for expanding the approaches that are being used to think about privacy when developing new targeting and metrics proposals. There is tremendous, ongoing technical work integrating differential privacy into metrics computation in which researchers are leveraging cutting-edge privacy-enhancing technologies to significantly improve people's concrete privacy [47, 90]. These efforts, however, cannot be the sum total of the solution. Specifically, future developments need to apply the same, policy-oriented analyses to metrics that are currently being applied to targeting. We advocate for the inclusion of group privacy notions, like attribute privacy, that explicitly account for privacy harms not covered by differential privacy. Namely, as we introduced in Section 3.2 and expanded on here, *what* data is leaked can be just as, if not more, important than *how much* information is revealed. This is especially true because advertisers can combine "private" metrics with already-known information about individual members of the audience, such as their identities [81].

However, applying attribute privacy to advertising metrics requires co-design across targeting and metrics protocols as, while targeting possesses the necessary information about data sensitivity, metrics does not. Distributional privacy notions naturally require information about the underlying distribution of users targeted by an advertisement which is not currently available to metrics protocols. Making metrics systems *targeting-aware* by giving it the audience information for its reports would allow for the use of definitions like attribute privacy and could, perhaps, even permit metrics protocols to monitor targeting for policy violations or discriminatory behavior. When targeting and metrics are run by different organizations, there may even be incentives to do this type of mutual monitoring. While there will no-doubt be significant technical (and even legal) hurdles in implementing such a vision, the result would be better alignment between the privacy preferences of users and the privacy properties of advertising ecosystems.

7 Conclusion

In this work we have taken a step back to study what notions of privacy are possible within advertising. We showed that any advertising system that is even minimally useful must also allow some amount of information leakage. Taking this as a given, we identify the sensitivity of data as an important consideration when it comes to managing this leakage—a decision which has significant implications on how future privacy-preserving advertising proposals should be designed.

Acknowledgments

This research was supported by the DARPA SIEVE program under Agreement No. HR00112020021 and by the National Science Foundation under Grants No. 1955270, 2209194, 2217770, 2228610, 2230670, and 2330065.

References

[1] Claire L Adida, Adeline Lo, Lauren Prather, and Scott Williamson. 2022. Refugees to the rescue? Motivating pro-refugee public engagement during the COVID-19 pandemic. *Journal of Experimental Political Science* 9, 3 (2022), 281–295.

[2] John Albert. 2023. Not a solution: Meta's new AI system to contain discriminatory ads. https://algorithmwatch.org/en/meta-discriminatory-ads/. Accessed January 2025.

[3] Muhammad Ali, Angelica Goetzen, Alan Mislove, Elissa M. Redmiles, and Piotr Sapiezynski. 2023. Problematic Advertising and its Disparate Exposure on Facebook. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 5665–5682. https://www.usenix.org/conference/usenixsecurity23/presentation/ali

[4] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. 2019. Measuring the Facebook advertising ecosystem. In *NDSS 2019-Proceedings of the Network and Distributed System Security Symposium*. 1–15.

[5] Apple. 2023. Private Ad Measurement. https://github.com/patcg-individual-drafts/private-ad-measurement. Accessed January 2025.

[6] Adrián Astorgano. 2023. From "Heavy Purchasers" of Pregnancy Tests to the Depression-Prone: We Found 650,000 Ways Advertisers Label You. https://themarkup.org/privacy/2023/06/08/from-heavy-purchasers-of-pregnancy-tests-to-the-depression-prone-we-found-650000-ways-advertisers-label-you. *The Markup* (2023). Accessed January 2025.

[7] Jordan Alexander Awan and Aleksandra Slavkovic. 2020. Differentially Private Inference for Binomial Data. *Journal of Privacy and Confidentiality* 10, 1 (Jan. 2020). https://doi.org/10.29012/jpc.725

[8] Michael Backes, Aniket Kate, Matteo Maffei, and Kim Pecina. 2012. Obliviad: Provably secure and practical online behavioral advertising. In *2012 IEEE Symposium on Security and Privacy*. IEEE, 257–271.

[9] Ziv Bar-Yossef. 2002. *The complexity of massive data set computations*. University of California, Berkeley.

[10] Kat Black. 2024. LinkedIn sued for tracking user health data. https://www.benefitspro.com/2024/10/29/linkedin-hit-with-wave-of-health-data-claims-under-california-privacy-law-412-177165/?slreturn=20250120132542. Accessed on 20 January 2025..

[11] Benjamin E. Borenstein and Charles R. Taylor. 2024. The effects of targeted digital advertising on consumer welfare. *Journal of Strategic Marketing* 32, 3 (2024), 317–332. https://doi.org/10.1080/0965254X.2023.2218865 arXiv:https://doi.org/10.1080/0965254X.2023.2218865

[12] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. 2016. Tales from the dark side: Privacy dark strategies and privacy dark patterns. *Proceedings on Privacy Enhancing Technologies* (2016).

[13] Sanaz Taheri Boshrooyeh, Alptekin Küpçü, and Öznur Özkasap. 2018. PPAD: Privacy preserving group-based advertising in online social networks. In *2018 IFIP Networking Conference (IFIP Networking) and Workshops*. IEEE, 1–9.

[14] Michael Braun, Bart de Langhe, Stefano Puntoni, and Eric M Schwartz. 2024. Leveraging Digital Advertising Platforms for Consumer Research. *Journal of Consumer Research* 51, 1 (05 2024), 119–128. https://doi.org/10.1093/jcr/ucad058 arXiv:https://academic.oup.com/jcr/article-pdf/51/1/119/57655198/ucad058.pdf

[15] Steven Brill. 2024. You Think You Know How Misinformation Spreads? Welcome to the Hellhole of Programmatic Advertising. https://www.wired.com/story/death-of-truth-misinformation-advertising/. Accessed January 2025.

[16] Moritz Büchi, Noemi Festic, and Michael Latzer. 2022. The chilling effects of digital dataveillance: A theoretical model and an empirical research agenda. *Big Data & Society* 9, 1 (2022), 20539517211065368.

[17] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. 2018. Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. In *USENIX Security 2018*, William Enck and Adrienne Porter Felt (Eds.). USENIX Association, 479–495.

[18] Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. 2017. Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning*. PMLR, 635–644.

[19] Giuseppe Calderonio, Mir Masood Ali, and Jason Polakis. 2024. Fledging Will Continue Until Privacy Improves: Empirical Analysis of Google's {Privacy-Preserving} Targeted Advertising. In *33rd USENIX Security Symposium (USENIX Security 24)*. 4121–4138.

[20] Third Division California Court of Appeals, First District. 2023. Liapes v. Facebook, Inc. https://casetext.com/case/liapes-v-facebook-inc.

[21] Ran Canetti. 2001. Universally Composable Security: A New Paradigm for Cryptographic Protocols. In *42nd FOCS*. IEEE Computer Society Press, 136–145. https://doi.org/10.1109/SFCS.2001.959888

[22] Clément L. Canonne. 2020. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library. 1–100 pages. https://doi.org/10.4086/toc.gs.2020.009

[23] Clément L Canonne, Gautam Kamath, Audra McMillan, Adam Smith, and Jonathan Ullman. 2019. The structure of optimal private tests for simple hypotheses. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*. 310–321.

[24] Juan Miguel Carrascosa, Jakub Mikians, Ruben Cuevas, Vijay Erramilli, and Nikolaos Laoutaris. 2015. I always feel like somebody's watching me: measuring online behavioural advertising. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*. 1–13.

[25] Claude Castelluccia, Mohamed-Ali Kaafar, and Minh-Dung Tran. 2012. Betrayed by your ads! reconstructing user profiles from targeted ads. In *Proceedings of the 12th International Conference on Privacy Enhancing Technologies* (Vigo, Spain) *(PETS'12)*. Springer-Verlag, Berlin, Heidelberg, 1–17. https://doi.org/10.1007/978-3-642-31680-7_1

[26] Eugene Y Chan and Jasmina Ilicic. 2019. Political ideology and brand attachment. *International Journal of Research in Marketing* 36, 4 (2019), 630–646.

[27] Farah Chanchary and Sonia Chiasson. 2015. User Perceptions of Sharing, Advertising, and Tracking. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, Ottawa, 53–67. https://www.usenix.org/conference/soups2015/proceedings/presentation/chanchary

[28] Salim Chouaki, Islem Bouzenia, Oana Goga, and Beatrice Roussillon. 2022. Exploring the online micro-targeting practices of small, medium, and large businesses. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–23.

[29] Wolfie Christl, Katharina Kopp, and Patrick Urs Riechert. 2017. Corporate surveillance in everyday life. *Cracked Labs* 6 (2017), 2017–10.

[30] US Federal Trade Comission. 2022. FTC Charges Twitter with Deceptively Using Account Security Data to Sell Targeted Ads . https://www.ftc.gov/news-events/news/press-releases/2022/05/ftc-charges-twitter-deceptively-using-account-security-data-sell-targeted-ads. Accessed on 20 January 2025..

[31] U.S. Congress. 1970. United States Code: Fair Housing, 42 U.S.C. §§3601 - 3619.

[32] Henry Corrigan-Gibbs and Dan Boneh. 2017. Prio: Private, Robust, and Scalable Computation of Aggregate Statistics. In *14th USENIX Symposium on Networked Systems Design and Implementation (NSDI 17)*. USENIX Association, Boston, MA, 259–282. https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/corrigan-gibbs

[33] Council of European Union. 2022. Regulation (EU) 2022/2065 of the European Parliament and of the Council (Digital Services Act). https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng. Accessed on 20 January 2025. Also see https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en..

[34] Matthew Crain and Anthony Nadler. 2019. Political manipulation and internet advertising infrastructure. *Journal of Information Policy* 9 (2019), 370–410.

[35] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz. 2018. Discrimination in Online Advertising: A Multidisciplinary Inquiry. *Conference on Fairness, Accountability, and Transparency* 81 (2018), 20–34.

[36] Soteris Demetriou, Whitney Merrill, Wei Yang, Aston Zhang, and Carl A. Gunter. 2016. Free for All! Assessing User Data Exposure to Advertising Libraries on Android. In *NDSS 2016*. The Internet Society.

[37] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *TCC 2006 (LNCS, Vol. 3876)*, Shai Halevi and Tal Rabin (Eds.). Springer, Heidelberg, 265–284. https://doi.org/10.1007/11681878_14

[38] Cynthia Dwork and Deirdre K Mulligan. 2013. It's not privacy, and it's not fair. *Stan. L. Rev. Online* 66 (2013), 35.

[39] ean Halliday. 2002. Gay Ride. https://adage.com/article/news/gay-ride/52730

[40] United States District Court for the District of Maryland. 2021. Opiotennione v. Bozzuto Mgmt. https://casetext.com/case/opiotennione-v-bozzuto-mgmt-co. Civil No. 20-1956 PJM.

[41] Avi Goldfarb and Catherine Tucker. 2011. Chapter 6 - Online Advertising. Advances in Computers, Vol. 81. Elsevier, 289–315. https://doi.org/10.1016/B978-0-12-385514-5.00006-9

[42] Google. 2021. Attribution Reporting for Web overview. https://developers.google.com/privacy-sandbox/private-advertising/attribution-reporting. Accessed: 21 January 2025.

[43] Google. 2024. Topics API. https://developers.google.com/privacy-sandbox/private-advertising/topics. Accessed January 2025.

[44] Google. 2025. Google Ads policies - Advertising Policies Help. https://support.google.com/adspolicy/. Accessed on 20 January 2025.

[45] Google. 2025. Use Lookalike segments to grow your audience. https://support.google.com/google-ads/answer/13541369?hl=en. Accessed on 20 January 2025..

[46] Matthew Green, Watson Ladd, and Ian Miers. 2016. A protocol for privately reporting ad impressions at scale. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 1591–1601.

[47] W3C Private Advertising Technology Community Group. 2024. Privacy-Preserving Attribution: Level 1. https://patcg.github.io/ppa-api/. Accessed January 2025.

[48] Saikat Guha, Bin Cheng, and Paul Francis. 2011. Privad: Practical privacy in online advertising. In *USENIX conference on Networked systems design and implementation*. 169–182.

[49] Julia Hanson, Miranda Wei, Sophie Veys, Matthew Kugler, Lior Strahilevitz, and Blase Ur. 2020. Taking Data Out of Context to Hyper-Personalize Ads: Crowdworkers' Privacy Perceptions and Decisions to Disclose Private Information. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376415

[50] Russell Heimlich. 2012. Internet Users Don't like Targeted Ads. https://www.pewresearch.org/short-reads/2012/03/13/internet-users-dont-like-targeted-ads/.

[51] Leon J Helsloot, Gamze Tillem, and Zekeriya Erkin. 2017. AHEad: privacy-preserving online behavioural advertising using homomorphic encryption. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.

[52] Leon J Helsloot, Gamze Tillem, and Zekeriya Erkin. 2018. BAdASS: Preserving privacy in behavioural advertising with applied secret sharing. In *International Conference on Provable Security*. Springer, 397–405.

[53] Miikka Hiltunen. 2021. Online Political Advertising and Disinformation during Elections: Regulatory Framework in the EU and Member States. *Helsinki Legal Studies Research Paper* 68 (2021).

[54] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for Discrimination in Algorithms Delivering Job Ads *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3767–3778. https://doi.org/10.1145/3442381.3450077

[55] Garrett A Johnson. 2023. Inferno: A guide to field experiments in online display advertising. *Journal of economics & management strategy* 32, 3 (2023), 469–490.

[56] Ari Juels. 2001. Targeted advertising... and privacy too. In *Cryptographers' Track at the RSA Conference*. Springer, 408–424.

[57] Kai Kaspar, Sarah Lucia Weber, and Anne-Kathrin Wilbers. 2019. Personally relevant online advertisements: Effects of demographic targeting on visual attention and brand evaluation. *PloS one* 14, 2 (2019), e0212419.

[58] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)* 39, 1 (2014), 1–36.

[59] Aleksandra Korolova. 2010. Privacy violations using microtargeted ads: A case study. In *2010 IEEE International Conference on Data Mining Workshops*. IEEE, 474–482.

[60] Sanne Kruikemeier, Minem Sezgin, and Sophie C. Boerman. 2016. Political Microtargeting: Relationship Between Personalized Advertising on Facebook and Voters' Responses. *Cyberpsychology, Behavior, and Social Networking* 19, 6 (2016), 367–372. https://doi.org/10.1089/cyber.2015.0652 arXiv:https://doi.org/10.1089/cyber.2015.0652 PMID: 27327063.

[61] Pedro Giovanni Leon, Ashwini Rao, Florian Schaub, Abigail Marsh, Lorrie Faith Cranor, and Norman Sadeh. 2015. Privacy and behavioral advertising: Towards meeting users' preferences. In *Symposium on usable privacy and security (SOUPS)*. 22–24.

[62] Yang Liu and Andrew Simpson. 2016. Privacy-preserving targeted mobile advertising: Formal models and analysis. In *Data Privacy Management and Security Assurance: 11th International Workshop, DPM 2016 and 5th International Workshop, QASA 2016, Heraklion, Crete, Greece, September 26-27, 2016, Proceedings 11*. Springer, 94–110.

[63] Nathalie Maréchal and Nick Doty. 2024. Brief – Defining Contextual Advertising. https://cdt.org/insights/brief-defining-contextual-advertising/. Accessed January 2025.

[64] Sandra C Matz, Michal Kosinski, Gideon Nave, and David J Stillwell. 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences* 114, 48 (2017), 12714–12719.

[65] Alex Mayyasi. 2016. How Subarus came to be seen as cars for lesbians. *The Atlantic* 22 (2016).

[66] Wei Meng, Ren Ding, Simon P Chung, Steven Han, and Wenke Lee. 2016. The Price of Free: Privacy Leakage in Personalized Mobile In-Apps Ads.. In *NDSS*. 1–15.

[67] Andrew C Miller. 2026. Invisible Allies: Algorithmic Consumer Profiling and the Rise of New Group Harms. https://ssrn.com/abstract=5254258. *Yale Journal of Law & Technology* (2026).

[68] Stephen Morris, Hannah Murphy, and Hannah McCarthy. 2024. Google and Meta ignored their own rules in secret teen-targeting ad deals. https://arstechnica.com/tech-policy/2024/08/google-and-meta-ignored-their-own-rules-in-secret-teen-targeting-ad-deals/. Accessed on 20 January 2025..

[69] Mozilla. 2024. Privacy-Preserving Attribution. https://support.mozilla.org/en-US/kb/privacy-preserving-attribution. Accessed: 21 January 2025.

[70] Deirdre K Mulligan, Colin Koopman, and Nick Doty. 2016. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (2016), 20160118.

[71] Helen Nissenbaum. 2004. Privacy as contextual integrity. *Wash. L. Rev.* 79 (2004), 119.

[72] Office of Public Affairs, US Department of Justice. 2022. Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising. https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known. Accessed on 20 January 2025..

[73] W3C PATCG. 2024. Interoperable Private Attribution (IPA). https://github.com/patcg-individual-drafts/ipa?tab=readme-ov-file. Accessed: 21 January 2025.

[74] Gonçalo Pestana, Iñigo Querejeta-Azurmendi, Panagiotis Papadopoulos, and Benjamin Livshits. 2020. THEMIS: Decentralized and Trustless Ad Platform with Reporting Integrity. arXiv:2007.05556 [cs.CR]

[75] Angelisa C. Plane, Elissa M. Redmiles, Michelle L. Mazurek, and Michael Carl Tschantz. 2017. Exploring User Perceptions of Discrimination in Online Targeted Advertising. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/plane. In *26th USENIX Security Symposium (USENIX Security 17)*. USENIX Association, Vancouver, BC, 935–951.

[76] Victor Le Pochat, Laura Edelson, Tom Van Goethem, Wouter Joosen, Damon McCoy, and Tobias Lauinger. 2022. An Audit of Facebook's Political Ad Policy Enforcement. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 607–624. https://www.usenix.org/conference/usenixsecurity22/presentation/lepochat

[77] Zahra Pooranian, Mauro Conti, Hamed Haddadi, and Rahim Tafazolli. 2021. Online advertising security: Issues, taxonomy, and future directions. *IEEE Communications Surveys & Tutorials* 23, 4 (2021), 2494–2524.

[78] Privacy Sandbox. 2025. Protected Audience API (formerly known as FLEDGE). https://github.com/WICG/turtledove/blob/main/FLEDGE.md.

[79] Filipe N. Ribeiro, Koustuv Saha, Mahmoudreza Babaei, Lucas Henrique, Johnnatan Messias, Fabricio Benevenuto, Oana Goga, Krishna P. Gummadi, and Elissa M. Redmiles. 2019. On Microtargeting Socially Divisive Ads: A Case Study of Russia-Linked Ad Campaigns on Facebook. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 140–149. https://doi.org/10.1145/3287560.3287580

[80] Sarita Schoenebeck, Cami Goray, Amulya Vadapalli, and Nazanin Andalibi. 2023. Sensitive Inferences in Targeted Advertising. *Nw. J. Tech. & Intell. Prop.* 21 (2023), 155.

[81] Jeremy Seeman and Daniel Susser. 2024. Between privacy and utility: On differential privacy in theory and practice. *ACM Journal on Responsible Computing* 1, 1 (2024), 1–18.

[82] Claire M Segijn and Joanna Strycharz. 2023. The ethical ramifications of surveillance in contemporary advertising for the industry, consumers, and regulators: current issues and a future research agenda. *International Journal of Advertising* 42, 1 (2023), 69–77.

[83] Sacha Servan-Schreiber, Kyle Hogan, and Srinivas Devadas. 2021. AdVeil: A Private Targeted Advertising Ecosystem. Cryptology ePrint Archive, Paper 2021/1032. https://eprint.iacr.org/2021/1032

[84] Shashi Shekhar, Michael Dietz, and Dan S Wallach. 2012. Adsplit: Separating smartphone advertising from applications. In *Presented as part of the 21st USENIX Security Symposium (USENIX Security 12)*. 553–567.

[85] Justin Sherman. 2021. Data brokers and sensitive data on us individuals. *Duke University Sanford Cyber Policy Program* 9 (2021).

[86] Daniel J Solove. 2005. A taxonomy of privacy. *U. Pa. l. Rev.* 154 (2005), 477.

[87] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. 2017. Pufferfish Privacy Mechanisms for Correlated Data. In *Proceedings of the 2017 ACM International Conference on Management of Data* (Chicago, Illinois, USA) *(SIGMOD '17)*. Association for Computing Machinery, New York, NY, USA, 1291–1306. https://doi.org/10.1145/3035918.3064025

[88] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Potential for discrimination in online targeted advertising. In *Conference on fairness, accountability and transparency*. PMLR, 5–19.

[89] Daniel Susser, Beate Roessler, and Helen Nissenbaum. 2019. Online manipulation: Hidden influences in a digital world. *Geo. L. Tech. Rev.* 4 (2019), 1.

[90] Pierre Tholoniat, Kelly Kostopoulou, Peter McNeely, Prabhpreet Singh Sodhi, Anirudh Varanasi, Benjamin Case, Asaf Cidon, Roxana Geambasu, and Mathias Lécuyer. 2024. Cookie Monster: Efficient On-Device Budgeting for Differentially-Private Ad-Measurement Systems. In *Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles*. 693–708.

[91] Aditya Srinivas Timmaraju, Mehdi Mashayekhi, Mingliang Chen, Qi Zeng, Quintin Fettes, Wesley Cheung, Yihan Xiao, Manojkumar Rangasamy Kannadasan, Pushkar Tripathi, Sean Gahagan, Miranda Bogen, and Rob Roudani. 2023. Towards Fairness in Personalized Ads Using Impression Variance Aware

Reinforcement Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. ACM, 4937–4947. https://doi.org/10.1145/3580305.3599916
[92] Vincent Toubiana, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas. 2010. Adnostic: Privacy preserving targeted advertising. In *Proceedings Network and Distributed System Symposium*.
[93] Ninth Circuit United States Court of Appeals. 2023. Vargas v. Facebook, Inc. https://casetext.com/case/vargas-v-facebook-inc-3. No. 21-16499.
[94] Roseline van Gogh, Michel Walrave, and Karolien Poels. 2020. Personalization in Digital Marketing: Implementation Strategies and the Corresponding Ethical Issues. *The SAGE Handbook of Marketing Ethics* (2020), 411.
[95] Giridhari Venkatadri, Athanasios Andreou, Yabing Liu, Alan Mislove, Krishna P. Gummadi, Patrick Loiseau, and Oana Goga. 2018. Privacy Risks with Facebook's PII-Based Targeting: Auditing a Data Broker's Advertising Interface. In *2018 IEEE Symposium on Security and Privacy*. IEEE Computer Society Press, 89–107. https://doi.org/10.1109/SP.2018.00014
[96] Giridhari Venkatadri, Elena Lucherini, Piotr Sapiezynski, and Alan Mislove. 2019. Investigating sources of PII used in Facebook's targeted advertising. *PoPETs* 2019, 1 (Jan. 2019), 227–244. https://doi.org/10.2478/popets-2019-0013
[97] Sandra Wachter. 2020. Affinity profiling and discrimination by association in online behavioral advertising. *Berkeley Technology Law Journal* 35, 2 (2020), 367–430.
[98] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L Mazurek, and Blase Ur. 2020. What Twitter knows: Characterizing ad targeting practices, user perceptions, and ad explanations through users' own Twitter data. In *29th USENIX Security Symposium (USENIX Security 20)*. 145–162.
[99] Gabriel Weinberg. 2019. What if We All Just Sold Non-Creepy Advertising? https://www.nytimes.com/2019/06/19/opinion/facebook-google-privacy.html. Accessed January 2025.
[100] John Wilander. 2021. Introducing Private Click Measurement, PCM. https://webkit.org/blog/11529/introducing-private-click-measurement-pcm/. Accessed: 21 January 2025.
[101] United States District Judge William S. Stickman IV. 2021. Civil Action No. 2:19-cv-450 : Popa v. Harriet Carter Gifts, Inc. Available at https://casetext.com/case/popa-v-harriet-carter-gifts-inc-1. Accessed on 20 January 2025..
[102] Abby K Wood and Ann M Ravel. 2017. Fool me once: Regulating fake news and other online advertising. *S. Cal. L. Rev.* 91 (2017), 1223.
[103] Joseph T Yun, Claire M Segijn, Stewart Pearson, Edward C Malthouse, Joseph A Konstan, and Venkatesh Shankar. 2020. Challenges and future directions of computational advertising measurement systems. *Journal of advertising* 49, 4 (2020), 446–458.
[104] Lex Zard. 2023. Consumer Manipulation via Online Behavioral Advertising. arXiv:2401.00205 [cs.CY] https://arxiv.org/abs/2401.00205
[105] Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. 2022. Attribute Privacy: Framework and Mechanisms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 757–766. https://doi.org/10.1145/3531146.3533139
[106] Ke Zhong, Yiping Ma, and Sebastian Angel. 2022. Ibex: Privacy-preserving ad conversion tracking and bidding. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 3223–3237.
[107] Ke Zhong, Yiping Ma, Yifeng Mao, and Sebastian Angel. 2023. Addax: A fast, private, and accountable ad exchange infrastructure. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. 825–848.
[108] Frederik Zuiderveen Borgesius, Judith Möller, Sanne Kruikemeier, Ronan Ó Fathaigh, Kristina Irion, Tom Dobber, Balazs Bodo, and Claes H de Vreese. 2018. Online political microtargeting: Promises and threats for democracy. *Utrecht Law Review* 14, 1 (2018), 82–96.

A Proof of Theorem 2

In this appendix, we provide a formal proof of Theorem 2. Suppose $\mathcal{A} = (\mathcal{A}_0, Env, \mathcal{A}_1)$ is an adversary such that $\mathsf{Adv}^{f_t,I,f_b,f_e,f_a,I,n}_{\mathcal{A},\mathcal{D}_0,f_s} > \frac{2}{3}$ for a non-private ad ecosystem. We proceed via a sequence of games to show that this implies that $\mathcal{A}$ still has advantage $\mathsf{Adv}^{f_t',\rho',f_b,f_e,f_a,f_r^\epsilon,n'}_{\mathcal{A},\mathcal{D}_0,f_s} > \frac{8}{15}$ in the private case. That is: the adversary can continue to have a similar advantage in the private case, so long as the campaign size is appropriately amplified.

Notation. In each game $\mathcal{G}_i$, we define $\mathsf{Adv}^i_{\mathcal{A}}$ to be the advantage of adversary $\mathcal{A}^i = (\mathcal{A}^i_0, Env^i, \mathcal{A}^i_1)$ in distinguishing between executions of the specified games starting from input distributions $\mathcal{D}_0$ versus $\mathcal{D}_1$. For the first and last games, our objective is to distinguish between distributions $\mathcal{D}_0$ and chosen distribution $\mathcal{D}_1$. However, some of our intermediate games are defined with different types of outputs that correspond to partial execution of a hybrid between the non-private and private ad ecosystems. Between each game we will show that the adversary's advantage is preserved with the appropriate increase to the campaign size—which corresponds to the sample complexity of distinguishing between the distributions.

Overview. To summarize our approach, the goal of this proof is to relate the adversary's advantage of distinguishing in the private case with their advantage of distinguishing in the non-private game in terms of the relative increase in the sample complexity required to distinguish. Recall that, in our setting, the sample complexity is the campaign size. The first game $\mathcal{G}_0$ represents distinguishing between the two distributions in the non-private setting, and the final game $\mathcal{G}_4$ is in the private setting. Each game $\mathcal{G}_i$ makes a slight modification to the advertising ecosystem from the game before it $\mathcal{G}_{i-1}$, and throughout this sequence of hybrids we show that the adversary still has a non-trivial advantage in distinguishing given a certain increase in the number of samples available. These games primarily leverage how the utility definitions require the two distributions to have certain properties to have been distinguishable in the non-private setting. We then leverage results and techniques from the distribution testing literature to derive the updated sample complexity and distinguishing advantage for each game.

Facts. To begin, we state a few facts relating Hellinger distance and sample complexity that we will reference throughout the proof:

Fact 1 (Folklore, restated in [23]). *The Hellinger distance characterizes the sample complexity $SC^{P,Q} = \Theta(1/H^2(P,Q))$ between two distributions P, Q.*

Fact 2 ([9], Theorem 4.7). *If a distinguisher for two distributions P, Q has error β, then the sample complexity for distingushing between these distributions is lower bounded as follows: $SC^{P,Q} > \frac{\ln(\frac{1}{4\beta})}{4 \cdot H^2(P,Q)}$*

Fact 3 ([23], Corollary 2.2). *The sample complexity for distinguishing between two distributions P, Q is upper bounded as follows: $SC^{P,Q} < \frac{1}{H^2(P,Q)}$.*

Fact 4 ([18], Theorem 2). *If the sample complexity $SC^{P,Q}$ for distinguishing two distributions P, Q with success probability $\geq 1-\beta$ is n, then the sample complexity for distinguishing between these distributions P, Q subject to ϵ-differential privacy (where $\epsilon < 1$) $SC^{P,Q}_\epsilon$ with success probability $\geq \frac{4}{5}(1-\beta) + \frac{1}{10}$ is bounded as follows: $n \leq SC^{P,Q}_\epsilon \leq \frac{10}{\epsilon \cdot H^2(P,Q)} = \frac{10n}{\epsilon}$.*

Recall that the Hellinger distance for two discrete probability distributions P, Q with domain $\mathcal{X}$ is given by

$$H^2(P,Q) = \frac{1}{2} \sum_{\mathcal{X}} \left(\sqrt{P(x)} - \sqrt{Q(x)} \right)^2$$

.

Here and elsewhere in the proof we will use $\sum_{\mathcal{X}}$ as shorthand for $\sum_{x \in \mathcal{X}}$.

Now we will detail the series of games:

$\mathcal{G}_0$ *(Non-private ecosystem).* This game is the non-private ecosystem, so $\mathsf{Adv}^0_{\mathcal{A}} = \mathsf{Adv}^{f_t,I,f_b,f_e,f_a,I,n}_{\mathcal{A},\mathcal{D}_0,f_s} > \frac{2}{3}$.

Let k be the number of `Report` messages that *Env* sends to $\mathcal{F}^{f_a,f_r}_{Metrics}$. Then, the output of metrics is a series of reports $\hat{r} = \{r_1, ..., r_k\}$. Let f_s be the processing function that the adversary $\mathcal{A}_1$ is using to distinguish over. [14] Thus, distinguishing on this output would indicate both that:

(1) There is a difference in the underlying distribution of user features in $\mathcal{D}_0$ and $\mathcal{D}_1$, and
(2) This difference corresponds to the features targeted by the two ads that are being A/B tested.

From the first observation that $\mathcal{D}_0$ and $\mathcal{D}_1$ are different, Fact 1 gives us that the Hellinger distance characterizes the sample complexity $SC^{\mathcal{D}_0,\mathcal{D}_1} = \Theta(1/H^2(\mathcal{D}_0, \mathcal{D}_1))$ between two distributions. From the second observation that the statistical test f_s is measuring some specific property of the distributions (i.e., A/B testing on a feature), we find that the distributions must differ in regards to that particular property.

$\mathcal{G}_1$ *(Switch to* $\mathcal{A}' = (\mathcal{A}_0, Env', \mathcal{A}'_1)$*):* In this game, we consider an adversary who receives the *input* of Metrics rather than the output. This is essentially a data processing inequality: anything that is performed by Metrics can instead be performed by the adversary itself if it is helpful in its distinguishing role.

In more detail, consider a modified *Env*' that does not interact with $\mathcal{F}^{f_a,f_r}_{Metrics}$, and uses the input to the metrics functionality as its output. In particular, the output distribution from *Env*' is a function of the total output from $\mathcal{F}^{f_b,f_e}_{Engagement}$. Now, we will show why $\mathcal{A}_1$ implies we can construct a $\mathcal{A}'_1$ that can distinguish here. If the output of $\mathcal{F}^{f_a,f_r}_{Metrics}$ is distinguishable, then its input must also be distinguishable with advantage at least $\mathsf{Adv}^0_{\mathcal{A}}$. Note that the input to metrics is the output of engagement, which is a list of browsing histories $\{(\mathtt{site}, \mathsf{ad}, \mathtt{conversion})\}$ for each user. Formally, we can say that for n users, the output of engagement is $\{\mathtt{browsing\text{-}history}_1, ..., \mathtt{browsing\text{-}history}_n\}$, where for user i with features $\mathtt{features}_i$:

$$\mathtt{browsing\text{-}history}_i = \{(\mathtt{site}_{i,1}, \mathsf{ad}_{i,1}, f_e(\mathtt{features}_i, \mathtt{site}_{i,1}, \mathsf{ad}_{i,1})), ...\}$$

where

$$\mathsf{ad}_{i,1} = f_t(\mathtt{active\text{-}campaigns}, \mathtt{features}_i, \mathtt{site}_{i,1}).$$

Additionally, we note that for these to result in a distinguishable set of reports, there must be some pair:

$$(\mathtt{audience}_1, \mathsf{ad}_1), (\mathtt{audience}_2, \mathsf{ad}_2) \in \mathtt{active\text{-}ads}$$

that had distinguishable metrics. Since metrics simply performs post-processing on the result of engagement, there exists an adversary $\mathcal{A}'_1$ that can distinguish this pair with advantage $\mathsf{Adv}^1_{\mathcal{A}} \geq \mathsf{Adv}^0_{\mathcal{A}}$.

[14] To provide some practical intuition, one could imagine that f_s is akin to Fisher's exact test being used to perform A/B testing on the reports outputted by metrics (which could be thought of as counting the number of clicks per ad).

$\mathcal{G}_2$ *(Add* (f'_t, ρ)*):* Targeting and Engagement utility mean that some information about the difference in the two distributions is still leaked, and we can use distribution testing to distinguish here. So, in this game, we replace (f_t, I) with (f'_t, ρ), where f'_t is α'_t useful with respect to *close* and ρ. As in the previous game, we focus on the task of $\mathcal{A}^2$ being able to distinguish the combined output of $\mathcal{F}^{f_b,f_e}_{Engagement}$ with respect to the pair:

$$(\mathtt{audience}_1, \mathsf{ad}_1), (\mathtt{audience}_2, \mathsf{ad}_2) \in \mathtt{active\text{-}ads}.$$

In particular, we want to show that by expanding the campaign size, we can construct an adversary that will be able to still distinguish with the same advantage. We leverage the fact that the size of the campaign determines the number of samples pulled from the distribution $\mathcal{D}_b$. Thus, by finding the sample complexity for distinguishing between the output distributions, we can determine the necessary campaign size. We note that the adversary could in theory construct a small campaign or one that is not relevant to the difference between the distributions, but our requirement that the adversary is able to distinguish eliminates the need to consider such scenarios.

We note that advertisers will try to use $\mathcal{F}^{f_t,\rho}_{Targeting}$ to deliver an ad dependent on a particular test feature. In this case, the addition of (f'_t, ρ) continues to allow targeting based on this feature, however, just less accurately (our analysis also holds for the case where $\alpha'_t = 0$). Formally, for the pair $(\mathtt{audience}_1, \mathsf{ad}_1), (\mathtt{audience}_2, \mathsf{ad}_2)$ and a feature vector representing a user x, if

$$close(\mathtt{audience}_1, x) - close(\mathtt{audience}_2, x) = \Delta_x,$$

then:

$$\Pr[\mathsf{ad}_1 \leftarrow f_t(\mathtt{active\text{-}campaigns}, \rho(x), \mathtt{site})] - \Pr[\mathsf{ad}_2 \leftarrow f_t(\mathtt{active\text{-}campaigns}, \rho(x), \mathtt{site})] \geq \alpha'_t \cdot \Delta_x,$$

where $\alpha'_t < \alpha_t$.

We note that the size of the campaign that is distinguishable in the non-private setting is n with success probability $\frac{5}{6}$ (since the advantage is $\frac{2}{3}$). Additionally, the distribution of the output of engagement is the distribution of ad conversions. In particular, we can look at the conversions for $(\mathtt{audience}_1, \mathsf{ad}_1)$ in comparison to $(\mathtt{audience}_2, \mathsf{ad}_2)$. We express this by

$$R_b(x) = \Pr[x \textit{ sampled from } \mathcal{D}_b] \Pr[\textit{show } x \ \mathsf{ad}_1] \Pr[x \textit{ clicks on } \mathsf{ad}_1] = \mathcal{D}_b(x) \cdot (\frac{1 + \alpha_t \cdot \Delta_x}{2}) \cdot (\alpha_e \cdot close(\mathsf{ad}_1, x)).$$

From Fact 1 we know $n = \Theta(1/H^2(R_0, R_1))$ and in particular from Facts 2 and 3 that

$$\frac{\ln(\frac{3}{2})}{4 \cdot H^2(R_0, R_1)} < n < \frac{1}{H^2(R_0, R_1)}.$$

For the private case, we have the updated distributions for the output of engagement:

$$R'_b(x) = \Pr[x \textit{ sampled from } \mathcal{D}_b] \Pr[\textit{show } x \ \mathsf{ad}_1] \Pr[x \textit{ clicks on } \mathsf{ad}_1] = \mathcal{D}_b(x) \cdot (\frac{1 + \alpha'_t \cdot \Delta_x}{2}) \cdot (\alpha_e \cdot close(\mathsf{ad}_1, x)).$$

Therefore, the campaign size needed for the adversary to distinguish would be

$$\frac{\ln(\frac{3}{2})}{4 \cdot H^2(R'_0, R'_1)} < n'_t < \frac{1}{H^2(R'_0, R'_1)}.$$

Note that $\mathcal{D}_0, \mathcal{D}_1, R_0, R_1, R'_0, R'_1$ all have the same domain $\mathcal{X}$, which is the universe of potential users.

In order to compute the bound on expansion factor $z = n_{t'}/n$ we must take an upper bound for the private case and a lower bound for the non-private case. Let $\gamma = \frac{4}{\ln(\frac{3}{2})}$, then

$$\begin{aligned}
z &= \frac{\gamma H^2(R_0, R_1)}{H^2(R'_0, R'_1)} \\
&= \frac{\gamma \frac{1}{2} \sum_{\mathcal{X}} \left(\sqrt{R_0(x)} - \sqrt{R_1(x)}\right)^2}{\frac{1}{2} \sum_{\mathcal{X}} \left(\sqrt{R'_0(x)} - \sqrt{R'_1(x)}\right)^2} \\
&= \gamma \frac{1}{2} \sum_{\mathcal{X}} \left(\sqrt{\mathcal{D}_0(x) \cdot \left(\frac{1 + \alpha_t \cdot \Delta_x}{2}\right) \cdot (\alpha_e \cdot \text{close}(\text{ad}_1, x))} - \right. \\
&\quad \left. \sqrt{\mathcal{D}_1(x) \cdot \left(\frac{1 + \alpha_t \cdot \Delta_x}{2}\right) \cdot (\alpha_e \cdot \text{close}(\text{ad}_1, x))} \right)^2 \\
&\quad / \frac{1}{2} \sum_{\mathcal{X}} \left(\sqrt{\mathcal{D}_0(x) \cdot \left(\frac{1 + \alpha'_t \cdot \Delta_x}{2}\right) \cdot (\alpha_e \cdot \text{close}(\text{ad}_1, x))} - \right. \\
&\quad \left. \sqrt{\mathcal{D}_1(x) \cdot \left(\frac{1 + \alpha'_t \cdot \Delta_x}{2}\right) \cdot (\alpha_e \cdot \text{close}(\text{ad}_1, x))} \right)^2 \\
&= \frac{\gamma \alpha_e \sum_{\mathcal{X}} \text{close}(\text{ad}_1, x)(\frac{1+\alpha_t \cdot \Delta_x}{2}) \left(\sqrt{\mathcal{D}_0(x)} - \sqrt{\mathcal{D}_1(x)}\right)^2}{\alpha_e \sum_{\mathcal{X}} \text{close}(\text{ad}_1, x)(\frac{1+\alpha'_t \cdot \Delta_x}{2}) \left(\sqrt{\mathcal{D}_0(x)} - \sqrt{\mathcal{D}_1(x)}\right)^2} \\
&= \frac{\gamma \sum_{\mathcal{X}} \text{close}(\text{ad}_1, x)(1 + \alpha_t \cdot \Delta_x) \left(\sqrt{\mathcal{D}_0(x)} - \sqrt{\mathcal{D}_1(x)}\right)^2}{\sum_{\mathcal{X}} \text{close}(\text{ad}_1, x)(1 + \alpha'_t \cdot \Delta_x) \left(\sqrt{\mathcal{D}_0(x)} - \sqrt{\mathcal{D}_1(x)}\right)^2} \\
&= \gamma \frac{A + \alpha_t B}{A + \alpha'_t B} = \gamma \frac{1 + \alpha_t K}{1 + \alpha'_t K}.
\end{aligned}$$

Thus, we have $n_{t'} < n \cdot \gamma \frac{1+\alpha_t K}{1+\alpha'_t K}$ where

$$A = \sum_{\mathcal{X}} \text{close}(\text{ad}_1, x) \left(\sqrt{\mathcal{D}_0(x)} - \sqrt{\mathcal{D}_1(x)}\right)^2$$

,

$$B = \sum_{\mathcal{X}} \text{close}(\text{ad}_1, x) \Delta_x \left(\sqrt{\mathcal{D}_0(x)} - \sqrt{\mathcal{D}_1(x)}\right)^2$$

and $K = \frac{B}{A}$.

As shown above we computed this ratio by simplifying down the expression

$$\frac{\gamma \frac{1}{2} \sum_{\mathcal{X}} \left(\sqrt{R_0(x)} - \sqrt{R_1(x)}\right)^2}{\frac{1}{2} \sum_{\mathcal{X}} \left(\sqrt{R'_0(x)} - \sqrt{R'_1(x)}\right)^2}.$$

We did this by expanding out the terms defining these distributions and identifying that the differences in the numerator and denominator both had like terms that could be factored out. Many of these like terms then canceled out between the numerator and denominator and we were left with a relatively simple expression of the ratio.

By looking at how the expression $\frac{1+\alpha_t K}{1+\alpha'_t K}$ grows with K and noting that $0 \leq K$ we can find additional bounds on n'_t. In particular, $1 \leq \frac{1+\alpha_t K}{1+\alpha'_t K} < \frac{\alpha_t}{\alpha'_t}$ so we have

$$n \cdot \gamma < n_{t'} < n \cdot \gamma \frac{\alpha_t}{\alpha'_t}.$$

The lower bound tells us that when K approaches 0, which indicates that $B << A$ (there is a very small difference in the closeness of users to the two audiences) then this ratio is dominated by γ. However, the upper bound tells us that when $A << B$ (there is a very large difference in the closeness of users to the two audiences) then this ratio is dominated by $\frac{\alpha_t}{\alpha_t}$. Thus, private targeting mechanisms make the most difference when there is a large difference in how close users are to the two potential audiences in question.

As a result, there exists an adversary can distinguish by increasing the number of samples (campaign size) to $n_{t'}$ to distinguish. They can use their knowledge of $\mathcal{D}_0, \mathcal{D}_1, \alpha_t, \alpha'_t, \rho, \alpha_e$ to compute this value. Thus $\text{Adv}^2_{\mathcal{A}} = 2/3 = \text{Adv}^1_{\mathcal{A}}$.

$\mathcal{G}_3$ (Distinguish on metrics again): In this game, we use the fact that metrics doesn't do anything currently, so it does not impact the adversary's advantage. Hence, we will switch back to interacting with *Env*, whose output is a function of the output of $\mathcal{F}^{f_a, f_r}_{Metrics}$, where we are still using I as our report creation function. Thus, like in the non-private case, reporting is transparently forwarding through the results from engagement. Thus, $\text{Adv}^3_{\mathcal{A}} = \text{Adv}^2_{\mathcal{A}}$.

$\mathcal{G}_4$ (Add f^{ϵ}_r): In this final game, we leverage the fact that metrics has utility, and therefore being able to still perform statistical tests in the private setting means the differences in the distributions remain detectable and we can use differentially-private distribution testing techniques to distinguish here. In more detail, we now use the reporting function f^{ϵ}_r (with respect to f_s) instead of the identity I. Because f^{ϵ}_r is α_r-utility-preserving with respect to f_s, this means that the impact of the reporting function on the statistical test cannot be more that α_r. As mentioned in Definition 5.4, α_r can be thought of as a bound on the change in the error rate for f_s.

Like in our previous games, our goal here is to show that by amplifying the size of the campaign, we can still allow for the adversary to distinguish with a noticeable advantage. Crucially, in order to use this to find the necessary sample complexity, we need to show that the two distributions are not the same, as otherwise the result will be undefined.

We know that the underlying $\mathcal{D}_0$ and $\mathcal{D}_1$ distributions are different and that from $\mathcal{G}_2$ that this difference is distinguishable on the output of engagement. Thus, the question is whether the reporting function may completely flatten this difference. However, we know that f^{ϵ}_r is bounded in how much it can impact the result of the statistical test f_s that is performed on the data. Since this test is measuring a property that is different between $\mathcal{D}_0$ and $\mathcal{D}_1$ (observation (2) from $\mathcal{G}_0$), that means that f^{ϵ}_r cannot completely flatten this difference. Thus, the distributions of interest here: $f^{\epsilon}_r(R'_0)$ and $f^{\epsilon}_r(R'_1)$ must have a non zero difference. (This is a slight abuse

of notation but $f_r^\epsilon(R'_b)$ is the final distribution from applying the reporting function to the engagement output distribution R'_b.)

We can make use of Fact 4 which gives us that the sample complexity for distingushing between these distributions $f_r^\epsilon(R'_0), f_r^\epsilon(R'_1)$ $SC_\epsilon^{P,Q}$ is $n'_t \leq SC_\epsilon^{R'_0,R'_1} \leq \frac{10n'_t}{\epsilon}$ with advantage $2(\frac{4}{5}(\frac{\text{Adv}^3_{\mathcal{A}}+1}{2})+\frac{1}{10})-1 = \frac{8}{10}\text{Adv}^3_{\mathcal{A}}$.

Thus, we can apply this and set the amplified campaign size to $n_r = \frac{10n'_t}{\epsilon} < \frac{10\frac{4}{\ln(\frac{3}{2})}n}{\epsilon} \cdot n \cdot \gamma\frac{1+\alpha_t K}{1+\alpha'_t K}$. This value is computable with the information used to construction R'_b as well as the ϵ parameter used in reporting. With this campaign size we have a distinguishing advantage $\text{Adv}^4_{\mathcal{A}} = \frac{8}{10}\text{Adv}^3_{\mathcal{A}}$.

The overall lemma follows by combining the bounds on each pair of adjacent games. This gives us a distinguishing advantage $p' = \text{Adv}^4_{\mathcal{A}} = \frac{8}{10}\text{Adv}^3_{\mathcal{A}} = \frac{8}{10}\frac{2}{3} = \frac{8}{15} = p - \frac{2}{15}$.

B Proof of Theorem 3

We will now show why any ad ecosystem where an adversary $\mathcal{A} = (\mathcal{A}_0, Env, \mathcal{A}_1)$ has advantage $\text{Adv}^{f_t,\rho,f_b,f_e,f_a,f_r,n}_{\mathcal{A},\mathcal{D}_0,f_s} > 0$ in distinguishing between two distributions $\mathcal{D}_0$ and $\mathcal{D}_1$, means that this ad ecosystem cannot achieve Dataset Attribute Privacy as defined in Definition 3.1. In this case, the audience of users is a set of feature vectors, where each attribute is just a single bit that is 1 if the user has that feature.

Assume for the sake of contradiction that this ad ecosystem has dataset attribute privacy where the private function $g(X_i)$ is a sum of the values of the i^{th} feature. Also, let f_s be the summary statistic F, which reveals how many users in the dataset had that feature.

From the fact that a distinguisher $\mathcal{A}$ exists, we know that the output of the environment is distinguishable. This directly implies that the output of $\mathcal{F}^{f_a,f_r}_{Metrics}$ is distinguishable between the two probability distributions $\mathcal{D}_0$ and $\mathcal{D}_1$. In particular, if k was the total number of reports that *Env* requested from $\mathcal{F}^{f_a,f_r}_{Metrics}$, the output it distinguishes on is a list of reports $\hat{r} = \{r_1, ..., r_k\}$. Let f_s be the processing function such that the output of the adversary is the output of $f_s(\hat{r})$. Since, $\hat{r} = \{r_1, ..., r_k\}$ is distinguishble, there must be some $r \in \hat{r}$ that is distinguishable.

By Definition 3.1 if attribute privacy were achieved for all parameters governing the distribution of users in $\mathcal{F}^{n,\mathcal{D}}_{Society}$ then the output of $f_s(\hat{r})$ should be independent of which distribution $\mathcal{F}^{n,\mathcal{D}}_{Society}$ sampled users from ($\mathcal{D}_0$ vs $\mathcal{D}_1$). However, $\mathcal{A}$ is able to distinguish on this output, which implies that there must be some report where the result is dependent on choice of distribution. Therefore, there must be some attribute for which attribute privacy is not preserved.

This is a contradiction to our claim that we had attribute privacy, and thus we conclude that the existence of an adversary that can distinguish between the underlying distributions in an ad ecosystem implies no attribute privacy.

C Proof of Theorem 1

Our proof of Theorem 1 follows from Theorem 2 and Theorem 3 in a straightforward manner. Suppose we had an ads ecosystem composed of instantiations of $\mathcal{F}^{f_t,\rho}_{Targeting}$, $\mathcal{F}^{f_b,f_e}_{Engagement}$, and $\mathcal{F}^{f_a,f_r}_{Metrics}$ that are *useful* (as defined by Definitions 5.1 to 5.4).

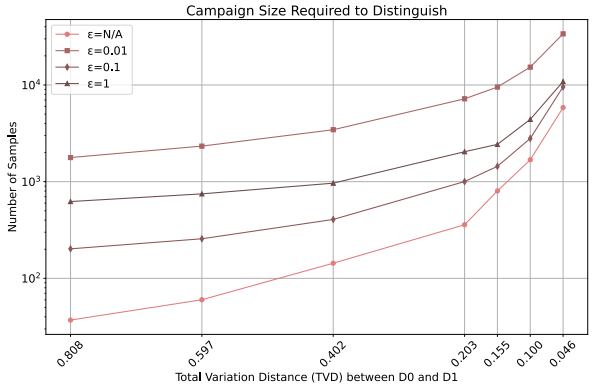

Figure 10: Impact of ϵ on sample complexity.

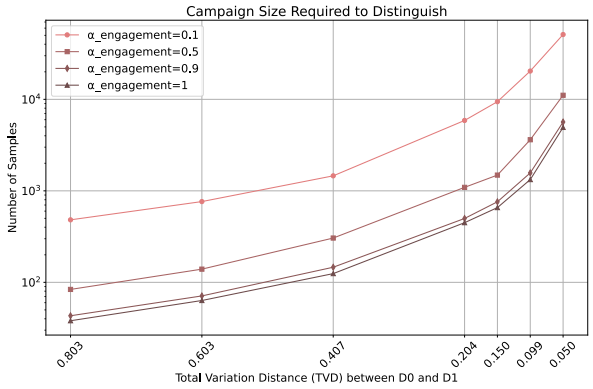

Figure 11: Impact of α-engagement on sample complexity.

By Theorem 2 we see that for such an ad ecosystem, we will have an adversary $\mathcal{A}$ that can distinguish between $\mathcal{D}_0$ and $\mathcal{D}_1$ with probability $\frac{8}{15}$. By, Theorem 3 we see that if such a distinguishing adversary exists, then this ad ecosystem will not satisfy attribute privacy for some attribute of that campaign's audience.

Thus, we conclude that any useful ads ecosystem for a given ad campaign will not satisfy attribute privacy for some attribute of that campaign's audience.

D Extended Empirical Evaluation

As shown in Figure 10, lower ϵ values increase the amount on noise added to the differentially-private metrics which reduces the confidence of the distinguisher at low sample sizes. As the number of samples increases, the signal to noise ratio improves and the sample complexity of the differentially-private lines approaches the baseline.

α-engagement impacts the overall click through rate so reducing this value reduces the overall click probability as shown in Figure 11. Consequently, this reduces the absolute difference in click probability between $\mathcal{D}_0$ and $\mathcal{D}_1$ making distinguishing more challenging. Generally, ad campaigns have very low click through

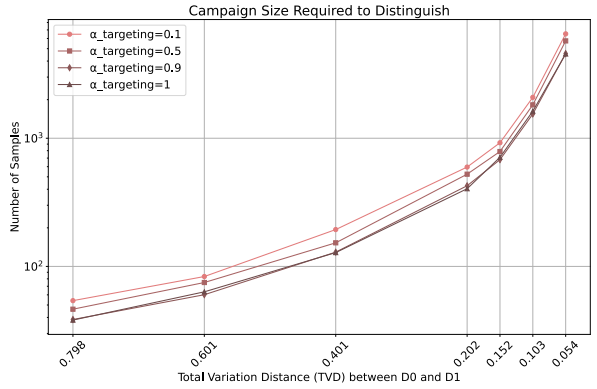

Figure 12: Impact of α-targeting on sample complexity.

rates (less than a percent) so low α-engagement values are the most realistic.

Figure 12 depicts how α-targeting has a less strong impact on overall sample complexity in distinguishing than either α-engagement or the ϵ value for differentially-private metrics. This is largely due to the way we use it as a modifier for $close()$ in the targeting utility definition—most users are not a perfect match for the ad feature vector and may be relatively close to both ads causing a small difference in probability of preferring one over the other even when α-targeting is 1. Alternative designs for targeting utility could have α-targeting instead indicate how likely targeting is to prefer the closer ad, regardless of *how* close it was and in this case α-targeting would have a stronger impact.

Additionally, we might expect the audience distribution and definition of $close()$ to affect how impactful α-targeting is. E.g., if the audience is well-specified and the users tend to share similar features aside from the test bit, then $close()$ might decide to weight that bit more strongly than the others when deciding which ad is more relevant.

E Background on Pufferfish Privacy

The Pufferfish privacy framework [58] is intended to capture scenarios where privacy is desired for sensitive data, and, moreover, the sensitive data might be correlated with some of the other features in the dataset. Whereas differential privacy treats all data as sensitive and therefore requires hiding all correlations, Pufferfish privacy is more flexible and allows for revealing some features but only up to the point that sensitive data remains hidden. Consider for example genetic traits or the transmission of disease: information about the hair color of a person's family members can allow for inferences of their own hair color and knowing that many people in a person's community have the flu is revealing of that person's health data. Resolving this under standard DP typically involves considering these correlated groups of people as a single entry. While this is effective, it introduces unmanageable levels of noise as the group size grows.

The Pufferfish framework allows for differentially-private statistics on correlated data without assuming that all entries are fully correlated. This allows it to achieve protection for correlated data at a more manageable level of noise. The framework consists of three parts:

(1) A set of secrets $\mathcal{S}$: this is the information that should **not** be revealed (or inferrable) by any output statistics. E.g., from the examples above, the secrets would be the flu status of any individual or their hair color.
(2) Pairs of potential secrets $\mathcal{Q} = (\mathcal{S} \times \mathcal{S})$: these are the values of secrets that should be indistinguishable given the output. For example: perhaps it should not be possible to distinguish between "Alice has the flu" and "Alice is healthy," or to distinguish between any combination of possible hair colors for Alice (e.g., brown vs. green hair, brown vs. blonde hair, etc). This list should be viewed as a denylist: any two events X and Y that do *not* form a pair (X, Y) in the list are (implicitly) allowed to be distinguishable by the adversary.
(3) A set of distributions Θ that could plausibly generate the dataset: Θ defines the correlations between the individual datapoints in the dataset. E.g., different Θ could specify varying levels of contagiousness for the flu or different probabilities that an individual has a certain hair color given the hair colors of their family members.

Definition E.1 (Definition 2.1 from Song et al. [87])**.** A privacy mechanism M is said to be ϵ-Pufferfish private in a framework $(\mathcal{S}, \mathcal{Q}, \Theta)$ if for all $\theta \in \Theta$ with X drawn from distribution θ, for all secret pairs $(s_i, s_j) \in \mathcal{Q}$, and for all $w \in \text{Range}(M)$, we have

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(X) = w | s_i, \theta)}{P_{M,\theta}(M(X) = w | s_j, \theta)} \leq e^{\epsilon} \tag{2}$$

when s_i and s_j are such that $P(s_i|\theta) \neq 0$, $P(s_j|\theta) \neq 0$.

Attribute privacy uses the Pufferfish framework to focus on privatizing the distribution of sensitive attributes within a dataset. There, the secrets $\mathcal{S}$ are the output of some function $g()$ over that sensitive attribute. Like most parameters in Pufferfish privacy, exactly what $g()$ is will be situational and differ between use cases. Here, we care about the fraction of the audience who possesses the sensitive attribute so $g()$ computes this value.