Having your Privacy Cake and Eating it Too: Platform-supported Auditing of Social Media Algorithms for Public Interest

BASILEAL IMANA, University of Southern California, USA
ALEKSANDRA KOROLOVA, Princeton University, USA
JOHN HEIDEMANN, USC/Information Science Institute, USA

Relevance estimators are algorithms used by major social media platforms to determine what content is shown to users and its presentation order. These algorithms aim to personalize the platforms' experience for users, increasing engagement and, therefore, platform revenue. However, at the large scale of many social media platforms, many have concerns that the relevance estimation and personalization algorithms are opaque and can produce outcomes that are harmful to individuals or society. Legislations have been proposed in both the U.S. and the E.U. that mandate auditing of social media algorithms by external researchers. But auditing at scale risks disclosure of users' private data and platforms' proprietary algorithms, and thus far there has been no concrete technical proposal that can provide such auditing. Our goal is to propose a new method for *platform-supported auditing* that can meet the goals of the proposed legislations. The first contribution of our work is to enumerate these challenges and the limitations of existing auditing methods to implement these policies at scale. Second, we suggest that limited, privileged access to *relevance estimators* is the key to enabling generalizable platform-supported auditing of social media platforms by external researchers. Third, we show platform-supported auditing need not risk user privacy nor disclosure of platforms' business interests by proposing an auditing framework that protects against these risks. For a particular fairness metric, we show that ensuring privacy imposes only a small constant factor increase (6.34× as an upper bound, and 4× for typical parameters) in the number of samples required for accurate auditing. Our technical contributions, combined with ongoing legal and policy efforts, can enable public oversight into how social media platforms affect individuals and society by moving past the privacy-vs-transparency hurdle.

1 INTRODUCTION

Algorithms that estimate relevance of online content to users are at the core of how social media platforms shape which information they show to their users. A major part of the platforms' value comes from their ability to deliver content in a personalized way – by choosing which organic content from a user's connections or which promoted content from advertisers should be shown. Popular platforms driven by relevance optimization algorithms include Facebook, Instagram, Twitter and TikTok. While the platforms do not publicly disclose in detail the specific algorithms and inputs to the algorithms they use to rank content by relevance, they acknowledge they do leverage sophisticated machine learning algorithms that harvest the extensive data they have on users [28, 49, 61, 79].

Social media platforms promise to provide a more meaningful experience for users, increasing user-engagement, and, therefore, platform revenue. Platforms consider such optimization proprietary. The combination of opaque algorithms that select what is shared with their operation at large scale has created increasing concern that relevance algorithms can produce outcomes that are harmful to individuals or society. Examples of harms include biased or discriminatory ad targeting and delivery [2, 5, 44, 51], amplification of hateful content [67], political polarization [36, 42, 43, 65, 70], and promotion of addictive behavior in teens [32].

Audits have shown that some of the hypothesized algorithmic harms are not merely theoretical by providing evidence of the side-effects of relevance optimization on how both organic content and promoted content (ads) are delivered. For example, several studies have shown relevance optimization algorithms used in delivering ads can result in biased, and even discriminatory, outcomes in contexts such as politics [3, 31], housing [2], and employment [2, 44, 50]. For organic content, studies have looked at how the shift from a reverse-chronological news feed to one that is curated by a relevance algorithm distorts information to which users are exposed [9, 35]. An internal audit by Twitter also showed its algorithmic newsfeed amplifies organic political content non-uniformly across the political spectrum [43]. These findings show greater transparency is needed for platforms to continue to innovate while staying accountable and building societal trust in their systems.

The importance of these risks has prompted policy or legal efforts aimed to increase transparency in social media platforms. Of many such proposals [16, 24, 25, 55, 68, 83], the U.S. Platform Accountability and Transparency Act (PATA [66, 83]) and the E.U. Digital Services Act (DSA [25]) are the most comprehensive, addressing broad algorithmic risks and platforms [63]. PATA has bipartisan support, but has not yet been debated in the U.S. Congress. On the other hand, the E.U. passed DSA and plans to start enforcement in 2024 [26, 27]. Both proposals mandate that platforms make data available to vetted, external researchers, who will conduct studies to audit platforms' algorithms and evaluate their alignment with societal and legal expectations. We call such proposals *platform-supported auditing*, and discuss how they extend existing auditing taxonomies in §4.1.

A critical concern outlined in both legislative proposals is the need to protect the privacy of a platform's users *and* its proprietary algorithms. While a desirable policy goal, prior research has shown increasing transparency without violating the privacy of users and the business interests of platforms presents technical challenges [1, 4, 6, 11, 43, 63]. Platforms such as Facebook have also cited privacy as a constraint on increasing their transparency efforts [15, 85]. Until now, no actionable proposals have been put forth for how to address the privacy concerns and implement the auditing policies practically and at scale.

Our first contribution is to enumerate limitations of existing auditing methods for implementing platform-supported auditing at scale (§2). We start with an overview of what DSA and PATA compel social media platforms to make available to external auditors, and the scope of the target platforms the proposals cover. We then enumerate the significant limitations and non-generalizability of existing external auditing methods to study algorithmic harms on these platforms. Specifically, although existing methods have been crucial to detecting how various social media platforms harm different demographic groups and our society at large, they do not generalize well to study multiple types of harms, demographic groups or platforms.

Our second contribution is to suggest that transparency of *relevance estimators* is the key to enabling a generalizable and actionable framework for platform-supported auditing (§3). Our proposal provides a plausible, practical approach to platform-supported auditing. While the DSA and PATA require auditing legislatively, they do not specify a mechanism; our approach is the first to meet this need. We show the importance of auditing relevance estimators by examining platforms' documentations that show they are the "brains" that shape delivery of every piece of organic and promoted content on social media. Despite being the core drivers, these algorithms are used across multiple social-media platforms with little transparency into their definitions of relevance or the specific inputs they use to optimize for it. We survey prior audits that indirectly measured how use of relevance estimators' can result in harmful outcomes to show a means to directly query and audit these algorithms is the key to increasing transparency and providing a meaningful path to verifying alignment with societal and legal expectations.

Our third contribution is to show platform-supported auditing need not risk user privacy nor disclosure of platforms' business interests. In §4, we propose an auditing framework that protects against these risks. Our framework uses the rigorous definition of Differential Privacy (DP) to protect private information about *audit participants* that may leak to the *auditor*. It also protects the platform by not exposing details of the ranking algorithm—the platform shares with the auditor only the privatized scores of the relevance estimator, not proprietary source code, models, training data or weights. At a high level (Figure 2), an auditor queries the algorithm with a trial content and a list of users whose sensitive demographic attributes are known to the auditor. The platform then calculates how relevant the content is to each user, applies a differentially private mechanism to protect information that the relevance scores may leak about the users, and returns a distribution of noisy scores to the auditor. Finally, the auditor uses the noisy scores and an applicable fairness metric to test for disparity between the distributions of relevance scores the algorithm assigns to different demographic groups. The auditor chooses the specific type of content, attribute of users and metric of fairness to use depending on the specific scenario and the type of bias or harm they are studying.

We show that the privacy guarantees in our framework increase the number of samples required for an accurate audit by only a small constant factor (§5). We theoretically analyze the trade-off between guaranteeing privacy and the minimum sample size required for auditing in one concrete scenario – bias in delivery of employment ads. For the specific fairness metric we studied, we find that the noise that the platform adds to guarantee DP increases the required sample size by only approximately a factor of 4 for reasonable auditing parameters, and with a strict upper bound of 6.34. Our contribution is application of standard DP algorithm to ad relevance estimators, a new application where DP enables viable privacy-utility trade-offs.

Overall, our technical contributions show a path exists from the proposed legislation to a realizable auditing system. While full implementation of our framework is future work and will require collaboration with a platform, conceptually demonstrating how to enable public oversight while protecting privacy is an important step forward. We summarize the limitations of our framework in §4.4, but as the first proposed solution for implementing DSA- and PATA-like laws, it provides a useful starting point for exploring a new solution space.

2 THE NEED FOR A GENERALIZABLE AUDITING FRAMEWORK

We discuss recent developments in policies that are pushing to increase transparency of digital platforms, the need for safeguards to protect privacy, and the insufficiency of existing auditing methods to practically implement these policies at scale.

2.1 Policy Pushes to Increase Transparency while Ensuring Privacy

As social media platforms increasingly shape economic, social and political discourse, new policies are being proposed to regulate them. We discuss two prominent pieces of legislation to mandate independent oversight and transparency research on platforms: Platform Accountability and Transparency Act (PATA [66, 83], proposed in the US) and Digital Services Act (DSA [25], to be enforced in the EU starting in 2024).

PATA was proposed in December 2021 to mandate platform support for independent research on algorithmic transparency [83]. The proposal covers all large platforms with at least 25 million unique monthly users. It mandates the platforms make data available to "qualified researchers" who will study how platforms negatively impact individuals and society. Only researchers and projects vetted and approved by the National Science Foundation (NSF) will be allowed to access platforms' data.

DSA was proposed in the EU in December 2020 to regulate digital platforms and services [25]. DSA covers a broader set of entities beyond large social media platforms, including online marketplaces and app stores. While DSA has a broader in scope than PATA, it similarly mandates platforms to allow scrutiny of their algorithms by "vetted researchers" (Article 31 [25]). DSA was approved and passed as a law in July 2022 [26, 27].

Both PATA and DSA recognize the need to ensure safeguards to protect privacy during platform auditing. PATA emphasizes user privacy, with the necessity to "establish reasonable privacy and cybersecurity safeguards" for user data, and to ensure the data platforms provide is "proportionate to the needs of the [...] researchers to complete the qualified research project" [83]. DSA acknowledges platform's desire for "protection of confidential information, in particular trade secrets" [25] when conducting audits. To mitigate the risks to users and platforms, both proposals require vetting auditors, their projects, and results before they are published. Prior to our work, no actionable technical proposals put forth methods to implement such auditor access while protecting users' privacy and platforms' proprietary algorithms.

Platforms themselves also often cite their need to protect user privacy as a handicap for their transparency and self-policing capabilities [1, 6]. For example, Facebook has argued laws such as the EU's GDPR constrain their efforts for making data available to researchers [85]. In line with this argument, Facebook has constrained transparency efforts through actions such as providing data without sufficient granularity and accuracy needed to conduct meaningful audits through its Ad Archive APIs [22, 65, 85], and shutting down accounts used for transparency research by NYU's Ad Observatory project [15]. In a partnership Facebook made with Social Science One, Facebook cited GDPR concerns and agreed to share data only using a differentially private mechanism [48]. Other social media platforms such as Twitter have also raised the challenges of sharing data for auditing for societal benefit while protecting the privacy of users [43].

Given policymakers' and platforms' concern about privacy, implementing PATA and DSA requires solving methodological challenges to increasing transparency while safeguarding the privacy of users. Our auditing framework (§4) suggests that these policy requirements can be made concrete and viable with our proposed methodology.

2.2 Existing External Auditing Methods are Insufficient

Until the present, societal and individual harms of social media algorithms have mostly been merely hypothesized or, in some cases, demonstrated by journalists and researchers through audits done independent of the platforms.

However, such fully external auditing methods are reaching hard limits in terms of what they can reliably and provably learn about the optimization algorithms' role; increasing public-interest researchers' calls for legislation and other transparency sources that can support their efforts [2, 3, 17, 21, 44, 65, 74]. Specifically, the fully external auditing methods face fundamental challenges accounting for confounding variables and using proxies for sensitive attributes of interest. As a result, they are difficult to generalize and have high cost. In addition, they are susceptible to platform interference (§2.3). We next expand on these challenges.

Confounding variables: The first challenge is controlling for temporal variables that confound measurements. These confounding factors are present because platforms' algorithms operate in an environment that is influenced by actions of both users and the algorithms themselves. These hidden variables make it difficult to attribute measured effects to decisions made by platforms' algorithms.

Auditing for bias in ad delivery provides an illustration of the challenge of accounting for confounding factors. Several factors can affect the delivery of an ad, such as market competition from other ads, and differences in platform's use or interaction patterns among users from various demographics. An external auditor must control for such factors to attribute differences in outcomes to the decisions made by relevance estimators. Designing auditing methods with such controls in place, however, is a laborious process that requires careful reasoning and creative hacks. It took many years of research effort to get from Sweeney's study that gave the first evidence of biased ad delivery in 2013 [75] to Ali and Sapiezynski et al.'s study that attributed such bias to the role the platform's algorithms play in 2019 [2], and the Imana et al.'s 2021 study [44] that established that the algorithms are not merely biased, but, in fact, discriminatory.

Similar factors can confound measurements of potential harms in personalized organic content delivery. For example, a recent study on Twitter used sock-puppet accounts to compare their reverse-chronological and personalized timelines, and showed Twitter's algorithms distort information that users get exposed to [9]. However, the study's authors identify the duration sock-puppet accounts stay logged-in for and the timeline scrolling capabilities as potential confounding factors that could possibly alter the study's conclusions [9]. Even Twitter's internal audit of disparate algorithmic amplification of political content shows the limits of current methods [43]. The study showed that their metric of amplification, which is based on number of impressions, demonstrates the presence of bias on Twitter, but that confounding factors prevent any conclusions about potential sources of this bias.

These examples demonstrate the limits of impression-based measurements for isolating algorithmic effects. To increase transparency beyond what we have already learned through existing external auditing methods, a new level of access is needed for auditors (§3).

Reliance on proxies: A second challenge is the need for an auditor to use proxies for demographic attributes that platforms do not collect or report. Auditors may be interested in studying the impact of a specific demographic feature on algorithmic personalization, but often conduct external audits by posing as a regular user or advertiser. Operating as a normal user or advertiser is relatively easy and allows audits without a platform support or knowledge, but it also means the auditor can only use data points that a platform makes available to any user. For example, in the context of ad delivery, some platforms may not report ad impression rates broken down by attributes such as gender, race or political affiliation. Past audits have worked around this challenge by using proxies for demographic attributes that platforms do not report [2, 3, 44]. However, such workarounds introduce measurement errors [44] and significantly limit the ability to vary the attributes.

Lack of generalizability: Another challenge is that existing external auditing methods are often not generalizable beyond the limited context which they were originally designed for. For example, the study described above that built upon Ali and Sapiezynski et al.'s work added new controls for job qualifications across genders [44]. The focus on job ads and gender makes the method not easily generalizable to other types of ads (Example: insurance ads) and demographic attributes (Example: race). This lack of generalizability is also directly related to the limitations of confounding variables and use of proxies discussed above. In order to work around these limitations, researchers often use one-off hacks that are experiment- or platform-specific. Examples include use of random phone numbers to generate a random custom ad audience [2], and use of public data sources such as voter data to build audiences with a specific demographic make up [2, 3, 74]. Such public data sources are extremely limited and subject individuals to participation in experiments without their knowledge.

Crowdsourced audits that rely on browser extensions do not generalize beyond desktop versions of platforms, a limitation given that the use of most social media is now primarily mobile-first. For example, 98.3% of Facebook users access the site using a phone app [46]. Furthermore, such extensions need to be customized for each platform, and need to be regularly maintained to adapt to changes on platforms' websites.

Cost of auditing: Existing external auditing methods can also incur high cost in terms of time and money. For ad delivery, the state-of-the-art method for auditing involves registering as an advertiser, running real ads, and measuring how they are delivered in real-time while controlling for hidden temporal factors [2, 44]. The monetary cost for this procedure can easily accumulate with repeated assessments of a platform to confirm results over time, increase statistical confidence, or vary study parameters. In addition, controlling for confounding factors and proxies for measuring delivery along sensitive attributes requires time for study design.

For studies of personalization of organic content, creation of sock-puppet accounts is expensive because it often requires separate hardware and phone number verification, and it takes time and effort to make a sock-puppet's account activity "realistic".

These challenges motivate our approach: by using platform-supported auditing centered on relevance estimators, we directly focus on platform choices, side-stepping confounding variables and proxies. Explicit platform support also avoids platform interference and minimizes cost, provided platforms collaborate, as we explore next.

2.3 Platforms Beginning to Favor Platform-supported Audits?

Platform-supported audits, of course, require support from the platform, so we next look at evolution of the platforms' responses to requests for auditing.

Pushback from platforms: Traditionally, a major challenge for external auditing methods has been pushback from platforms, often citing privacy concerns or violation of their terms of service.

External audits collect data either through interfaces the platforms provide or by using tools such as customized scrapers and browser extensions. Regular website changes complicate long-term maintenance of tools that track platforms [9]. Facebook has resisted external auditing by explicitly blocking accounts used to conduct audits [15], tweaking its APIs to break auditing tools [57], and threatening legal actions against researchers who scrape data from its platform [41].

A change of heart? Recently platforms have released data or provided APIs to researchers, suggesting platforms themselves may be interested in some form of platform-supported auditing. Platform support allows them to manage auditing, and perhaps preempt adversarial black-box audits, lawsuits, and explicit regulation.

In a historic settlement with the US Department of Justice (DOJ), Facebook announced in June 2022 that it will work towards de-biasing its algorithms used for delivering job, housing and credit ads [7, 77]. The settlement requires Facebook to work with a vetted external entity to verify the changes implemented to its algorithms are compliant with the non-discrimination goals set by the settlement, a compliance structure similar to platform-supported methods proposed in PATA and DSA.

Platforms are also establishing programs to provide vetted researchers with access to their data and algorithms. In July 2022, YouTube announced "YouTube Researcher Program" (YRP), which promises to provide academic researchers with data, tools, and support for studying YouTube and its impact [89]. A year earlier, Facebook announced the "Facebook Open Research & Transparency" (FORT) initiative, that provides privacy-protected datasets and APIs for researchers, so that "the public can learn more about Facebook's impact on the world from credible and independent academic sources" [30]. In 2021, Twitter held an algorithmic bias bounty challenge, through which Twitter made the code for their image salience algorithm suspected of bias [88] available to researchers [14].

These steps are promising responses that approach legislative requirements, suggesting platforms are considering explicit support of methods that increase transparency of their influence on individuals and society.

However, for both YRP and FORT, current data available to researchers through the APIs is limited to public data corpora, such as public videos, pages, posts and comments [30, 89]. While such access is an important first step for helping understanding how the platforms shape public discourse, we argue in §3 that it is also important for platforms to provide a means to studying their relevance estimator algorithms. We hope our work encourages platforms to expand these first efforts to allow researchers to study how their algorithms shape access to content.

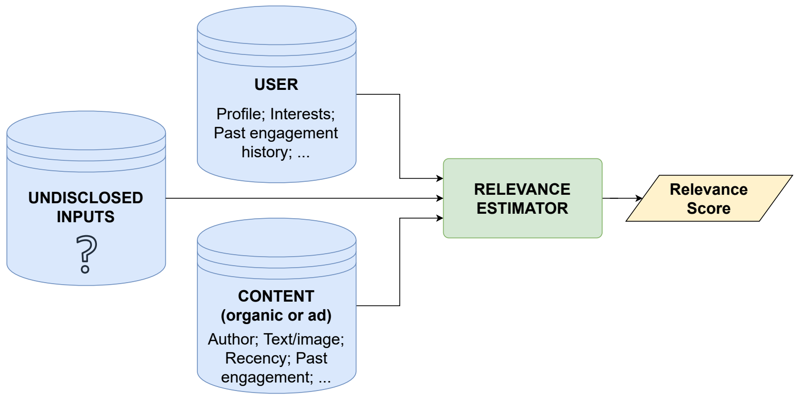

Fig. 1. Social media platforms use relevance estimators to score every content using many factors as input, only some of which are publicly documented.

3 RELEVANCE ESTIMATORS ARE THE KEY TO INCREASING TRANSPARENCY

In this section, we show transparency of relevance estimators is the key to enabling generalizable auditing of platforms for potential harms. To support this claim, we first document the importance of relevance estimators for content prioritization. We then survey studies that have shown harmful outcomes that result from use of these algorithms.

3.1 Relevance Estimators: "Brains" of Social Media Platforms

Relevance estimators are the main drivers that shape every piece of content shown to users. Prior work and platforms' documentation show this importance and that these algorithms are opaque to external auditors.

Given the vast amount of potential content shared on social media, relevance estimators have become responsible for selecting which content is shown on a user's timeline and in what order, and which is omitted or deprioritized. For example, for organic content, it maybe selecting and ranking posts from a user's friends or pages they follow, or for promoted content, it may be running auctions for ads that are currently competing for a particular user's attention. Platforms may mix both organic posts and ads in the content stream. Facebook's algorithmic newsfeed dates back to 2007 [60], and Twitter and Instagram deployed such personalization in 2016 [47]. Before deploying these algorithms, the platforms used ordering that was mostly reverse chronological.

For organic content, these algorithms ultimately boil down to *relevance scores* that will determine the selection and order of content shown at the top of users' news-feed. For example, Facebook makes a number of predictions about how likely a user is to engage with posts, and will "add these predictions up into a *relevancy score*" to order the posts [28]. Instagram follows a similar approach [61]. Similarly, Twitter uses a number of "consolidated signals to develop a *relevancy score*" [49] that it uses to determine which tweets to show on top of the feed. LinkedIn also uses an algorithm that "*scores* tens of thousands of posts and ranks the most *relevant* at the top of the feed" [59]. TikTok mixes content from both followed accounts and others using algorithms that optimize for "effective relevance" as a "secret sauce" [80].

Relevance estimators are used in ad auctions in addition to organic content. Ads are selected using auctions; these auctions consider relevance as a factor predictive of user engagement. These predictions are combined with other factors, such as the bid and budget the advertiser set for the ad, to determine the auction winner. Different platforms use different terminologies to refer to these predictions. For example, Facebook, LinkedIn and Twitter refer to them as "Estimated action rates" [29], "Relevancy scores" [53], and "Quality scores" [82], respectively. But they all have very similar purposes in that they are applied as modifiers to bids to determine which ad wins an auction. Therefore, an ad with the highest bid may not win an auction if it is given a low relevance score by the algorithmic prediction.

Platforms provide little transparency into optimization algorithms, neither for organic nor promoted content. As summarized in Figure 1, publicly available documentation gives a high-level description that platforms use information about the content itself, the author of the content, and user's profile data. However, the specific types of algorithms and inputs to those algorithms are not disclosed. Facebook, Instagram and LinkedIn use thousands of factors to estimate relevance of posts [28, 59, 61]. Similarly, Twitter's documentation shows they use advanced machine learning algorithms to predict relevance, where the "list of considered features and their varied interactions keeps growing" [49].

The importance of relevance estimators to organic content and ad delivery lead us to place them at the center of our mechanism for external auditing in §4.

3.2 Relevance Estimators Can Cause Various Forms of Algorithmic Harms

Prior studies by researchers and journalists have raised concerns about how relevance estimators on social media platforms can lead to potentially harmful outcomes. We next show examples of audits where controlled experiments show biased or harmful outcomes that result from platform choices.

Ad delivery is an area where there is substantial evidence for relevance optimization resulting in bias and discrimination. Starting with Sweeney's empirical study in 2013, researchers hypothesized that platform-driven choices result in discriminatory ad delivery across demographic groups [75]. In 2019, this hypothesis was confirmed by Ali and Sapiezynski et al. by showing Facebook's relevance algorithms skewed delivery of job and housing ads by gender and race, even when an advertiser targets a gender- and race- balanced audience [2]. A subsequent study by Imana et al. controlled for job qualifications, a legally excusable source of skew, to demonstrate the role of Facebook's relevance algorithms may be violating U.S. anti-discrimination laws [44]. These examples provide evidence that opaque optimization algorithms result in discriminatory delivery of opportunity ads for certain demographic groups.

Facebook's academic work [19] and Facebook's public statements [7] in response to the recent settlement with the US Department of Justice [77] both also acknowledge the need to ensure its algorithms for opportunity ads are not biased. Both of these serve as additional evidence that harms of relevance estimators that prior studies pointed out are well grounded.

Delivery of organic content is another area where past audits have found evidence for bias. Twitter conducted an internal audit on its algorithms used to curate timelines, and found that its platform amplifies right-leaning political tweets more than moderate ones [43]. The study suggests the difference may be attributable to Twitter's ranking models assigning higher relevance scores to the right-leaning tweets. Another external audit on Twitter by Bartley et al. also did a more general comparison of algorithmic and reverse chronological timelines on Twitter and showed the algorithmic timeline distorts information that users are shown [9].

Besides such internal and external audits, investigations done by journalists have corroborated that the potential harms of algorithms used by platforms are not merely theoretical. A recent prominent example is "the Facebook files", an investigation done by Wall Street Journal into leaked internal Facebook documents. Among other findings, the investigation showed how changes in 2018 to make "the platform healthier" by focusing on relevance and engagement caused its algorithms

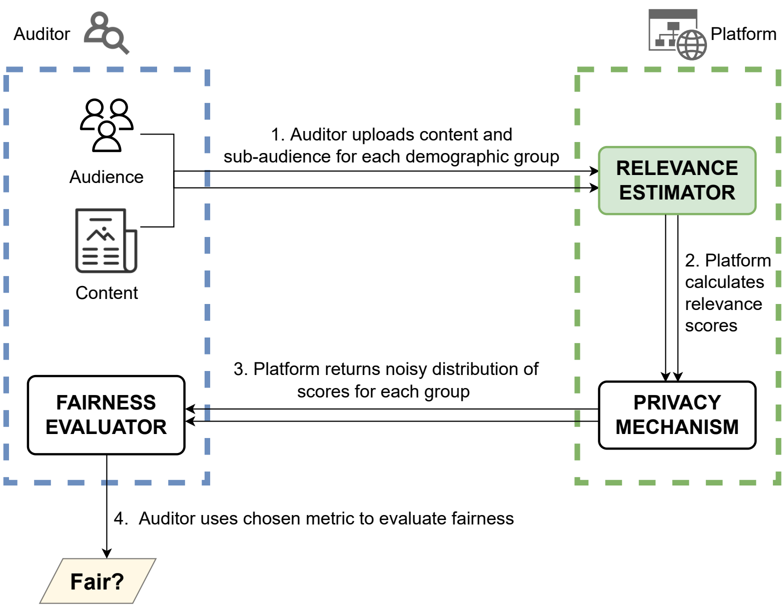

Fig. 2. Privacy-preserving platform-supported auditing framework for evaluating fairness of relevance estimators.

to promote objectionable content [78]. They report cases where Facebook's algorithms have led teenagers to harmful content [32] and spread hateful posts [67].

These findings underscore the need for public oversight. The goal of such oversight will be to ensure relevance estimators that optimize for business objectives take societal interests into account. Platform-supported auditing is an important part of confirming progress towards this goal.

4 PRIVACY-PRESERVING PLATFORM-SUPPORTED AUDITING

We next describe our approach to platform-supported auditing and how it addresses risks to the privacy of users and business interests of platforms.

4.1 Overview and Context

Our proposal for *platform-supported auditing*, shown in Figure 2, has four high-level steps: (1) an auditor selects a trial content and an audience whose demographic attributes are known to the auditor, and uploads the content and sub-audience for each demographic group separately; (2) for each group, the platform calculates relevance scores that estimate how relevant the content is to each user in the group; (3) the platform then applies a privacy mechanism and returns to the auditor a noisy distribution of the scores for each demographic group; (4) finally, the auditor evaluates the fairness of the scores assigned to different demographic groups using an applicable metric of fairness. We discuss each of these steps in more detail in §4.3.

Our platform-supported auditing framework is a new category that extends the classical Sandvig taxonomy of auditing [71]. Their taxonomy defines five types of audits: source code audit, survey-based audit, scraping audit, and crowdsourced audit. Other types of audits recently proposed include *everyday algorithm audits* [73] and *cooperative audits* [87]. Our platform-supported auditing defines a new type of audit because, unlike methods that use only features or APIs publicly available to regular users and advertisers, it requires a privileged and auditor-specific interface. And unlike methods that require access to platforms' proprietary source code, it only requires query-level access to the output of their algorithms. We further discuss how platform-supported auditing compares to the existing taxonomy of methods in §6.

4.2 Privacy and Business Risks of Platform-Supported Auditing

Our approach is designed to minimize risks to the privacy of platform users and to platform's proprietary information. As discussed in §2.1, protecting against these risks is an important goal of PATA and DSA, and is also a concern that platforms identify as a constraint to enabling transparency and auditability. We next discuss the potential risks of providing query access to relevance estimators and the need for ensuring rigorous privacy protection when their outputs are shared.

Relevance scores may leak private user data based on which they are calculated. As discussed in §3.1, platforms calculate relevance scores based on users' personal profile data and their historical engagements with the platform. The relevance of each particular content to each user may reveal information about the user that the auditor otherwise would not know. For example, when a platform finds content about disability support or insurance highly relevant to a given user, that result suggests the user may be disabled or is caring for a disabled person. Similar real-life examples from other contexts include Target's predictive algorithms for sending relevant coupons leaking a teenager's pregnancy [37], and Facebook's "People You May Know" feature that suggest connections Facebook deems relevant revealing private information about a user [38]. Even if relevance scores are aggregated in some fashion, prior work has shown similar aggregate outputs of personalization systems, combined with auxiliary information about users can leak private information [10, 13, 86]. Therefore, our auditing method must limit the potential to make such inferences.

In §4.3.3, we show how our framework protects the privacy of users, when privacy protection is defined as ensuring a Differential Privacy (DP) guarantee on any *data that is shared with the auditor*. Protection from other privacy violations, such as how the platform itself may violate privacy of users through training and deployment of personalization algorithms, is outside the scope of our framework. DP is the current gold standard for protecting privacy of individuals, while providing useful statistical computations on their data [20]. DP provides a rigorous guarantee that each individual subject's participation in the audit has negligible impact on their privacy. A differentially-private mechanism will specifically protect *the privacy of the users participating in the audit*, while providing aggregate information about the relevant estimator, which the auditor can use to asses fairness.

In addition to risks to platform users, platforms themselves would like to minimize what details of their algorithms they share. Our framework minimizes information it requires the platforms to share about their algorithms and data by providing only query (rather than source-code) access to auditors, asking to share only aggregate relevance metrics, while preserving the confidentiality of the source code, how those metrics are computed and what inputs and training data they use.

4.3 Steps of Platform-Supported Auditing

We next describe each of the four steps of platform-supported auditing in detail.

4.3.1 Auditor Uploads Content and Audience. The auditor first will select a trial content and a customized audience. The content is specific to the platform under study. For example, a job ad for LinkedIn or a political Tweet for Twitter. The audience is a list of users whose demographic attributes are known to the auditor. The auditor selects the content and demographics based on the specific platform and the type of algorithmic harm they are studying. The auditor then uploads the content and a sub-audience for each demographic group to the platform. Major platforms already have an infrastructure for advertisers to upload audience and content which, with some modifications, can be used for auditing purposes.

The type of content and the demographic make up of the users depends on what type of harm or bias the auditor is interested in studying. For example, one may wish to study whether LinkedIn's ad delivery algorithms deliver STEM job ads in a biased way that reflects historical under-representation of women in the field. To perform the study, the auditor may use a STEM job ad as the content, and a sample of men and women as the audience. The auditor will query the platform using the audience from each demographic group to evaluate whether the platforms' relevance estimators assign higher scores to men compared to women.

The auditor will specify the audience by uploading a custom list of specific people whose demographic attributes are known to the auditor. One way the auditor can build such an audience is by recruiting volunteers for the study. Recruiting volunteers has two main advantages over prior methods that use publicly available datasets, such as voter data [2, 3, 44]. First, each participant gets the opportunity to consent the use of their data for auditing purposes. Second, participants can provide additional attributes that may not be present in public data but could be useful for auditing. An example is a job qualification attribute that is useful for auditing delivery of employment ads. Existing recent studies such as The Markup's Citizen Browser [76] and Mozilla's Rally [62] show that users are willing to opt-in and provide data to reputable efforts that aim to hold platforms accountable. In addition to such external efforts, platforms themselves can provide support for recruiting users [1, 54].

Having the auditor specify a custom audience is also advantageous over letting the platform itself pick an audience. It helps protect the privacy of users since it does not require platforms, nor gives them the excuse, to collect sensitive demographic attributes. This advantage also addresses the challenges around collecting and securing sensitive attributes of users that companies often identify as one of the major obstacles to auditing for fairness [1, 4, 11].

Popular social media platforms have an infrastructure for advertisers to upload custom audiences. These existing features are currently designed for use by advertisers to run advertisement campaigns that retarget their customers. An advertiser may upload information such as names, email addresses and phone numbers of their customers and the platform tries to match this information with user profiles on the platform.

These existing features can serve as a starting point for platforms to build a similar interface that can be used for platform-supported auditing. In the existing custom-audience features, not all people in the audience may match to user profiles on the platform. In a prior work, we have found that such partial matches can be a source of error when using custom audiences for auditing [44]. A possible modification for supporting accurate auditing is for the platform to allow the auditor to upload unique identifiers (for example: Facebook usernames) for the accounts of people that are participating in the audit. The auditor can collect these identifiers when recruiting volunteers who will participate in the study.

4.3.2 Platform Calculates Relevance Scores. The platform then calculates how relevant the content is to each user in the custom audience. Relevance estimation on the platform boils down to a *relevance score* for each user, which is the platform's prediction of how likely the user is to engage with the content. The platform will not report the raw scores to the auditor as they may reveal private information about its users' past engagement history. Instead, the platform builds statistics that summarizes the distribution of the scores (For example, a histogram or a CDF), and adds privacy protections (discussed in §4.3.3), before returning the statistics to the auditor.

Platforms use many factors for estimating relevance but not all are applicable for auditing context. For example, during a normal usage of the platform, the estimators may use as inputs factors such as what time of the day it is and for how long a user has been logged in [28]. For such temporal variables that are only applicable in the context of a user browsing the site, the platform must keep them constant for all users in the audience. This control allows the auditor to evaluate bias in the relevance estimators that may arise from the historical data that the platform has about users and not from temporal factors.

4.3.3 Platform Applies Privacy Mechanism and Returns DP-protected Scores. The platform then applies a differentially private mechanism to the statistics of relevance scores calculated and returns the noisy statistics to the auditor. The mechanism will provide a rigorous guarantee that the data the auditor gets ensures differential privacy for individuals participating in the audit. We use the following definition of DP, where neighboring databases are defined as differing in one person u's data: $D_1 = D$ and $D_2 = D \cup \{u\}$ for some database D.

Definition 4.1 (ϵ-Differential Privacy[20])**.** Given $\epsilon > 0$, a randomized mechanism M is ϵ-differentially private if for any two neighboring databases D_1 and D_2, and for any subset $S \in R$ of outputs,

$$Pr[M(D_1) \in S] \leq e^{\epsilon} * Pr[M(D_2) \in S],$$

where the probability is taken over the random coin tosses of M.

Auditors can approximate tests for group-fairness metrics using a binned histogram of relevance scores without access to individual scores. One method to share the binned histogram while preserving privacy is using the Laplace Mechanism [20]. The platform can independently add noise drawn from the Laplace distribution to each of the bins in the histogram. Since presence or absence of a single user changes each bin's count by at most one, adding noise from Laplace distribution $Lap(1/\epsilon)$ independently to each bin ensures the mechanism is $\epsilon-$differentially private [20]. The platform then returns the noisy histogram counts back to the auditor.

Any arbitrary post-processing to output of a differentially private mechanism does not reverse the privacy protection. Therefore, auditor can use the noisy scores to apply any post-processing computations to test for fairness without incurring additional privacy cost to the platform or its users.

4.3.4 Auditor Evaluates Fairness of Relevance Scores. Finally, the auditor uses the noisy distribution of scores to test whether there is a disparity between the relevance scores the algorithm assigns to different demographic groups. The specific metric of fairness depends on the type of algorithmic bias the auditor is interested in testing for. For example, to study bias in the delivery of employment ads, the auditor may use Equality of Opportunity as a metric for fairness since the definition takes qualification of people into account, which is a relevant factor for the context of employment [34]. We further explore this scenario in our theoretical result in §5.

4.4 Trust Model and Limitations

We next discuss the trust model we use to evaluate the privacy and business interest risks of our approach. The efficiency our approach assumes a legal framework in which both the platforms and auditors work in good faith, or potential tests for non-cooperation.

Platforms: One major assumption of our framework is that the platform will truthfully collaborate with auditors and ensure audits are done accurately and effectively. The platform must provide auditors access to the same algorithms that are used in production, truthfully executing them on the audience the auditors upload and reporting relevance scores accurately (modulo privacy modifications). This assumption was not stated in prior auditing methods that do not use a platform's support. Even for such methods, platforms have the means to know they are being audited as the audiences and methodologies auditors typically used are publicly documented. Examples included

North Carolina's voter datasets used as data source for demographic attributes [74], Facebook ad accounts used to audit ad delivery [2, 3, 44], and browser extensions used for collecting data from Facebook [64, 76].

Assuming the platform truthfully collaborates with auditors is a strong assumption, but there are four reasons we think it is appropriate. First, the consequences of non-compliance are significant when auditing is part of an official legal framework, as it would be in the context of a DSA- or PATA-like law or a legal settlement, such as Facebook's settlement with the US Department of Justice [7]. For example, Volkswagen faced significant legal and financial repercussions as a result of their violation of emissions regulations [45].

Second, platforms also have the incentive to minimize inadvertent errors in order to avoid tarnishing their public image and potential legal liability. Two cases, both involving Facebook, serve as an example of this. In the first case, Facebook made inadvertent errors in sharing data to external researchers as part of its Social Science One program [81]. This preventable error undermined academic work that was based on the data [81], tarnishing Facebook's efforts to be a leader in increasing transparency. In the second case, Facebook mistakenly inflated potential reach estimates for ads, and is currently being sued as a result [33].

Third, simply formalizing auditing and involving two parties often adds sufficient oversight to discourage abuse. For example, corporate financial accounting is not immune to fraud, but the levels of non-compliance are small enough that it is a very useful and powerful tool.

Finally, as discussed in §2.3, there is evidence that the platforms themselves may be moving towards supporting audits through giving external researchers privileged access to their data and algorithms.

Auditors: The platform must also trust researchers doing the independent audit. One risk for abuse is misuse of the auditing interface to harm a platform's business. Both the DSA and PATA provide rules to ensure only vetted researchers will be allowed to perform audits on social media platforms [25, 83]. In both proposals, an assigned regulatory body will screen researchers and their projects before they are allowed to audit a platform's system or data [25, 83]. Platform-initiated transparency efforts such as Facebook's FORT, Social Science One, and YouTube's Researcher Program also all have approaches for vetting researchers [30, 48, 89]. Such screening processes will minimize the risk that comes from malicious auditors, and the platforms' implementations show that the platforms themselves believe this risk can be overcome.

Another risk is misuse of sensitive data that auditors collect from users who are participating in an audit. Similar to the risk for platforms discussed above, having only vetted researchers conduct audits helps reduce the risk for users. In addition, under our proposed framework, users would be voluntarily providing their data, unlike prior methods that used public voter data. In these prior methods, users were not even aware their data was being used for experiments.

5 SAMPLE SIZE REQUIRED FOR AUDITING RELEVANCE ESTIMATORS WITH PRIVACY

We next present the key technical result of this paper by applying our framework to one use-case: a study of discrimination in employment ad delivery. We show that the addition of differential privacy to the auditing pipeline does not prevent an auditor from achieving the same statistical confidence as without privacy protections, provided the sample audience is increased by a small constant factor. This result supports our claim that it is feasible to both audit for fairness *and* protect user privacy and platforms' business interests.

5.1 Setup and Assumptions: Bias in Delivery of Employment Ads

Auditing social media platforms for fairness while preserving privacy is a goal that desirable in multiple scenarios. We study one scenario: assessing discrimination in delivery of employment ads. Our problem formulation is general, although specific scenarios place additional requirements, like the role of qualifications in job advertisements. Extending our approach to other types of ads may require identifying similar factors reflecting allowable preferences.

We consider the case where an auditor wishes to confirm delivery of job ads is unbiased relative to a factor such as gender or race. To evaluate this question, the auditor will examine a platform's relevance estimator for different groups with specific demographic attributes. This scenario is motivated by prior third-party audits that have indirectly measured the role of relevance optimization in biased job ad delivery [2, 44].

We first introduce formal notations for the scenario. Let X represent a set of all users on a platform and let A be the range of values for a sensitive attribute (For example, $A = \{\text{black, white, ...}\}$ for race). Let $Q = \{0, 1\}$ represent binary options for qualification of a user to a given job ad (1 if the user is qualified, 0 – otherwise). Let $R_j(x)$ be the relevance estimator that calculates the relevance score of the job ad j to a given user $x \in X$. We assume a specific ad j and omit the subscript j throughout. And let Y be a small finite set of discrete relevance scores (we describe how to extend Y to the continuous case at the end of this section).

In practice, the external auditor cannot have access to a complete list of all of the platform's users (X), so the auditor recruits a sample (S) of users to perform the audit. The auditor uses a random sample set $S = \{(x_1, a_1, q_1), (x_2, a_2, q_2), ..., (x_n, a_n, q_n)\}$ drawn i.i.d. from X. In that case each subset $S_{a,q}$ is also i.i.d. in $X_{a,q}$, where $S_{a,q}$ and $X_{a,q}$ represent subsets with given values of a and q. We discuss implications of this assumption at the end of this section.

Following the steps in Figure 2, the auditor first queries the platform's relevance estimator using each subset S_a and j (step 1). The platform then applies R to every user in S_a (step 2) and builds a histogram H of the scores, grouped by possible range of relevance scores in Y. It then independently adds noise drawn with a Laplacian distribution $Lap(\frac{1}{\epsilon})$ to each of the bins in H, where ϵ represents the level of differential privacy desired. The platform returns the noisy histogram counts back to the auditor (step 3).

Finally, the auditor tests for fairness of the scores assigned using Equality of Opportunity as a definition of fairness (step 4). Equality of Opportunity is an established fairness notion in the algorithmic fairness literature, and is applicable to job ads as it allows for taking into account the qualification of users [34].

Definition 5.1 (derived from Equality of Opportunity [34]). A relevance estimator function R satisfies equality of opportunity:

$$Pr_{(x,a',q)\sim X}[R(x) = y | a' = a \wedge q = 1]$$
$$= Pr_{(x,a',q)\sim X}[R(x) = y | q = 1] \text{ for all } a \in A \text{ and } y \in Y,$$

where the probability is taken over the choices of samples from X and the random coin tosses of R.

We modify Hardt et al.'s formulation by using the group of qualified people ($q = 1$) to represent the "advantaged outcome" group [34]. The advantaged outcome in our case is that a person sees a job ad because they are qualified for the job. In addition, in our formulation, the outcome space Y is not binary but a finite set of discrete values.

To test for this metric, the auditor must know whether each user is qualified for the job being advertised. For convenience, we introduce the following notation:

$$P_{a,y}(R) = Pr_{(x,a',q)\sim X}\,[R(x) = y | a' = a \wedge q = 1] \tag{1}$$

$P_{a,y}(R)$ represents the likelihood that a qualified individual from a specific demographic group a receives a relevance score y. The auditor expects this likelihood to be equal across demographic groups if the platform's algorithm is unbiased.

We relax strict equality of the above term since any real-world observation may have small noise or variation. We will use a relaxation from prior work [72], that allows a small additive error α as maximum allowed fairness gap (FG) between any two demographic groups. We change the relaxation to use α instead of ϵ because we use ϵ as a privacy parameter.

Definition 5.2 (α-fairness [72])**.** We define a relevance estimator function R to be α-fair if:

$$FG(R) = \max_{a_1,a_2 \in A, y \in Y} |P_{a_1,y}(R) - P_{a_2,y}(R)| \leq \alpha$$

Since the auditor has only access to an independent sample of users (S), the measure of $P_{a,y}$ the auditor gets empirically is given by:

$$\overline{P}_{a,y}(R,S) = \frac{1}{n_{a,q}} \sum_{i=1}^{|S|} \mathbb{1}\{R(x_i) = y \wedge a_i = a \wedge q_i = 1\} \tag{2}$$

where $\mathbb{1}\{.\}$ is an indicator function selecting qualified members from group a that are assigned a score y, and $n_{a,q}$ is the number of qualified members in S from group a. The equation requires that $n_{a,q} > 0$, an assumption we discuss at the end of this section.

Let $n_{a,q,y}$ be the number of qualified people in S from group a that got assigned a score y. We can also rewrite $\overline{P}_{a,y}(R,S)$ as:

$$\overline{P}_{a,y}(R,S) = \frac{n_{a,q,y}}{n_{a,q}} \tag{3}$$

We next consider the value of $\overline{P}_{a,y}(R,S)$ after it is distorted by noise to preserve privacy. From Equation 3, $n_{a,q}$ is already known to the auditor so the quantity the platform wishes to protect is $n_{a,q,y}$, which represents each bin in the histogram that the platform computes. The platform applies Laplace mechanism by adding noise drawn from $r \sim Lap(\frac{1}{\epsilon})$ to each count $n_{a,y,q}$ to guarantee ϵ-DP [20]. Let $P^*_{a,y}(R,S)$ represent the noisy value the platform calculates:

$$P^*_{a,y}(R,S,\epsilon) = \frac{n_{a,q,y} + r}{n_{a,q}} = \overline{P}_{a,y}(R,S) + \frac{r}{n_{a,q}} \quad \text{s.t.} \quad r \sim Lap(\frac{1}{\epsilon}) \tag{4}$$

Extending a formulation in prior work [72] by adding a new privacy parameter, the empirical fairness gap (EFG) is given below (we give both the private and non-private cases). A large EFG between two demographic groups implies unfairness.

$$EFG(R,S) = \max_{a_1,a_2 \in A, y \in Y} |\overline{P}_{a_1,y}(R,S) - \overline{P}_{a_2,y}(R,S)|$$

$$EFG(R,S,\epsilon) = \max_{a_1,a_2 \in A, y \in Y} |P^*_{a_1,y}(R,S,\epsilon) - P^*_{a_2,y}(R,S,\epsilon)|$$

The auditor checks $EFG(R,S,\epsilon) \leq \alpha$ to test whether a relevance estimator R is fair. To analyze the sample size needed to perform this test with high statistical confidence, we will use the following definition that allows a small δ probability of failure over the randomness in R and possible choices of samples in S.

Definition 5.3 ((α, δ)-fairness [72])**.** We define R to be (α, δ)-fair with high probability with respect to S if:

$$Pr\,[EFG(R,S) \leq \alpha]$$

$$= Pr\left[\max_{a_1,a_2\in A, y\in Y} |\overline{P}_{a_1,y}(R,S) - \overline{P}_{a_2,y}(R,S)| \leq \alpha\right] > 1-\delta$$

We extend this definition for the case where an ϵ-DP mechanism is applied to outputs of R to protect privacy of users.

Definition 5.4 ((α, δ, ϵ)-fairness). We define R to be $(\alpha, \delta, \epsilon)$-fair with respect to S where an ϵ-DP mechanism is applied to outputs of R if:

$$\begin{aligned} &Pr\left[EFG(R,S,\epsilon) \leq \alpha\right] \\ &= Pr\left[\max_{a_1,a_2\in A, y\in Y} |P^*_{a_1,y}(R,S,\epsilon) - P^*_{a_2,y}(R,S,\epsilon)| \leq \alpha\right] > 1-\delta \end{aligned}$$

The formulation in this and the following sections assumes Y is a set of discrete values. Equation 1 and Equation 2 can be extended to the case where Y is a continuous space by choosing a different indicator function and comparing CDFs of relevance scores:

$$\begin{aligned} P_{a,y}(R) &= Pr_{(x,a',q)\sim X}\left[R(x) > y | a' = a \land q = 1\right] \\ \overline{P}_{a,y}(R,S) &= \frac{1}{n_{a,q}} \sum_{i=1}^{|S|} \mathbb{1}\{R(x_i) > y \land a_i = a \land q_i = 1\} \end{aligned} \tag{5}$$

Assumptions: Our approach makes several assumptions to avoid degenerate cases. We describe these next so that an auditor can design a robust experiment and may verify, post-audit, that the assumptions are met.

Equality of Opportunity (EoO) metric (Definition 5.1) adapts to unequal numbers of qualified individuals from different groups, but it cannot handle cases when *no* one or *very few* in the population with specific attributes are qualified for the job being advertised. The first degenerate case occurs when $n_{a,q} = 0$ in the denominator in Equation 2. Another case is when only a few individuals are qualified from one group, and very many individuals are qualified from a second group (Example: $n_{a_1,q} = 1$ and $n_{a_2,q} = 1$ million). In this case, EoO requires selecting all or none of the 1 million people in a_2 to match the inclusion or exclusion of the only individual in a_1. Our Theorem 1 guarantees that, for realistic parameters, $n_{a,q}$ is not small and that such degenerate cases do not occur. Moreover, the auditor may verify, post-audit, that the assumptions about $n_{a,q}$ were met.

Second, we assume samples in each demographic group are independent and identically distributed. We recognize that there maybe confounding factors that may induce bias, such as the location the audience is chosen from or difference in how active users are on the platform. The auditor can anticipate some of these factors and control for them but only the platform has the data to verify independence. In our result, we assume independence only within samples in a group, so we do not expect this limitation to decrease the observable differences in fairness across groups. This assumption is common in nearly all statistical studies, and is aimed to be achieved by following best practices in subject selection.

Third, we assume there is some way to randomly sample users. This mechanism may be provided by the platform, or the auditor may use some external source of users (in which case we require that will not induce its own bias). We recognize that sampling users from social media and encouraging them to share their data with the auditor may be difficult, but prior studies have met this requirement satisfactorily (for example, the work of Citizen Browser [76]). We therefore place this problem outside the scope of this paper.

5.2 Result: Minimum Sample Size Required for Auditing with Privacy

Building on the background in the prior section, we give the following theoretical result: we show that, for employment ad delivery use-case, auditing with differential privacy guarantee increases the number of samples required for auditing, but only by a small constant factor.

Theorem 1. An audit relying on a differentially privatized output of a relevance estimator R is $(\alpha, \delta, \epsilon)$-fair under equality-of-opportunity provided that, compared to the non-private case, an additional factor of S_{dp} samples are measured. We show that $4\ln(3)/\ln(2) = 6.34$ is an upper bound for S_{dp} and that 4 is a better estimate for S_{dp} under typical auditing parameters.

Formally, for an auditor to verify R is $(\alpha, \delta, \epsilon)$-fair with respect to a sample set S, assuming $\epsilon > \alpha/2$, the condition $EFG(R, S, \epsilon) \leq \alpha$ and the following condition on the minimum number of samples must hold:

$$\min_{a \in A} n_{a,q} \geq \frac{8}{\alpha^2} \ln \frac{3|A||Y|}{\delta} \tag{6}$$

where $S = \{(x_1, a_1, q_1), (x_2, a_2, q_2), ..., (x_n, a_n, q_n)\} \sim X$ and $n_{a,q}$ is the number of people in S with sensitive attribute $a \in A$ and who are qualified for the job being advertised. α and δ are knobs that control the level of fairness and statistical confidence, respectively.

To prove this theorem, we first show with the case of auditing relevance scores when a privacy mechanism is not used. We then analyze by what factor the required number of samples increases when a differentially private mechanism is applied.

Lemma 5.1. Without any guarantees of privacy, the following minimum number of samples is required to verify whether R is (α, δ)-fair with respect to a sample set S:

$$\min_{a \in A} n_{a,q} \geq \frac{2}{\alpha^2} \ln \frac{2|A||Y|}{\delta} \tag{7}$$

where $S = \{(x_1, a_1, q_1), (x_2, a_2, q_2), ..., (x_n, a_n, q_n)\} \sim X$ and $n_{a,q}$ is the number of people in S with sensitive attribute $a \in A$ and who are qualified for the job being advertised.

For the non-private case, the proof directly follows from prior work by Segal et al. on auditing machine learning models using cryptographic techniques [72]. In Appendix A, we extend their proof with consideration of qualification as an additional attribute.

We next consider sample size for the private case, where the auditor receives a noisy histogram of relevance scores because the platform applies a differentially-private mechanism.

Lemma 5.2. With privacy, the following minimum number of samples is needed to verify whether R is $(\alpha, \delta, \epsilon)$-fair with respect to a sample set S,

$$\min_{a \in A} n_{a,q} \geq \frac{8}{\alpha^2} \ln \frac{3|A||Y|}{\delta} \tag{8}$$

where $S = \{(x_1, a_1, q_1), (x_2, a_2, q_2), ..., (x_n, a_n, q_n)\} \sim X$ and $n_{a,q}$ is the number of people in S with the sensitive attribute $a \in A$ and are qualified for the job being advertised.

Proof. At a high level, the proof works by first defining a bad event that we want to happen with very low probability and then conditioning on this event not happening to derive the sample size needed to guarantee $(\alpha, \delta, \epsilon)$-fairness. The bad event is when there is error in the value for $P_{a,y}$ that the auditor calculates empirically. We have two sources of error: sampling error and error due to noise added to protect privacy.

Now, consider the following "bad" event where the error between the value the auditor calculates $P^*_{a,y}(R,S)$ and the true $P_{a,y}(R)$ is above some threshold $t > 0$:

$$\text{“Bad”} : \left|P^*_{a,y}(R,S) - P_{a,y}(R)\right| = \left|\left(\overline{P}_{a,y}(R,S) + \frac{r}{n_{a,q}}\right) - P_{a,y}(R)\right| > t$$

Conditioning on the event that the total error for the bad event does not exceed t, we get a lower bound for a sample size that satisfies (α,δ)-fairness using the following value of t (see Appendix A):

$$t = \frac{\alpha}{2} \tag{9}$$

We bound the probability of the above bad event for all groups in A and possible outputs in Y:

$$Pr\left[\exists a \in A \text{ and } y \in Y : \left|\left(\overline{P}_{a,y}(R,S) + \frac{r}{n_{a,q}}\right) - P_{a,y}(R)\right| > t\right] \leq \delta$$

By applying the triangle inequality, it is sufficient (but not necessary) to bound the probability that each of the two sources of errors exceed $t/2$:

$$Pr\left[\left|\overline{P}_{a,y}(R,S) - P_{a,y}(R)\right| > \frac{t}{2}\right] + Pr\left[\left|\frac{r}{n_{a,q}}\right| > \frac{t}{2}\right] \tag{10}$$

Since we require the samples in S are chosen i.i.d., $\overline{P}_{a,y}(R,S)$ is unbiased estimator of $P_{a,y}(R)$, i.e, $E[\overline{P}_{a,y}(R,S)] = P_{a,y}(R)$ (We prove this in Appendix C). Therefore, we can apply Hoeffding's inequality to the first term (sampling error) to simplify it to $2\exp(\frac{-n_{a,q}t^2}{2})$.

We then apply a known tail bound for the Laplace distribution (for $r \sim Lap(B) : Pr[|r| \geq t] < exp(\frac{-t}{B})$) to the second term (privacy error) to simplify it to $\exp(\frac{-n_{a,q}t\epsilon}{2})$. We then take a union bound over all possible values of a and y:

$$\begin{aligned}
&Pr[\exists a \in A \text{ and } y \in Y : \text{Bad event occurs}] \\
&\leq \sum_{a\in A}\sum_{y\in Y} Pr\left[\left|\left(\overline{P}_{a,y}(R,S) + \frac{r}{n_{a,q}}\right) - P_{a,y}(R)\right| > t\right] \\
&\leq \sum_{a\in A}\sum_{y\in Y} Pr\left[\left|\overline{P}_{a,y}(R,S) - P_{a,y}(R)\right| > \frac{t}{2}\right] + Pr\left[\left|\frac{r}{n_{a,q}}\right| > \frac{t}{2}\right] \\
&\leq \sum_{a\in A}\sum_{y\in Y} \left(2\exp(\frac{-n_{a,q}t^2}{2}) + \exp(\frac{-n_{a,q}t\epsilon}{2})\right) \\
&= \sum_{a\in A} |Y| \left(2\exp(\frac{-n_{a,q}t^2}{2}) + \exp(\frac{-n_{a,q}t\epsilon}{2})\right) \\
&\leq |A||Y| \left(2\exp(\frac{-n_{min}t^2}{2}) + \exp(\frac{-n_{min}t\epsilon}{2})\right) \\
&\leq |A||Y| \left(3\exp(\frac{-n_{min}t^2}{2})\right) \leq \delta
\end{aligned}$$

where n_{min} is the smallest $n_{a,q}$ across all groups $S_{a,q}$. The last step above uses the fact that $\epsilon > t$ to simplify the term. This fact follows from Equation 9 and uses the assumption from Theorem 1 that $\epsilon > \frac{\alpha}{2}$. Rearranging the term and then plugging in $t = \frac{\alpha}{2}$, we get the following lower bound for n_{min}:

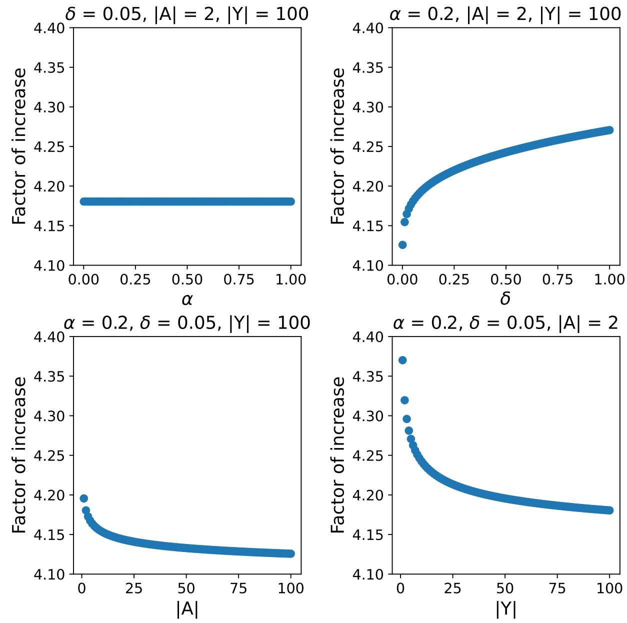

Fig. 3. Relationship between auditing parameters (α, δ, $|A|$, and $|Y|$) and the factor of increase in sample size.

$$\min_{a \in A} n_{a,q} = n_{min} \geq \frac{2}{t^2} \ln \frac{3|A||Y|}{\delta} = \frac{8}{\alpha^2} \ln \frac{3|A||Y|}{\delta}$$

□

We next give the following upper bound on the factor by which number of samples increase when a privacy mechanism is added to conclude the proof of the theorem.

Lemma 5.3. Compared to the non-private case (Lemma 5.1), at most 6.34 times as many samples are needed to perform the audit with differential privacy guarantees (Lemma 5.2)).

$$\frac{\frac{8}{\alpha^2} \ln \frac{3|A||Y|}{\delta}}{\frac{2}{\alpha^2} \ln \frac{2|A||Y|}{\delta}} \leq 4 * \frac{\ln(3)}{\ln(2)} \approx 6.34 \tag{11}$$

We prove the above lemma in Appendix B. While this is a strict upper bound, for reasonable auditing parameters, the overhead is much lower, around 4. Figure 3 shows these relationships for the four parameters: α, δ, $|A|$, and $|Y|$. For all parameters, the factor of increase stays close to 4, lower than the true upper bound of 6.34. We omit ϵ from the plots because the upper bound stays the same for any $\epsilon > \frac{\alpha}{2}$. Since a typical value for α will be close to 0, this constraint allows for small values of ϵ that provide reasonable privacy guarantees.

As an example of this more typical upper bound, say the auditor sets the fairness gap to $\alpha = 0.2$, a comparable parameter to the 4/5ths rule that is commonly applied to test for adverse impact [23]. Assume there are 2 demographic groups ($|A| = 2$) and that relevance scores range from 1 to 100 ($|Y| = 100$), and assume the auditor would like to evaluate fairness with 95% confidence ($\delta = 0.05$). Then, the auditor needs a minimum of 1,879 samples from each demographic group to do the evaluation with privacy guarantees, compared to 450 samples without privacy, which is a 4.17x increase. Such a sample size is reasonable compared to cohorts of several thousands of users used in prior external audits performed on social media platforms [2, 44], and is at the same order of magnitude achieved by current opt-in crowdsourcing efforts [76].

This small constant factor represents the increase in number of samples that ensuring the protection of differential privacy requires. More importantly, it demonstrates that ensuring privacy need not be a barrier to implementation of platform-supported auditing.

6 RELATED WORK

As algorithmic decision making systems have become ubiquitous, there is a growing call for auditing them for potential harmful behavior. We highlight below such work on methods for algorithmic auditing, their use on social media and their trade-offs with privacy.

Methods for Algorithmic Auditing: Audits can be either internal, performed by employees of companies with direct access to their systems, or external, performed by independent third-party entities with usually only user-level access to the systems. We highlight how the platform-supported auditing framework we propose compares to existing auditing methods.

Sandvig et al. provides an overview and taxonomy of external algorithmic auditing methods [71]. The taxonomy identifies five categories for types of audits: source code audit, survey-based audit, scraping audit, sock puppet audit, and crowdsourced audit. Using this taxonomy, a recent literature review categorized past algorithmic audits done on Internet platforms [8]. Our proposal for platform-supported auditing would extend this taxonomy of audits. It differs from source code audits because it only requires that auditors to have query access to algorithms' output without access to the underlying code. It differs from the other four types of methods because it requires a privileged and auditor-specific query interface.

There are newer proposals for auditing that do not directly fit into the Sandvig taxonomy. In *everyday algorithm auditing*, users of social media platforms identify problematic behavior on social media platforms through their normal, day-to-day interactions with the platforms [73]. Their case studies show the power of everyday users to identifying problematic algorithms without a centralized and organized audit study. Our proposal differs because, first, it assumes the auditor has the technical expertise regarding algorithmic fairness and privacy. Second, it enables studying harms that users cannot identify through their day-to-day use of the platforms. An example of such harm is discriminatory ad delivery, as a user cannot know which ads they were targeted with but were not shown.

A second recent proposal is *software-supported auditing* for augmenting the effectiveness of crowdsourced audits by using automation to choose auditing parameters such as audit prompts, and sample size [56]. While Matias' work rigorously estimates sample sizes, they do not analyze how adding a privacy guarantee changes the sample size required, something we add in §5.

Reisman et al. proposes a framework for performing algorithmic impact assessments and enumerates challenges around them [69]. Among other recommendations, they identify the need for external auditors to have meaningful access to periodically assess the impact of algorithms but they do not suggest how auditing can be done while protecting privacy. Metaxa et al. emphasizes the need to evaluate the role of personalization when auditing algorithmic systems [58]. Our work provides a concrete proposal for how to implement an audit of social media platforms' personalization algorithms while safeguarding the privacy of users.

An audit of Pymetrics, a startup that offers a job candidate screening service, performed by external researchers in 2020 proposes a new *cooperative audit* framework, where the target platform gives the auditor special access to its source code and data [87]. This framework is similar to our work in that it requires platform collaboration. Our framework differs in that it requires only query access to the platform's algorithms, and does not require access to underlying proprietary source code and data; furthermore, it protects the privacy of the individuals participating in the audit.

Use of Algorithm Audits on Social Media: Several studies have investigated the role of social media algorithms in biased delivery of both organic content and promoted ads. Sweeney's empirical study of Google Search ads [75] was the first to hypothesize that platform-driven decisions can lead to discriminatory ad delivery; a hypothesis strengthened by evidence from subsequent works [18, 31, 50]. Ali and Sapiezynski et al. confirmed this hypothesis by showing Facebook's algorithms skew delivery of job and housing ads by gender and race even when an advertiser targets a neutral audience [2]. In our prior work, we showed how to control for job qualifications on Facebook and LinkedIn, providing evidence that skew on Facebook may be discriminatory under U.S. law [44]. While these studies successfully identified harms, each has limitations we discuss in §2.2. The new method we propose can be used to audit societal impacts of ad delivery algorithms while accounting for user privacy and other limitations.

Audits have also evaluated how social media algorithms bias delivery of organic content. A sock-puppet study of Facebook's newsfeed, with a focus on content generated leading up to the Italian election in 2018, shows the algorithms cause ranking bias [35]. A similar sock-puppet audit compared reverse-chronological and algorithmic timelines on Twitter to show the platform's algorithms distort content that is shown to users [9]. An internal audit by Twitter also looked at the effect of algorithmic timelines on political content and found their algorithms amplify content unequally across the political spectrum [43]. These studies quantify biases by comparing algorithmic and chronological timelines. Although we do not apply our work to bias in organic content, our framework is generalizable to studying where such biases may arise from.

Algorithmic Auditing and Privacy: Auditing for fairness while protecting privacy of users is also an active area of research that our work contributes to. Segal et al. proposed a privacy-preserving framework for certifying the fairness of machine learning models through an interactive test [72]. Their framework protects privacy of auditors' query inputs by using secure computation to ensure the model owner does not see the data in the queries. In contrast, our method assumes user data is already known to the platform, as is in the case of social media platforms. Our framework focuses on protecting the information query outputs leak about users or the platforms' algorithms to the auditor.

Other studies at the intersection of auditing and privacy have also looked at addressing privacy and other challenges around use of demographic data. Studies by Holstein [40] and later by Andrus [4] interviewed practitioners from a wide range of industries to map out such challenges and normative questions around collection, inference, and use of sensitive demographics attributes of users for fairness efforts [4, 40]. Similarly, Bogen et al. discusses the challenges around access to demographic attributes that arise due to different laws and inconsistent practices across different domains such as credit, employment, and health [12]. Platforms like Meta are actively working to address these challenges with new mechanisms for internal studies of the impact of sensitive attributes while protecting privacy [1, 6]. Our proposal sidesteps these challenges as it does not require platforms to collect or store sensitive attributes; they only need to be known by the external auditor. Similar to our work, Veale et al.'s proposes use of a trusted third-party entity to collect demographic data of users of an algorithmic system and later used the data for auditing the system [84]. Our proposal differs in that it does not require collection of demographic attributes of all users, but just enough number needed to conduct an audit.

7 IMPLICATIONS AND FUTURE WORK

Privacy concerns have hindered increasing transparency into operation of social media platforms. Our work addresses this challenge by showing it is feasible to audit relevance estimators, the "brains" of social media platforms, without violating the privacy of their users or revealing proprietary details of the platforms' algorithms.

Our proposal for platform-supported auditing gives a practical framework for implementing policies outlined in DSA and PATA. Our framework focuses on these proposals as both are promising efforts to increasing transparency of social media platforms and their algorithms' role in influencing individuals and shaping societal discourse. Compared to prior proposals in the U.S. [16, 55, 68], PATA is the most comprehensive in terms of the large platforms it covers [63]. Even if PATA's ultimate fate is uncertain, the EU-centric DSA that has already been passed as law may influence future policies in the U.S. and beyond, similar to the way EU's GDPR has shaped the global privacy landscape [52]. As an example, YouTube's announcement of the YouTube Researcher Program for researchers in more than 50 countries came on the heels of the passing of the DSA [89].

The scope of our framework has limitations that are potential avenues for future work. For example, DSA's proposal covers platforms and services other than social media that are outside the scope of our study. Also, within social media platforms, our work focuses on how organic and promoted content is delivered on users' newsfeed, a place where users consume most of their content. However, there are other features, such as Trends on Twitter, chosen by platform's algorithms, which we do not address in our work but are worth studying for potential harms such as misinformation.

Another potential direction for future work is exploring how platform-supported auditing can be adopted to study other forms of algorithmic harms. Our example use case focuses on auditing for discrimination in job ad delivery. A potential direction is exploring privacy mechanisms and metrics of fairness that are applicable for performing audits in other contexts, such as political or hateful content, while safeguarding privacy of users.

Our work assumes audits will be conducted under a legal framework that incentivizes platform to act in good faith (§4.4), but another area of future work is to relax this assumption and add technical methods that look for accidental errors or intentional non-compliance by platforms. Correlation of data has detected lapses in the past [81]. Technical methods, combined with the legal incentives proposed in DSA and PATA, would provide even stronger guarantees that audits are accurate and complete.

8 CONCLUSION

Auditing social media platforms for public interest is an active and pressing area of academic research, policy-making and legislation. To address concerns raised by prior audits, legislations have been proposed to mandate auditing by external researchers without compromising privacy of platform users and business interest of platforms. We propose a platform-supported auditing framework that has safeguards for protecting against these risks. The center of our mechanism is *increasing transparency of relevance estimators*, which are the core drivers of both organic and promoted content choice and prioritization on social media. Our analysis shows privacy-preserving auditing of relevance estimators can be implemented with high statistical confidence, provided that the sample size is increased by a small constant factor. Our findings offer a novel technical solution for how to practically implement public oversight of social media companies, a core goal the proposed legislations are pushing for.

9 ACKNOWLEDGMENTS

This work was funded in part by a grant from The Rose Foundation and NSF awards #1916153, #1956435, and #1943584.

REFERENCES

[1] Alao, R., Bogen, M., Miao, J., Mironov, I., and Tannen, J. How Meta is working to assess fairness in relation to race in the U.S. across its products and systems. Technical Report https://ai.facebook.com/research/publications/how-meta-

is-working-to-assess-fairness-in-relation-to-race-in-the-us-across-its-products-and-systems, 2021.

[2] Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., and Rieke, A. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing* (2019).

[3] Ali, M., Sapiezynski, P., Korolova, A., Mislove, A., and Rieke, A. Ad delivery algorithms: The hidden arbiters of political messaging. In *14th ACM International Conference on Web Search and Data Mining* (2021).

[4] Andrus, M., Spitzer, E., Brown, J., and Xiang, A. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021), FAccT '21, Association for Computing Machinery, p. 249–260.

[5] Asplund, J., Eslami, M., Sundaram, H., Sandvig, C., and Karahalios, K. Auditing race and gender discrimination in online housing markets. *Proceedings of the International AAAI Conference on Web and Social Media 14*, 1 (May 2020), 24–35.

[6] Austin, R. L. Race Data Measurement and Meta's Commitment to Fair and Inclusive Products. Meta blog. https://about.fb.com/news/2021/11/inclusive-products-through-race-data-measurement, 2021. Accessed: 2022-04-20.

[7] Austin, R. L. Expanding our work on ads fairness. https://about.fb.com/news/2022/06/expanding-our-work-on-ads-fairness/, 2022.

[8] Bandy, J. Problematic machine behavior: A systematic literature review of algorithm audits, 2021.

[9] Bartley, N., Abeliuk, A., Ferrara, E., and Lerman, K. Auditing algorithmic bias on twitter. In *13th ACM Web Science Conference 2021* (New York, NY, USA, 2021), WebSci '21, Association for Computing Machinery, p. 65–73.

[10] Beigi, G., and Liu, H. A survey on privacy in social media: Identification, mitigation, and applications. *ACM/IMS Trans. Data Sci. 1*, 1 (2020).

[11] Bogen, M., Rieke, A., and Ahmed, S. Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020), FAT* '20, Association for Computing Machinery, p. 492–500.

[12] Bogen, M., Rieke, A., and Ahmed, S. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020).

[13] Calandrino, J. A., Kilzer, A., Narayanan, A., Felten, E. W., and Shmatikov, V. "you might also like:" privacy risks of collaborative filtering. In *2011 IEEE Symposium on Security and Privacy* (2011), pp. 231–246.

[14] Chowdhury, R., and Williams, J. Introducing Twitter's first algorithmic bias bounty challenge. https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge, 2021.

[15] Clark, M. Research cannot be the justification for compromising people's privacy. https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/, 2021.

[16] Clarke, Y. Algorithmic Accountability Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231, 2021.

[17] Datta, A., Datta, A., Makagon, J., Mulligan, D. K., and Tschantz, M. C. Discrimination in online personalization: A multidisciplinary inquiry. *FAT* (2018).

[18] Datta, A., Tschantz, M. C., and Datta, A. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 1 (2015).

[19] Do, V., Corbett-Davies, S., Atif, J., and Usunier, N. Online certification of preference-based fairness for personalized recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2022), vol. 36, pp. 6532–6540.

[20] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography* (Berlin, Heidelberg, 2006), S. Halevi and T. Rabin, Eds., Springer Berlin Heidelberg, pp. 265–284.

[21] Edelson, L., Chuang, J., Franklin Fowler, E., Franz, M., and Ridout, T. N. Universal digital ad transparency. *Available at SSRN 3898214* (2021).

[22] Edelson, L., Sakhuja, S., Dey, R., and McCoy, D. An analysis of United States online political advertising transparency. *arXiv preprint arXiv:1902.04385* (2019).

[23] EEOC. Uniform guidelines on employment selection procedures, 29 c.f.r. §1607.4(d), 2018.

[24] European Commission. Artificial intelligence act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206, 2021.

[25] European Commission. The Digital Services Act: Ensuring a safe and accountable online environment. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM%3A2020%3A825%3AFIN, 2021.

[26] European Commission. Digital Services Act: Commission welcomes political agreement on rules ensuring a safe and accountable online environment. https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545, 2022.

[27] European Commission. Digital services package: Commission welcomes the adoption by the european parliament of the eu's new rulebook for digital services. https://ec.europa.eu/commission/presscorner/detail/en/IP_22_4313, 2022.

[28] Facebook. How Facebook distributes content. https://www.facebook.com/business/help/718033381901819?id=

208060977200861.
[29] Facebook. About ad auctions. https://www.facebook.com/business/help/430291176997542?id=561906377587030, 2021.
[30] Facebook. Facebook Open Research & Transparency. https://fort.fb.com/, 2022.
[31] Gelauff, L., Goel, A., Munagala, K., and Yandamuri, S. Advertising for demographically fair outcomes. *arXiv preprint arXiv:2006.03983* (2020).
[32] Georgia Wells, J. H., and Seetharaman, D. Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show. https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739, 2021.
[33] Graham, M. Facebook knew ad metrics were inflated, but ignored the problem to make more money, lawsuit claims. https://www.cnbc.com/2021/02/18/facebook-knew-ad-metrics-were-inflated-but-ignored-the-problem-lawsuit-claims.html, 2021.
[34] Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NIPS* (2016).
[35] Hargreaves, E., Agosti, C., Menasche, D., Neglia, G., Reiffers-Masson, A., and Altman, E. Biases in the facebook news feed: A case study on the italian elections. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (Aug 2018).
[36] Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z., and Wojcieszak, M. Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations, 2022.
[37] Hill, K. How Target figured out a teen girl was pregnant before her father did. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/, 2012.
[38] Hill, K. 'People You May Know:' A Controversial Facebook Feature's 10-Year History. https://gizmodo.com/people-you-may-know-a-controversial-facebook-features-1827981959, 2018.
[39] Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 58*, 301 (1963), 13–30.
[40] Holstein, K., Wortman Vaughan, J., Daumé, H., Dudik, M., and Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), Association for Computing Machinery, p. 1–16.
[41] Horwitz, J. Facebook seeks shutdown of nyu research project into political ad targeting. https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533, 2020.
[42] Horwitz, J., and Seetharaman, D. Facebook executives shut down efforts to make the site less divisive. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499, 2020.
[43] Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjera, A., and Hardt, M. Algorithmic amplification of politics on twitter. https://cdn.cms-twdigitalassets.com/content/dam/blog-twitter/official/en_us/company/2021/rml/Algorithmic-Amplification-of-Politics-on-Twitter.pdf, 2021.
[44] Imana, B., Korolova, A., and Heidemann, J. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021* (New York, NY, USA, 2021), Association for Computing Machinery, p. 3767–3778.
[45] Jung, J. C., and Sharon, E. The Volkswagen emissions scandal and its aftermath. *Global Business and Organizational Excellence 38*, 4 (2019), 6–15.
[46] Kemp, S. Digital 2021: Global overview report. https://datareportal.com/reports/digital-2021-global-overview-report, 2021.
[47] Kiberd, R. Why 2016 was the year of the algorithmic timeline. https://www.vice.com/en/article/bmvbaw/why-2016-was-the-year-of-the-algorithmic-timeline, 2021.
[48] King, G., and Persily, N. Unprecedented Facebook URLs Dataset now Available for Academic Research through Social Science One. https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one, 2020.
[49] Koumchatzky, N., and Andryeyev, A. Using deep learning at scale in Twitter's timelines. https://blog.twitter.com/engineering/en_us/topics/insights/2017/using-deep-learning-at-scale-in-twitters-timelines, 2017.
[50] Lambrecht, A., and Tucker, C. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science 65*, 7 (2019), 2966–2981.
[51] Lecuyer, M., Spahn, R., Spiliopolous, Y., Chaintreau, A., Geambasu, R., and Hsu, D. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (New York, NY, USA, 2015), CCS '15, Association for Computing Machinery, p. 554–566.
[52] Linden, T., Khandelwal, R., Harkous, H., and Fawaz, K. The privacy policy landscape after the gdpr. *Proceedings on Privacy Enhancing Technologies 2020*, 1 (2020), 47–64.
[53] LinkedIn. How the LinkedIn ads auction works — and how you can benefit. https://www.linkedin.com/business/marketing/blog/linkedin-ads/how-the-linkedin-ads-auction-works-and-how-you-can-benefit, 2021.
[54] LinkedIn. Together we can improve equal access to opportunity. https://members.linkedin.com/equal-access, 2022.
[55] Markey, E., and Matsui, D. Algorithmic justice and online platform transparency act of 2021. https://www.markey.

senate.gov/imo/media/doc/ajopta.pdf, 2021.
[56] Matias, J. N., Hounsel, A., and Feamster, N. Software-supported audits of decision-making systems: Testing google and facebook's political advertising policies. In *The 25th ACM Conference on Computer-Supported Cooperative Work and Social Computing* (2021).
[57] Merrill, J. B., and Tobin, A. Facebook moves to block ad transparency tools — including ours. https://www.propublica.org/article/facebook-blocks-ad-transparency-tools, 2019.
[58] Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., and Sandvig, C. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction 14*, 4 (2021), 272–344.
[59] Mohamed, A., and Li, Z. Community-focused feed optimization. https://engineering.linkedin.com/blog/2019/06/community-focused-feed-optimization, 2021.
[60] Mohan, P. Facebook's news feed just turned 10. https://www.fastcompany.com/4018352/facebooks-news-feed-just-turned-10, 2016.
[61] Mosseri, A. Shedding More Light on How Instagram Works. https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works, 2021.
[62] Mozilla. Mozilla rally. https://rally.mozilla.org/.
[63] Nonnecke, B., and Carlton, C. Eu and us legislation seek to open up digital platform data. *Science 375*, 6581 (2022), 610–612.
[64] NYU Ad Observatory. Explore Facebook political ads. https://adobservatory.org/.
[65] Papakyriakopoulos, O., Tessono, C., Narayanan, A., and Kshirsagar, M. How algorithms shape the distribution of political advertising: Case studies of facebook, google, and tiktok. *arXiv preprint arXiv:2206.04720* (2022).
[66] Persily, N. A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act. *Journal of Online Trust and Safety 1*, 1 (October 2021).
[67] Purnell, N., and Horwitz, J. Facebook services are used to spread religious hatred in india, internal documents show. https://www.wsj.com/articles/facebook-services-are-used-to-spread-religious-hatred-in-india-internal-documents-show-11635016354, 2021.
[68] Racine, K. Stop discrimination by algorithms act of 2021. https://oag.dc.gov/sites/default/files/2021-12/DC-Bill-SDAA-FINAL-to-file-.pdf, 2021.
[69] Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now* (2018).
[70] Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., and Meira, W. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020), FAT* '20, Association for Computing Machinery, p. 131–141.
[71] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. Auditing algorithms : Research methods for detecting discrimination on internet platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (2014).
[72] Segal, S., Adi, Y., Pinkas, B., Baum, C., Ganesh, C., and Keshet, J. Fairness in the eyes of the data: Certifying machine-learning models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA, 2021), AIES '21, Association for Computing Machinery, p. 6–935.
[73] Shen, H., DeVos, A., Eslami, M., and Holstein, K. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proc. ACM Hum.-Comput. Interact. 5*, CSCW2 (Oct. 2021).
[74] Speicher, T., Ali, M., Venkatadri, G., Ribeiro, F. N., Arvanitakis, G., Benevenuto, F., Gummadi, K. P., Loiseau, P., and Mislove, A. Potential for discrimination in online targeted advertising. In *Proceedings of Machine Learning Research* (2018), S. A. Friedler and C. Wilson, Eds.
[75] Sweeney, L. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue* (2013).
[76] The Markup. The Citizen Browser Project—Auditing the Algorithms of Disinformation. https://themarkup.org/citizen-browser, 2020.
[77] The US Department of Justice. 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, 2022.
[78] The Wall Street Journal. The Facebook Files - A Wall Street Journal Investigation. https://www.wsj.com/articles/the-facebook-files-11631713039, 2021.
[79] TikTok. How TikTok recommends videos For You. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you.
[80] TikTok. Relevance is the new reach. https://www.tiktok.com/business/en-US/blog/relevance-is-the-new-reach.
[81] Timberg, C. Facebook made big mistake in data it provided to researchers, undermining academic work. The

Washington Post https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/, 2021.
[82] Twitter. The Twitter ads auction. https://business.twitter.com/en/help/troubleshooting/bidding-and-auctions-faqs.html, 2021.
[83] U.S. Senate. Platform accountability and transparency act. https://www.coons.senate.gov/imo/media/doc/text_pata_117.pdf, 2021.
[84] Veale, M., and Binns, R. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society 4*, 2 (2017), 2053951717743530.
[85] Vermeulen, M. The keys to the kingdom. https://knightcolumbia.org/content/the-keys-to-the-kingdom, 2021.
[86] Weinsberg, U., Bhagat, S., Ioannidis, S., and Taft, N. BlurMe: Inferring and obfuscating user gender based on ratings. RecSys '12, Association for Computing Machinery, p. 195–202.
[87] Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., and Polli, F. Building and auditing fair algorithms: A case study in candidate screening. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2021).
[88] Yee, K., Tantipongpipat, U., and Mishra, S. Image cropping on twitter: Fairness metrics, their limitations, and the importance of representation, design, and agency. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW2 (2021), 1–24.
[89] YouTube. YouTube researcher program. https://research.youtube/, 2022.

A NUMBER OF SAMPLES REQUIRED FOR AUDITING WITHOUT PRIVACY

In §5.2, we suggest Lemma 5.1 holds as lower bound for number of samples required for auditing without privacy. Here we give a detailed proof for Lemma 5.1. Our proof follows Segal et al.'s work [72], with modifications to adopt it to our use-case that considers qualification as an additional attribute of users.

Proof. An auditor uses a sample set S of users to perform an audit. Consider the following "bad" event where the sampling error is above some threshold $t > 0$:

$$\overline{P}_{a,y}(R,S) \text{ is bad if}: \ |\overline{P}_{a,y}(R,S) - P_{a,y}(R)| > t.$$

We would like to bound the probability of this event for all demographic groups in A and possible outputs in Y:

$$Pr[\exists a \in A \text{ and } y \in Y : \overline{P}_{a,y}(R,S) \text{ is bad}] \leq \delta$$

We use union bound followed by Hoeffding's concentration bound:

$$\begin{aligned}
&Pr[\exists a \in A \text{ and } y \in Y : \overline{P}_{a,y}(R,S) \text{ is bad}] \\
&\leq \sum_{a\in A}\sum_{y\in Y} Pr\left[\overline{P}_{a,y}(R,S) \text{ is bad}\right] \\
&= \sum_{a\in A}\sum_{y\in Y} Pr\left[|\overline{P}_{a,y}(R,S) - P_{a,y}(R)| > t\right] \\
&\leq \sum_{a\in A}\sum_{y\in Y} 2\exp(-2n_{a,q}t^2) \\
&= \sum_{a\in A} |Y| 2\exp(-2n_{a,q}t^2) \\
&\leq |A||Y|2\exp(-2n_{min}t^2)
\end{aligned}$$

where n_{min} is the number of people in a group in S that has least number of qualified people. We want the above probability to be small, i.e., $|A||Y|2\exp(-2n_{min}t^2) \leq \delta$. Rearranging, we get the following bound on n_{min}:

$$n_{min} \geq \frac{1}{2t^2} \ln \frac{2|A||Y|}{\delta} \tag{12}$$

We next derive the value of t needed to guarantee (α, δ)-fairness. Based on Definition 5.2, it is sufficient to show that, for any pair $a_1, a_2 \in A$ and any $y \in Y$, the fairness gap is bounded by α:

$$|P_{a_1,y}(R) - P_{a_2,y}(R)| \leq \alpha$$

Conditioning on the above bad event not occurring, we start with $|P_{a_1,y}(R) - P_{a_2,y}(R)|$ and apply triangle inequality. In the second inequality below, we add the term $(EFG(R, S) - |\overline{P}_{a_1,y}(R, S) - \overline{P}_{a_2,y}(R, S)|)$ because it is positive (based on definition of EFG).

$$\begin{aligned}
&|P_{a_1,y}(R) - P_{a_2,y}(R)| \\
&\leq |P_{a_1,y}(R)| + |P_{a_2,y}(R)| \\
&\leq |P_{a_1,y}(R)| + |P_{a_2,y}(R)| + (EFG(R, S) - |\overline{P}_{a_1,y}(R, S) - \overline{P}_{a_2,y}(R, S)|) \\
&\leq |P_{a_1,y}(R)| + |P_{a_2,y}(R)| + EFG(R, S) - (|\overline{P}_{a_1,y}(R, S)| + |\overline{P}_{a_2,y}(R, S)|) \\
&= |P_{a_1,y}(R)| - |\overline{P}_{a_1,y}(R, S)| + |P_{a_2,y}(R)| - |\overline{P}_{a_2,y}(R, S)| + EFG(R, S) \\
&\leq t + t + EFG(R, S)
\end{aligned}$$

We want $2t + EFG(R, S) \leq \alpha$. Therefore, $t \leq \frac{\alpha - EFG(R,S)}{2}$. For $EFG(R, S) \leq \alpha$, $t \leq \frac{\alpha - EFG(R,S)}{2} \leq \frac{\alpha}{2}$ holds. Plugging in the value of $t = \frac{\alpha}{2}$ to Equation 12 gives us a lower bound for the number of samples needed:

$$n_{min} \geq \frac{2}{\alpha^2} \ln \frac{2|A||Y|}{\delta}$$

□

B UPPER BOUND ON INCREASE OF NUMBER OF SAMPLES

In this appendix we give a proof for Lemma 5.3 to provide an upper bound on the factor by which number of samples increase when a privacy mechanism is added.

Proof. Let $P = \ln \frac{|A||Y|}{\delta}$. Then:

$$\frac{\frac{8}{\alpha^2} \ln \frac{3|A||Y|}{\delta}}{\frac{2}{\alpha^2} \ln \frac{2|A||Y|}{\delta}} = 4 \left(\frac{\ln \frac{3|A||Y|}{\delta}}{\ln \frac{2|A||Y|}{\delta}} \right) = 4 \left(\frac{\ln 3 + \ln \frac{|A||Y|}{\delta}}{\ln 2 + \ln \frac{|A||Y|}{\delta}} \right) = 4 \left(\frac{\ln 3 + P}{\ln 2 + P} \right)$$

Because δ is a probability and A and Y cannot be empty, we know $|A| \geq 1$, $|Y| \geq 1$, and $\delta \leq 1$. Therefore, it is always the case that $\frac{|A||Y|}{\delta} \geq 1$ and $P \geq 0$.

Now, consider $f(P) = 4 \left(\frac{\ln 3 + P}{\ln 2 + P} \right)$. Because $P \geq 0$, $f(P)$ is maximized when $P = 0$, and monotonically decreases as P increases. Therefore, $f(P) \leq f(0)$ for all $P \geq 0$. Finally:

$$\frac{\frac{8}{\alpha^2} \ln \frac{3|A||Y|}{\delta}}{\frac{2}{\alpha^2} \ln \frac{2|A||Y|}{\delta}} = 4 \left(\frac{\ln 3 + P}{\ln 2 + P} \right) = f(P) \leq f(0) = 4 \times \frac{\ln(3)}{\ln(2)} \approx 6.34$$

□

C APPLYING HOEFFDING'S

In §5.2, we use Hoeffding's inequality to bound Equation 10. Here we give a proof for why we can apply Hoeffding's inequality even in the presence of potential bias in R.

From Equation 10, we would to apply Hoeffding's bound to the following sampling error term:

$$Pr\left[\left|\overline{P}_{a,y}(R,S) - P_{a,y}(R)\right| > \frac{t}{2}\right]$$

Hoeffding's inequality gives an upper bound on the probability that the sum of bounded random variables deviates from its expected value [39].

To apply Hoeffding's, we need to show sampling is i.i.d. and that we are summing bounded random variables. An auditor can sample i.i.d. in several ways: the platform may provide sampling or the auditor may use an external source of a unique set of users. Based on Equation 2, $\overline{P}_{a,y}(R,S)$ is a sum of $n_{a,q}$ indicator variables defined on each sample in $S_{a,q}$. Indicator variables can only hold a value of 0 or 1, so they are bounded. The remaining requirement we need to show to apply Hoeffding's is:

$$E[\overline{P}_{a,y}(R,S)] = P_{a,y}(R) \tag{13}$$

The goal of the auditor is to test for potential bias that is correlated with some sensitive attribute. We next show $\overline{P}_{a,y}(R,S)$ is unbiased estimator of $P_{a,y}(R)$ (by showing Equation 13 holds) even in the presence of bias per group as long as the samples in $S_{a,q}$ are i.i.d.

Bias that is an additive, constant factor. Consider the following formulation that takes such bias into account:

$$R(x) = T(x) + b_a \tag{14}$$

where b_a is a constant bias for a user with attribute a, and $T(x)$ is a random variable reflecting that individual x's history.

As mentioned before, $S_{a,q}$ represent subset of S with given values of a and q. We consider the subset of qualified individuals so $q = 1$. Let $\overline{S_{a,q}}$ represent the complement of $S_{a,q}$.

$$\begin{aligned}
E[\overline{P}_{a,y}(R,S)] &= E\left[\frac{1}{n_{a,q}}\sum_{i=1}^{|S|}\mathbb{1}\{R(x_i) = y \wedge a_i = a \wedge q_i = 1\}\right] \\
&= E\left[\frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|}\mathbb{1}\{R(x_i) = y\} + \sum_{i=1}^{|\overline{S_{a,q}}|} 0\right] \quad \text{.....separate } S_{a,q} \text{ and } \overline{S_{a,q}} \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|} E\left[\mathbb{1}\{R(x_i) = y\}\right] \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|} 0 * Pr[R(x_i) \neq y] + 1 * Pr[R(x_i) = y] \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|} Pr[R(x_i) = y] \qquad (15) \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|} Pr[T(x_i) + b_a = y]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} Pr[T(x_i) = y - b_a] \\
&= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} P_{a,y-b_a}(T) \qquad \text{.....by i.i.d. assumption} \\
&= \frac{n_{a,q}}{n_{a,q}} P_{a,y-b_a}(T) \\
&= P_{a,y-b_a}(T) \\
&= P_{a,y}(R) \qquad \text{.....plug in Equation 14 in Equation 1}
\end{aligned}
$$

Therefore, we can apply Hoeffding's for samples in a group even if the group attribute induces an additive bias.

Bias that is a multiplicative, constant factor. One can follow similar steps to show Equation 13 holds for the case a multiplicative constant bias. Let

$$R(x) = T(x) * b_a \tag{16}$$

where b_a is a constant.

$$
\begin{aligned}
E[\overline{P}_{a,y}(R,S)] &= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} Pr[R(x_i) = y] \qquad \text{.......by Equation 15} \\
&= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} Pr[T(x_i) * b_a = y] \\
&= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} Pr[T(x_i) = \frac{y}{b_a}] \\
&= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} P_{a,\frac{y}{b_a}}(T) \qquad \text{.......by i.i.d. assumption} \\
&= \frac{n_{a,q}}{n_{a,q}} P_{a,\frac{y}{b_a}}(T) \\
&= P_{a,\frac{y}{b_a}}(T) \\
&= P_{a,y}(R) \qquad \text{.....plug in Equation 16 in Equation 1}
\end{aligned}
$$

Bias that is a random variable (not a constant). Consider bias that is a discrete random variable and is an additive factor. Let $R(x) = T(x) + B_a$ where B_a is a discrete random variable. We would like to show Equation 13 holds for this case. We look at each side of the equation separately:

$$
\begin{aligned}
E[\overline{P}_{a,y}(R,S)] &= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} Pr[R(x_i) = y] \qquad \text{.......by Equation 15} \\
&= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} Pr[T(x_i) + B_a = y]
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n_{a,q}} \sum_{i=1}^{|S_{a,q}|} \sum_{b} Pr[T(x_i) = y - b] * Pr[B_a = b] \\
&= \frac{1}{n_{a,q}} \sum_{b} \sum_{i=1}^{|S_{a,q}|} Pr[T(x_i) = y - b] * Pr[B_a = b] \\
&= \frac{1}{n_{a,q}} \sum_{b} Pr[B_a = b] \sum_{i=1}^{|S_{a,q}|} Pr[T(x_i) = y - b] \\
&= \frac{1}{n_{a,q}} \sum_{b} Pr[B_a = b] \sum_{i=1}^{|S_{a,q}|} P_{a,y-b}(T) \qquad \text{.......by i.i.d. assumption} \\
&= \frac{1}{n_{a,q}} \sum_{b} Pr[B_a = b] * n_{a,q} * P_{a,y-b}(T) \\
&= \sum_{b} Pr[B_a = b] * P_{a,y-b}(T)
\end{aligned} \tag{17}
$$

$$
\begin{aligned}
P_{a,y}(R) &= Pr_{(x,a',q)\sim X}\,[R(x) = y | a' = a \wedge q = 1] \\
&= Pr_{(x,a',q)\sim X}\,[T(x) + B_a = y | a' = a \wedge q = 1] \\
&= \sum_{b} Pr[B_a = b] * Pr_{(x,a',q)\sim X}\,[T(x) = y - b | a' = a \wedge q = 1] \\
&= \sum_{b} Pr[B_a = b] * P_{a,y-b}(T)
\end{aligned} \tag{18}
$$

Since Equation 17 and Equation 18 are equal, $E[\overline{P}_{a,y}(R,S)] = P_{a,y}(R)$.