HAVING YOUR PRIVACY CAKE AND EATING IT TOO: PLATFORM-SUPPORTED AUDITING OF SOCIAL MEDIA ALGORITHMS FOR PUBLIC INTEREST

Basileal Imana
Center for Information Technology Policy, Princeton University
Princeton, New Jersey, USA
imana@princeton.edu

Aleksandra Korolova
Department of Computer Science and School of Public and International Affairs, Princeton University
Princeton, New Jersey, USA
korolova@princeton.edu

John Heidemann
Information Sciences Institute, University of Southern California
Los Angeles, California, USA
johnh@isi.edu

ABSTRACT

Social media platforms curate access to information and opportunities, and so play a critical role in shaping public discourse today. The opaque nature of the algorithms these platforms use to curate content raises societal questions. Prior studies have used black-box methods led by experts or collaborative audits driven by everyday users to show that these algorithms can lead to biased or discriminatory outcomes. However, existing auditing methods face fundamental limitations because they function independent of the platforms. Concerns of potential harmful outcomes have prompted proposal of legislation in both the U.S. and the E.U. to mandate a new form of auditing where vetted external researchers get privileged access to social media platforms. Unfortunately, to date there have been no concrete technical proposals to provide such auditing, because auditing at scale risks disclosure of users' private data and platforms' proprietary algorithms. We propose a new method for *platform-supported auditing* that can meet the goals of the proposed legislation. The first contribution of our work is to enumerate the 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 ($6.34\times$ as an upper bound, and $4\times$ 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. The comprehensive details of our work can be found in our full paper Imana et al. (2023).

1 INTRODUCTION

Social media platforms are no longer just digital tools that connect friends and family—today they play a critical role in shaping public discourse and moderating access to information and opportunities. Platforms such as Facebook, Instagram, Twitter, LinkedIn and TikTok have become the new search engines (Kalley Huang, 2022; Sarah Perez, 2022), helping individuals find important life opportunities such as jobs (Hosain, 2021; Nikolaou, 2014), and are often sources of news and advice (Shearer & Gottfried, 2017). Content curation on these platforms is done by *algorithms that estimate relevance of content to users*, which raises fundamental questions about their societal implications. For example: Do these algorithms discriminate against certain demographic groups? Do they amplify political content in a way that threatens democracy? Does optimization for relevance, often defined via engagement, emphasize extreme viewpoints, dividing society?

Researchers and journalists have performed audits in a systematic way to uncover harm on platforms. Examples of their findings include biased or discriminatory ad targeting and delivery (Lecuyer et al., 2015; Ali et al., 2019; Asplund et al., 2020; Imana et al., 2021), amplification of hateful content (Purnell & Horwitz, 2021), political polarization (Ribeiro et al., 2020; Horwitz & Seetharaman, 2020; Huszár et al., 2022; Haroon et al., 2022; Papakyriakopoulos et al., 2022), and promotion of addictive behavior in teens (Georgia Wells & Seetharaman, 2021). Although existing methods that auditors use have been crucial in uncovering harms and driving change, they are reaching hard limits in terms of what they can reliably and provably learn about the role of platforms' algorithms (Ali et al., 2019; 2021; Imana et al., 2021). We expand on limitations of existing methods in §2.2.

The importance of these risks has prompted policy and legal efforts aimed to increase transparency in social media platforms. Of many such proposals, the U.S. Platform Accountability and Transparency Act (PATA (Coons et al., 2022)) and the E.U. Digital Services Act (DSA (European Commission, 2022c)) are the most comprehensive, addressing broad algorithmic risks and platforms (Nonnecke & Carlton, 2022). PATA has bipartisan support, but as of October 2023, has not yet been debated in the U.S. Congress. On the other hand, the E.U. passed DSA and the law is already in effect (European Commission, 2022a). 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*.

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 (Bogen et al., 2020a; Alao et al., 2021; Huszár et al., 2022). Platforms such as Meta have also citepd privacy as a constraint on increasing their transparency efforts (Clark, 2021; Vermeulen, 2021). Existing proposals or implementations for providing privacy protection require researchers to sign strict NDAs or operate in "clean rooms" (Persily, 2021). Clean rooms, secure environments where researchers are monitored during data access, and NDAs may provide strong limits on data disclosure, but they impose a significant burden on auditors.

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. 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). 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. 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. We theoretically 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 do so by analyzing the trade-off between guaranteeing privacy and the minimum sample size required for auditing in one concrete scenario – bias in delivery of employment ads.

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 Explicit Platform Support

Before we describe our platform-supported auditing framework, we discuss why explicit platform support is needed in light of recent policy developments.

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. Two prominent pieces of legislation that mandate independent oversight and transparency research on platforms are: Platform Accountability and Transparency Act (PATA (Coons et al., 2021), proposed in the US) and Digital Services Act (DSA (European Commission, 2022b), already enforced in the EU). Both DSA and PATA mandate platform support for independent research on algorithmic transparency by allowing vetted academic researchers to have access to the platforms' data.

Both proposals recognize the tension between mandating audits and protecting privacy of users and platforms 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" (Coons et al., 2021). DSA acknowledges platform's desire for "protection of confidential information, in particular trade secrets" (European Commission, 2022b) when conducting audits. To mitigate the risks to users and platforms, both proposals require vetting auditors, their projects, and results before they are published. Platforms themselves also often citep their need to protect user privacy as a handicap for their transparency and self-policing capabilities (Austin, 2021; Alao et al., 2021; Vermeulen, 2021; Huszár et al., 2022). Prior to our work, no actionable technical proposals put forth methods to implement such auditor access while protecting users' privacy and platforms' proprietary algorithms.

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 that can support their efforts such as PATA and DSA. 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. We expand on each of these challenges in Appendix A

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. We discuss our motivation for focusing on relevance estimators in §3, and then our new framework in §4.

Platform-supported audits, of course, require support from the platform, so we give an overview of the evolution of the platforms' responses to requests for auditing in §A.1.

3 Relevance Estimators are the Key to Increasing Transparency

We suggest that giving auditors query access to study relevance estimators is the key to address the limitations of existing methods discussed in the previous section. Relevance estimators are the main drivers that shape every piece of content shown to users. Prior work and platforms' documentation show the importance of these algorithms and how they are currently 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. Facebook's algorithmic newsfeed dates back to 2007 (Mohan, 2016), and Twitter and Instagram deployed such personalization in 2016 (Kiberd, 2021).

Relevance estimators are used to personalize both organic and promoted content. 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 (Facebook; Mosseri, 2021; Koumchatzky & Andryeyev, 2017; Mohamed & Li, 2021; TikTok, 2021). Relevance estimators are used in ad auctions as well. An ad with the highest bid may not win an auction if it is given a low relevance score by the algorithmic prediction (Facebook, 2021; LinkedIn, 2021; Twitter, 2021).

Platforms provide little transparency into their optimization algorithms, neither for organic nor promoted content. 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 (Facebook; Mosseri, 2021; Mohamed & Li, 2021; Koumchatzky & Andryeyev, 2017). However, the specific types of algorithms and inputs to those algorithms are not disclosed.

The importance of relevance estimators to both organic content and ad delivery, and the lack of transparency into how they operate, leads us to place them at the center of our mechanism for auditing.

4 A Privacy-preserving Platform-Supported Auditing Framework

In the previous section, we discussed why adding platform support for auditing relevance estimators is the key to increasing transparency. We next describe how platforms can practically implement such access while safeguarding the privacy of users and platforms.

4.1 Overview

Our proposal for *platform-supported auditing* allows an auditor to evaluate whether, for a given piece of content, the platform's relevance estimator scores that content with bias reflecting protected attributes such as gender or race. The framework is summarized in Figure 1 and 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 demographic 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 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.

Figure 1 summarizes the four steps and captures two key properties of our framework: *generalizability* and *privacy protection*. The two steps on the left side (blue, dotted box) are generalizable because the auditor can vary the audience, content and fairness metric based on the specific problem they are studying. The two steps on the right side (green, solid box) are kept private from the

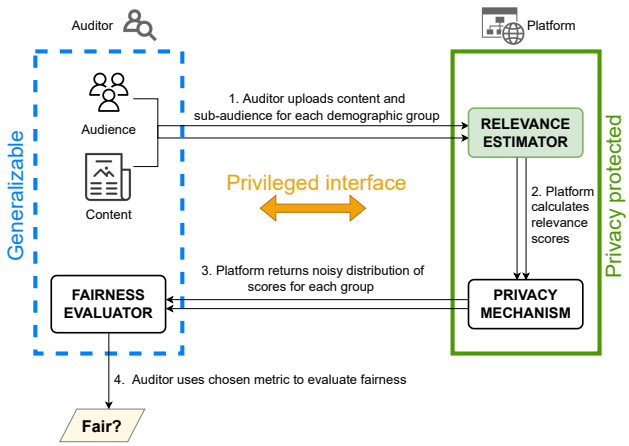

Figure 1: Privacy-preserving platform-supported framework for auditing relevance estimators.

auditor because the platform does not reveal users' private data or proprietary information about its algorithms source code, models or weights.

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, 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. 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 (Calandrino et al., 2011; Weinsberg et al., 2012; Beigi & Liu, 2020). Therefore, our auditing method must limit the potential to make such inferences.

In §4.3, 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*. DP is the current gold standard for protecting privacy of individuals, while providing useful statistical computations on their data (Dwork et al., 2006). 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

Having shown the privacy risks of revealing relevance scores, we next propose our auditing framework that protects against these risks. Our framework has four high-level steps which we describe in detail next.

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 both the content and the audience 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.

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 (Example: a histogram or CDF), and adds privacy protections (discussed next), before returning the statistics to the auditor.

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 guarantee that the data the auditor gets ensures differential privacy for individuals participating in the audit.

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 (Dwork et al., 2006). 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 (Dwork et al., 2006). The platform then returns the noisy histogram counts back to the auditor.

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 (Hardt et al., 2016). We further explore this scenario in our theoretical result.

4.4 TRUST MODEL AND LIMITATIONS

Our proposed framework requires some level of trust for both auditors and platforms because it relies on giving auditors privileged access to platforms' algorithms. The efficiency of our approach assumes a legal framework, such as DSA or PATA, in which both the platforms and auditors work in good faith, or potential tests for non-cooperation. Due to space, we expand on the trust model we use to evaluate the privacy and other risks of our approach, and its limitations in Appendix B.

5 SAMPLE SIZE 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. Due to space limitation, we only informally state our theorem here. We formally state and prove our result in Appendix C.

Theorem 1. An audit relying on a differentially privatized output of a relevance estimator is 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.

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 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. (Racine, 2021; Clarke, 2021; Markey & Matsui, 2021), PATA is the most comprehensive in terms of the large platforms it covers (Nonnecke & Carlton, 2022). 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 (Linden et al., 2020). 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 (YouTube, 2022).

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 (Timberg, 2021). Technical methods, combined with the legal incentives proposed in DSA and PATA, would provide even stronger guarantees that audits are accurate and complete.

7 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 interests of platforms. 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.

REFERENCES

Rachad Alao, Miranda Bogen, Jingang Miao, Ilya Mironov, and Jonathan Tannen. 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, November 2021.

Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 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*, November 2019.

Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Ad delivery algorithms: The hidden arbiters of political messaging. In *14th ACM International Conference on Web Search and Data Mining*, March 2021.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 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*, FAccT '21, New York, NY, USA, March 2021. ISBN 9781450383097. doi: 10.1145/3442188.3445888. URL https://doi.org/10.1145/3442188.3445888.

Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. Auditing race and gender discrimination in online housing markets. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1), May 2020. URL https://ojs.aaai.org/index.php/ICWSM/article/view/7276.

Roy L. Austin. 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, November 2021.

Roy L. Austin. Expanding our work on ads fairness. https://about.fb.com/news/2022/06/expanding-our-work-on-ads-fairness/, June 2022.

Jack Bandy. Problematic machine behavior: A systematic literature review of algorithm audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), apr 2021. doi: 10.1145/3449148. URL https://doi.org/10.1145/3449148.

Nathan Bartley, Andres Abeliuk, Emilio Ferrara, and Kristina Lerman. Auditing algorithmic bias on twitter. In *13th ACM Web Science Conference 2021*, WebSci '21, pp. 65–73, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383301. doi: 10.1145/3447535.3462491. URL https://doi.org/10.1145/3447535.3462491.

Ghazaleh Beigi and Huan Liu. A survey on privacy in social media: Identification, mitigation, and applications. *ACM/IMS Trans. Data Sci.*, 1(1), 2020. ISSN 2691-1922. doi: 10.1145/3343038. URL https://doi.org/10.1145/3343038.

Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. Awareness in practice: Tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 492–500, New York, NY, USA, 2020a. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372877. URL https://doi.org/10.1145/3351095.3372877.

Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020b.

Joseph A. Calandrino, Ann Kilzer, Arvind Narayanan, Edward W. Felten, and Vitaly Shmatikov. "You might also like:" Privacy risks of collaborative filtering. In *2011 IEEE Symposium on Security and Privacy*, pp. 231–246, 2011. doi: 10.1109/SP.2011.40.

Rumman Chowdhury and Jutta Williams. Introducing Twitter's first algorithmic bias bounty challenge. https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge, July 2021.

Mike Clark. 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/, August 2021.

Yvette Clarke. Algorithmic Accountability Act of 2019. https://www.congress.gov/bill/116th-congress/house-bill/2231, April 2021.

Chris Coons, Rob Portman, and Amy Klobuchar. Platform Accountability and Transparency Act. https://www.coons.senate.gov/imo/media/doc/text_pata_117.pdf, 2021.

Chris Coons, Rob Portman, Amy Klobuchar, and Bill Cassidy. Platform Accountability and Transparency Act. https://www.coons.senate.gov/news/press-releases/senator-coons-colleagues-introduce-legislation-to-provide-public-with-transparency-of-social-media-platforms, December 2022.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, (1), 2015. doi: https://doi.org/10.1515/popets-2015-0007. URL https://content.sciendo.com/view/journals/popets/2015/1/article-p92.xml.

Alicia DeVos, Aditi Dhabalia, Hong Shen, Kenneth Holstein, and Motahhare Eslami. Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517441. URL https://doi.org/10.1145/3491102.3517441.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In Shai Halevi and Tal Rabin (eds.), *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

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

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, July 2022a.

European Commission. The Digital Services Act (DSA). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065&qid=1666857835014, October 2022b.

European Commission. The Digital Services Act: Ensuring a safe and accountable online environment. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32022R2065, 2022c.

Facebook. How Facebook distributes content. https://www.facebook.com/business/help/7180333819018l9?id=208060977200861.

Facebook. About ad auctions. https://www.facebook.com/business/help/430291176997542?id=561906377587030, October 2021.

Facebook. Facebook Open Research & Transparency. https://fort.fb.com/, 2022.

Lodewijk Gelauff, Ashish Goel, Kamesh Munagala, and Sravya Yandamuri. Advertising for demographically fair outcomes, 2020. URL https://arxiv.org/abs/2006.03983.

Jeff Horwitz Georgia Wells and Deepa Seetharaman. 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, September 2021.

Megan Graham. 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, February 2021.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems (NIPS)*, 2016.

Eduardo Hargreaves, Claudio Agosti, Daniel Menasche, Giovanni Neglia, Alexandre Reiffers-Masson, and Eitan Altman. 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. doi: 10.1109/asonam.2018.8508659. URL http://dx.doi.org/10.1109/ASONAM.2018.8508659.

Muhammad Haroon, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, Zubair Shafiq, and Magdalena Wojcieszak. Youtube, the great radicalizer? auditing and mitigating ideological biases in youtube recommendations, 2022. URL https://arxiv.org/abs/2203.10666.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: https://doi.org/10.1080%2F01621459.1963.10500830. URL https://repository.lib.ncsu.edu/bitstream/handle/1840.4/2170/ISMS_1962_326.pdf;jsessionid=43EFBA9451E7E87617C394D0B81306E7?sequence=1.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702. URL https://doi.org/10.1145/3290605.3300830.

Jeff Horwitz. 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, October 2020.

Jeff Horwitz and Deepa Seetharaman. 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, May 2020.

Md Sajjad Hosain. Integration of social media into hrm practices: a bibliometric overview. *PSU Research Review*, April 2021.

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022. doi: 10.1073/pnas.2025334119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2025334119.

Basileal Imana, Aleksandra Korolova, and John Heidemann. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021*, pp. 3767–3778, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. URL https://doi.org/10.1145/3442381.3450077.

Basileal Imana, Aleksandra Korolova, and John Heidemann. Having your privacy cake and eating it too: Platform-supported auditing of social media algorithms for public interest. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1), 04 2023. doi: 10.1145/3579610. URL https://doi.org/10.1145/3579610.

Jae C. Jung and Elizabeth Sharon. The Volkswagen emissions scandal and its aftermath. *Global Business and Organizational Excellence*, 38(4):6–15, 2019. doi: https://doi.org/10.1002/joe.21930. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/joe.21930.

Kalley Huang. For Gen Z, TikTok Is the New Search Engine. https://www.nytimes.com/2022/09/16/technology/gen-z-tiktok-search-engine.html, September 2022.

Simon Kemp. Digital 2021: Global overview report. https://datareportal.com/reports/digital-2021-global-overview-report, 2021.

Roisin Kiberd. 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, October 2021.

Gary King and Nathaniel Persily. 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, February 2020.

Nicolas Koumchatzky and Anton Andryeyev. 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.

Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of STEM career ads. *Management Science*, 65(7):2966–2981, 2019.

Mathias Lecuyer, Riley Spahn, Yannis Spiliopolous, Augustin Chaintreau, Roxana Geambasu, and Daniel Hsu. Sunlight: Fine-grained targeting detection at scale with statistical confidence. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, pp. 554–566, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450338325. doi: 10.1145/2810103.2813614. URL https://doi.org/10.1145/2810103.2813614.

Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. The privacy policy landscape after the gdpr. *Proceedings on Privacy Enhancing Technologies*, 2020(1):47–64, 2020. doi: doi:10.2478/popets-2020-0004. URL https://doi.org/10.2478/popets-2020-0004.

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, October 2021.

Edward Markey and Doris Matsui. Algorithmic justice and online platform transparency act of 2021. https://www.markey.senate.gov/imo/media/doc/ajopta.pdf, 2021.

J. Nathan Matias, Austin Hounsel, and Nick Feamster. 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.

Jeremy B. Merrill and Ariana Tobin. Facebook moves to block ad transparency tools — including ours. https://www.propublica.org/article/facebook-blocks-ad-transparency-tools, January 2019.

Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends® in Human–Computer Interaction*, 14(4):272–344, 2021. ISSN 1551-3955. doi: 10.1561/1100000083. URL http://dx.doi.org/10.1561/1100000083.

Ali Mohamed and Zheng Li. Community-focused feed optimization. https://engineering.linkedin.com/blog/2019/06/community-focused-feed-optimization, 2021.

Pavithra Mohan. Facebook's news feed just turned 10. https://www.fastcompany.com/4018352/facebooks-news-feed-just-turned-10, September 2016.

Adam Mosseri. Shedding More Light on How Instagram Works. https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works, 2021.

Ioannis Nikolaou. Social networking web sites in job search and employee recruitment. *International Journal of Selection and Assessment*, 22(2):179–189, 2014. doi: https://doi.org/10.1111/ijsa.12067. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ijsa.12067.

Brandie Nonnecke and Camille Carlton. EU and US legislation seek to open up digital platform data. *Science*, 375(6581):610–612, 2022. doi: 10.1126/science.abl8537. URL https://www.science.org/doi/abs/10.1126/science.abl8537.

NYU Ad Observatory. Explore Facebook political ads. https://adobservatory.org/.

Orestis Papakyriakopoulos, Christelle Tessono, Arvind Narayanan, and Mihir Kshirsagar. How Algorithms Shape the Distribution of Political Advertising: Case Studies of Facebook, Google, and TikTok. *arXiv preprint arXiv:2206.04720*, 2022.

Nathaniel Persily. A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act. *Journal of Online Trust and Safety*, 1(1), October 2021. doi: 10.54501/jots.v1i1.22. URL https://tsjournal.org/index.php/jots/article/view/22.

Newley Purnell and Jeff Horwitz. 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, October 2021.

Karl Racine. Stop discrimination by algorithms act of 2021. https://oag.dc.gov/sites/default/files/2021-12/DC-Bill-SDAA-FINAL-to-file-.pdf, 2021.

Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now*, 2018.

Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pp. 131–141, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372879. URL https://doi.org/10.1145/3351095.3372879.

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. Auditing algorithms : Research methods for detecting discrimination on internet platforms. In *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 2014.

Sarah Perez. Google exec suggests Instagram and TikTok are eating into Google's core products, Search and Maps. https://techcrunch.com/2022/07/12/google-exec-suggests-instagram-and-tiktok-are-eating-into-googles-core-products-search-and-maps/?tpcc=tcplustwitter, July 2022.

Shahar Segal, Yossi Adi, Benny Pinkas, Carsten Baum, Chaya Ganesh, and Joseph Keshet. Fairness in the eyes of the data: Certifying machine-learning models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. doi: 10.1145/3461702.3462554. URL https://doi.org/10.1145/3461702.3462554.

Elisa Shearer and Jeffrey Gottfried. News Use Across Social Media Platforms 2017. *Pew Research Center*, September 2017. doi: 20.500.12592/255tqv. URL https://policycommons.net/artifacts/617725/news-use-across-social-media-platforms-2017/1598576/.

Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), October 2021. doi: 10.1145/3479577. URL https://doi.org/10.1145/3479577.

Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of Machine Learning Research*, 2018. URL http://proceedings.mlr.press/v81/speicher18a.html.

Latanya Sweeney. Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 2013. ISSN 1542-7730. doi: 10.1145/2460276.2460278. URL https://doi.org/10.1145/2460276.2460278.

The Markup. The Citizen Browser Project—Auditing the Algorithms of Disinformation. https://themarkup.org/citizen-browser, 2020.

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, June 2022.

TikTok. Relevance is the new reach. https://www.tiktok.com/business/en-US/blog/relevance-is-the-new-reach, September 2021.

Craig Timberg. 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/, September 2021.

Twitter. The Twitter ads auction. https://business.twitter.com/en/help/troubleshooting/bidding-and-auctions-faqs.html, October 2021.

Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017. doi: 10.1177/2053951717743530. URL https://doi.org/10.1177/2053951717743530.

Mathias Vermeulen. The keys to the kingdom. https://knightcolumbia.org/content/the-keys-to-the-kingdom, July 2021.

Udi Weinsberg, Smriti Bhagat, Stratis Ioannidis, and Nina Taft. BlurMe: Inferring and obfuscating user gender based on ratings. RecSys '12, pp. 195–202, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450312707. doi: 10.1145/2365952.2365989. URL https://doi.org/10.1145/2365952.2365989.

Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and auditing fair algorithms: A case study in candidate screening. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. doi: 10.1145/3442188.3445928. URL https://doi.org/10.1145/3442188.3445928.

YouTube. YouTube researcher program. https://research.youtube/, July 2022.

A LIMITATIONS OF EXISTING AUDITING METHODS

We expand on the limitations of existing auditing methods that we describe in §2.2.

Confounding variables: The first challenge is controlling for 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 may drive differences in ad delivery to individuals from different demographic groups, such as different levels of market competition from other ads for members of different groups, as well as differences in platform's use or interaction patterns among users from various demographics. An external auditor aiming to isolate the role of the relevance estimator for differences in outcomes between demographic groups must control for such factors. Designing auditing methods with such controls in place, however, is a laborious process that requires careful reasoning and creative hacks. In particular, it took many years of research effort to get from Sweeney's study that gave the first evidence of biased ad delivery in 2013 (Sweeney, 2013) to Ali and Sapiezynski et al.'s study that attributed such bias to the role the platform's algorithms play in 2019 (Ali et al., 2019), and the Imana et al.'s 2021 study (Imana et al., 2021) 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 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 (Bartley et al., 2021). However, the study identifies the duration sock-puppet accounts stay logged-in for and the timeline scrolling capabilities as potential confounding factors that could possibly alter the conclusions (Bartley et al., 2021). Even Twitter's internal audit of disparate algorithmic amplification of political content, for political right compared to political left, shows the limits of current methods (Huszár et al., 2022). 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).

Opacity to end-users: Another challenge is that the effects of platform algorithms are often not obvious or visible to end-users. Collaborative methods that rely on end-users' day-to-day experiences may not be able to detects harms that are invisible or unnoticeable to users (DeVos et al., 2022; Shen et al., 2021). We discuss this limitation in more detail in Appendix G.

Reliance on proxies: A third 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 (Ali et al., 2019; 2021; Imana et al., 2021). However, such workarounds introduce measurement errors (Imana et al., 2021) 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, we carried out a study aiming to ascertain whether job ad delivery algorithms are discriminatory that built upon Ali and Sapiezynski et al.'s work, but adding new controls for job qualifications across genders that required additional knowledge about gender composition of current employees of several companies (Imana et al., 2021). The use of additional data on employers and the gender of their employees means this method does not directly generalize to auditing for discrimination in ad delivery of other types of ads (for example, housing ads) and along other demographic attributes

(such as 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 (Ali et al., 2019), and use of public data sources such as voter data to build audiences with a specific demographic make up (Speicher et al., 2018; Ali et al., 2019; 2021). Such public data sources are extremely limited and subject individuals to participation in experiments without their knowledge.

On the other hand, crowdsourced audits that rely on browser extensions do not easily generalize beyond desktop versions of platforms, a significant limitation to their applicability given that most people today access social media through their phones. For example, 98.3% of Facebook users access it using a phone app (Kemp, 2021). 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: Finally, existing external auditing methods can also incur high costs in terms of both 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 confounding factors (Ali et al., 2019; Imana et al., 2021). 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".

A.1 Platforms Beginning to Favor Platform-supported Audits?

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 (Bartley et al., 2021). Facebook has resisted external auditing by explicitly blocking accounts used to conduct audits (Clark, 2021), tweaking its APIs to break auditing tools (Merrill & Tobin, 2019), and threatening legal actions against researchers who scrape data from its platform (Horwitz, 2020).

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.

Platforms are establishing programs to provide vetted researchers with access to their data and algorithms. 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 (Austin, 2022; The US Department of Justice, 2022). 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. Other platforms such as Youtube, Facebook and Twitter have also recently started initiatives to provide data to external researchers (YouTube, 2022; Facebook, 2022; Chowdhury & Williams, 2021). 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.

Current data available to researchers through these programs are limited to public data corpora, such as public videos, pages, posts and comments (YouTube, 2022; Facebook, 2022). 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.

B TRUST MODEL AND LIMITATIONS OF OUR FRAMEWORK

In this section, we expand on the trust model of our framework that we briefly described in §4.4.

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 (Speicher et al., 2018), Facebook ad accounts used to audit ad delivery (Ali et al., 2019; 2021; Imana et al., 2021), and browser extensions used for collecting data from Facebook (NYU Ad Observatory; The Markup, 2020).

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 (Austin, 2022). For example, Volkswagen faced significant legal and financial repercussions as a result of their violation of emissions regulations (Jung & Sharon, 2019).

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 (Timberg, 2021). This preventable error undermined academic work that was based on the data (Timberg, 2021), 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 (Graham, 2021).

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 §A.1, 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 (Coons et al., 2021; European Commission, 2022b). 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 (Coons et al., 2021; European Commission, 2022b). Platform-initiated transparency efforts such as Facebook's FORT, Social Science One, and YouTube's Researcher Program also all have approaches for vetting researchers (King & Persily, 2020; Facebook, 2022; YouTube, 2022). 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.

C SAMPLE SIZE REQUIRED FOR AUDITING RELEVANCE ESTIMATORS WITH PRIVACY

In §5, we informally stated our main technical result that shows it is feasible to both audit for fairness *and* protect user privacy and platforms' business interests. We next formally state our result and prove our claim.

C.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 job qualifications in employment ads. 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 the relevance scores a platform's relevance estimator will assign to 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 (Ali et al., 2019; Imana et al., 2021).

C.1.1 SETUP AND DEFINITIONS:

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 = {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 1, the auditor first queries the platform's relevance estimator using each subset S_a and ad 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 (Hardt et al., 2016).

Definition C.1 (derived from Equality of Opportunity (Hardt et al., 2016))**.** A relevance estimator function R satisfies equality of opportunity:

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

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 (Hardt et al., 2016). 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 (Segal et al., 2021), 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 C.2 (α-fairness (Segal et al., 2021))**.** 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 (Dwork et al., 2006). 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 (Segal et al., 2021) 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 C.3 ((α,δ)-fairness (Segal et al., 2021))**.** We define R to be (α,δ)-fair with high probability with respect to S if:

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

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

Definition C.4 ((α,δ,ϵ)-fairness)**.** We define R to be (α,δ,ϵ)-fair with respect to S where an ϵ-DP mechanism is applied to outputs of R if:

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

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:

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

C.1.2 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 C.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 2 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. Examples from prior work include repeating audits on various audience partitions, and varying locations that users are chosen from (Ali et al., 2019; Imana et al., 2021).

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 (The Markup, 2020)). We therefore place this problem outside the scope of this paper.

C.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 2. 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 C.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 (Segal et al., 2021). In Appendix D, 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 C.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 D):

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

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 F). 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 2 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}:

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

□

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 C.3. Compared to the non-private case (Lemma C.1), at most 6.34 times as many samples are needed to perform the audit with differential privacy guarantees (Lemma C.2)).

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

We prove the above lemma in Appendix E. While this is a strict upper bound, for reasonable auditing parameters, the overhead is much lower, around 4. Figure 2 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, which are known to provide reasonable privacy guarantees (Dwork & Roth, 2014).

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 (EEOC, 2018). 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 (Ali et al., 2019; Imana et al., 2021), and is at the same order of magnitude achieved by current opt-in crowdsourcing efforts (The Markup, 2020).

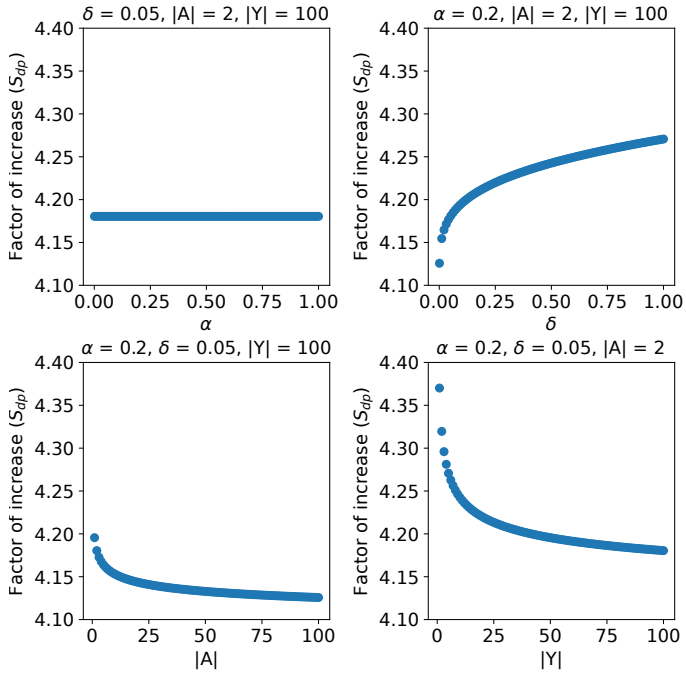

Figure 2: Relationship between auditing parameters (α, δ, $|A|$, and $|Y|$) and the factor of increase in sample size.

D Proof Number of Samples Required for Auditing Without Privacy

Lemma C.1 references a proof for a lower bound for number of samples required for auditing without privacy. Our proof follows Segal et al.'s work (Segal et al., 2021), 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 C.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:

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

□

E Upper-bound on Increase in Number of Samples

In this appendix we give a proof for Lemma C.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:

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

□

F Applying Hoeffding's in the Presence of Potential Bias

In Appendix C, 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 (Hoeffding, 1963).

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 $S_{a,q}$.

$$
\begin{aligned}
E[\overline{P}_{a,y}(R,S)] &= E\left[\frac{1}{n_{a,q}}\sum_{i=1}^{|S|}\mathbb{1}\{R(x_i)=y \wedge a_i=a \wedge q_i=1\}\right] \\
&= E\left[\frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|}\mathbb{1}\{R(x_i)=y\} + \sum_{i=1}^{|\overline{S_{a,q}}|}0\right] \quad \text{.....separate } S_{a,q} \text{ and } \overline{S_{a,q}} \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|}E\left[\mathbb{1}\{R(x_i)=y\}\right] \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|}0 * Pr[R(x_i)\neq y] + 1 * Pr[R(x_i)=y] \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|}Pr[R(x_i)=y] = \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|}Pr[T(x_i)+b_a=y] \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|}Pr[T(x_i)=y-b_a] \\
&= \frac{1}{n_{a,q}}\sum_{i=1}^{|S_{a,q}|}P_{a,y-b_a}(T) \qquad \text{.....by i.i.d. assumption} \\
&= \frac{n_{a,q}}{n_{a,q}}P_{a,y-b_a}(T) \\
&= P_{a,y-b_a}(T) \\
&= P_{a,y}(R) \qquad \text{.....plug in Equation 14 in Equation 1}
\end{aligned}
$$

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

$$R(x) = T(x) * b_a \tag{15}$$

where b_a is a constant.

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

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:

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

$$
\begin{aligned}
P_{a,y}(R) &= Pr_{(x,a',q)\sim X}\left[R(x) = y | a' = a \wedge q = 1\right] \\
&= Pr_{(x,a',q)\sim X}\left[T(x) + B_a = y | a' = a \wedge q = 1\right] \\
&= \sum_b Pr[B_a = b] * Pr_{(x,a',q)\sim X}\left[T(x) = y - b | a' = a \wedge q = 1\right] \\
&= \sum_b Pr[B_a = b] * P_{a,y-b}(T) \qquad (17)
\end{aligned}
$$

Since Equation 16 and Equation 17 are equal, $E[\overline{P}_{a,y}(R,S)] = P_{a,y}(R)$.

G 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 (Sandvig et al., 2014). 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 (Bandy, 2021). 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 (Shen et al., 2021). 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 (Matias et al., 2021). 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 (Reisman et al., 2018). 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 (Metaxa et al., 2021). 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 (Wilson et al., 2021). 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 (Sweeney, 2013) was the first to hypothesize that platform-driven decisions can lead to discriminatory ad delivery; a hypothesis strengthened by evidence from subsequent works (Datta et al., 2015; Gelauff et al., 2020; Lambrecht & Tucker, 2019). 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 (Ali et al., 2019). 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 (Imana et al., 2021). 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 (Hargreaves et al., 2018). 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 (Bartley et al., 2021). 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 (Huszár et al., 2022). 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 (Segal et al., 2021). 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 (Holstein et al., 2019) and later by Andrus (Andrus et al., 2021) 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 (Andrus et al., 2021; Holstein et al., 2019). 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 (Bogen et al., 2020b). 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 (Alao et al., 2021; Austin, 2021). 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 (Veale & Binns, 2017). 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.