Enforcing Demographic Coherence: A Harms Aware Framework for Reasoning about Private Data Release

Anonymous author(s)

Abstract

The technical literature about data privacy largely consists of two complementary approaches: formal definitions of conditions 'sufficient' for privacy preservation and attacks that demonstrate privacy breaches. Differential privacy is an accepted standard in the former sphere. However, differential privacy's powerful adversarial model and worst-case guarantees may make it too stringent in some situations, especially when achieving it comes at a significant cost to data utility . Meanwhile, privacy attacks aim to expose real and worrying privacy risks associated with existing data release processes but often face criticism for being unrealistic. Moreover, the literature on attacks generally does not identify what properties are necessary to defend against them.

We address the gap between these approaches by introducing *demographic coherence*, a condition inspired by privacy attacks that we argue is necessary for data privacy. This condition captures privacy violations arising from inferences about individuals that are incoherent with respect to the demographic patterns in the data. Our framework focuses on confidence rated predictors, which can in turn be distilled from almost any data-informed process. Thus, we capture privacy threats that exist even when no attack is explicitly being carried out. Our framework not only provides a condition with respect to which data release algorithms can be analysed but suggests natural experimental evaluation methodologies that could be used to build practical intuition and make tangible assessment of risks. Finally, we argue that demographic coherence is weaker than differential privacy: we prove that every differentially private data release is also demographically coherent, and that there are demographically coherent algorithms which are not differentially private.

Contents

1 Introduction

[1] Inspired by high-profile demonstrations of the limitations of heuristic approaches to data privacy [55, 41, 10, 9, 12], there is an emerging consensus among researchers and practitioners that formal data privacy approaches are a necessary component of responsible data release pipelines. Differential privacy (DP) [25] in particular has become the accepted standard guarantee, seeing deployment in both industry [28, 5, 22, 57] and government [1]. Formal approaches are often treated as *sufficient* for ensuring data subject's privacy, in that they are "one-size-fits-all," *i.e.,* data curators can apply best practice protections without needing to consider the intricacies of each deployment.

In practice, however, there are significant barriers to making most formal privacy notions an easy "apply and forget" process, stemming from the fact that, unlike encryption, the privacy guarantees are not binary and require striking a delicate balance between the benefits of privacy preservation and the cost to accuracy [4]. Within the context of DP in particular, this tension is represented by the selection of parameters ε and δ which modulate the accuracy-privacy tradeoff. Reasoning about this tradeoff is a central challenge in the deployment of DP, in no small part due to the fact that it can be unclear what the necessary level of privacy is and what a failure of DP would even mean. Moreover, the generality of DP means that it is naturally very abstract, making it easy to lose sight of the concrete privacy harms against which DP is intended to protect [17]. This confluence of factors can give rise to a common critique of DP: the lack of clear guidance during parameter selection makes it unclear what "good" parameter choices are, leading some practitioners to make choices that are seen as overly conservative while others make overly permissive choices that fail to concretely protect against concerning attacks.

To more clearly illustrate the problem, consider a data curator charged with managing a privatized data release. Having determined that DP is the "right" guarantee for their setting, they must now pick parameters for a data release that maximize accuracy without risking harm to data subjects. Looking at existing deployments, they will find a wide variety of parameter choices in existing deployments, offering little direct guidance; e.g., Khavkin and Toch [42] document industry deployments with ε choices between 0.1 and >10 (recall that ε is in the exponent of the privacy-accuracy tradeoff). In the hopes of making the harms against which they must protect more concrete, the curator might study the known attack methodologies [55, 41, 10, 9, 12] and the results of testing concrete DP mechanisms against state-of-the-art attack implementations [38, 37, 46, 54]. While this highlights the risks of data sharing in general, and testing results may justify the deployment of specific DP mechanisms, they do not indicate what is *necessary* to defend against specific attack methodologies beyond such indirect arguments. Moreover, recontextualizating these attacks to their own context can be challenging (e.g., *What axillary information do adversaries have? What are the harms associated with failure?*) and there is no promise that attacks will not improve with time. As such, it is not clear if the results of these experiments provide actionable guidance within the curator's context or how to operationalize these result if the data contained within a data release will remain meaningfully sensitive for a long time. Left without any other support, the curator will inevitably be reduced to making a heuristic decision, possibly choosing to be highly conservative in order to account for uncertainty (sacrificing accuracy) or defaulting to loose parameters that maximize accuracy. Note that while we described this process for DP, these challenges will inevitably arise for any privatization method with a privacy-accuracy tradeoff.

Necessary privacy conditions. Caught between abstract privacy conditions and the heuristic nature of existing attack work, we explore an intermediary design philosophy: *necessary* conditions. Within this approach, we can formally define a set of properties that any private data release should

[1] All proofs of theoretical claims and the code used in this paper can be found in this repository.

guarantee without needing to provide a single, unifying, sufficient condition. These necessary conditions can be seen as giving formal procedures for recognizing when an attacker has inflicted a particular type of harm. Critically, because the necessary conditions need not be all-encompassing, they can be more easily contextualized allowing for more clear application of domain experts' intuition about acceptable levels of risk. At the same time, specifying necessary conditions with mathematical rigor (and universal quantifiers, in particular) embraces the formality of sufficient conditions and supports rigorous analysis.

Of course, thinking in terms of necessary conditions has always been implicit in the practice of parameter selection, albeit, usually informally. In the process described above, the data curator might have leveraged the literature on attacks to set an internal understanding of some necessary conditions. For example, the curator might identify membership inference [33] as an attack against which they would want to protect. The existing literature (e.g., [16, 17, 54, 33, 27, 26]), however, does not provide much support as the curator translates this implicit intuition into decisions. By giving this curator a set of explicit and formal necessary conditions, we can hope to generate a concrete methodology for justifying parameter choices by identifying parameter regimes that could enable specific harms.

A new necessary condition: Demographic Coherence. We take a step towards this reality by developing a non-trivial, new necessary condition. Our approach is rooted in three key insights: (1) privacy harms are increasingly going to come in the form of inferences at the hands of predictive algorithms.[2] That is, we should be interested in the predictions that these algorithms make about people—and the decisions organizations may make based on these predictions—even when predictive algorithms are not intentionally designed with causing harm in mind; (2) we should consider the confidence with which an algorithm can make predictions, because simply increasing the *confidence* that an individual or a group has a certain attribute may be enough to result in harm; and (3) the harms associated with breaches of privacy are not experienced uniformly among members of a population. This means that, if not defined carefully, an aggregate measure across an entire population could easily "hide" effects on vulnerable subgroups by averaging them away.

The resulting notion, which we call *demographic coherence*, is intentionally designed to be *ergonomic*[3] in many different contexts. For example, we provide sufficient formalism to enable rigorous analysis and provable realization, all while keeping the specific harms against which demographic coherence protects compellingly salient. Additionally, we provide a vision as to how demographic coherence can support the type of intuition building required to set real-world parameters. Critically, as a necessary condition, we imagine demographic coherence as one of a set of conditions which a data curator might want to jointly consider when making deployment decisions.

1.1 Our Contributions

In this work we make the following contributions:

- **Demographic Coherence.** In this work we introduce *demographic coherence*, an analytical framework for reasoning about the privacy provided by data release algorithms. Demographic coherence has the following qualities:

[2] In this work we intentionally us the term "algorithm" broadly to capture, *e.g.,* informal decision-making process made by humans that might not be explicitly codified as algorithms in the traditional sense.

[3] We use ergonomic in this context to mean ease of use by many different stakeholders. We intentionally move away from the term "usable," as this typically focuses only on end-users and we are interested in ease of use from a more diverse set of communities.

  - *Captures predictive harms.* Demographic coherence builds on conceptual tools from generalization [59, 24, 18, 35, 8] and multicalibration [31, 43] to (1) evaluate the risk of predictive harms distributionally without relying on measuring accuracy with respect to an unknown (and possibly unknowable) ground truth and (2) evaluate the risk of predictive harms local to the different subgroups within a population. Evaluating risks distributionally allows the framework to remain applicable even when ground truth is unavailable, and evaluating risks for different subgroups allows the framework to identify effects specific to vulnerable subgroups.
  - *Lends itself to experimental auditing.* Demographic coherence has a natural translation to an experimental setup for comparing the effects of various algorithms for privacy preserving data release. In addition, demographic coherence is measured by computing a distance metric over two distributions, which facilitates quantification of the concrete risk. We focus on demographic coherence when instantiated with Wasserstein distance.
  - *Lends itself to analytical arguments.* Finally, the formalism we build supports rigorous analytical arguments about algorithms. For example, we show that all algorithms with bounded max-information are also coherence enforcing.
- **Demographic coherence enforcement is achievable.** We prove that demographic coherence enforcement is achievable, showing parameter conversions under which any pure differentially private (pure-DP) algorithm and any approximate differentially private (approx-DP) algorithm enforce demographic coherence. See Section 2 for an overview.

1.2 Related Work

The study of privacy-preserving data release broadly falls into two categories: demonstration of potential harms via concrete attacks, and the development of formal methodologies that provide robust guarantees. These two approaches provide complementary insights. Formal approaches provide a concrete path to implementing privacy-protections, and the motivation for their use is derived from attacks. In particular differential privacy provably protects against membership inference (*e.g.,* [33, 27, 53, 60]), reconstruction (*e.g.,* [23, 14, 30, 11]), and reidentification (*e.g.,* [55, 44]), as shown by Dwork et. al. [26]. In practice, however, there are fundamental challenges in using attacks to guide the many choices one must make when implementing privacy protections. These challenges arise from (1) identifying successful attacks, (2) identifying realistic attacks, and (3) determining the privacy protections *necessary* to prevent the attacks being considered.

Evaluating the success of an attack. In using attacks to motivate formal methodologies, one must start by demonstrating the extent of potential vulnerabilities. For example, membership inference is an attack that relates directly to the definition of differential privacy—however, the potential to infer membership in a dataset isn't a convincing vulnerability in the case of large data collection efforts like the US decennial Census. Therefore, differential privacy frequently derives its motivations from re-identification and reconstruction attacks. Still, the success of these attacks is difficult to evaluate.[4] In recent work, Dick et al. implemented a reconstruction attack [21] along with robust evaluations of its success. Their work has since been cited by the US Census Bureau's chief scientist as evidence that "database reconstruction does compromise confidentiality" [40]. The key insight in their evaluation comparing the results of the reconstruction to a baseline in which reconstruction is conducted with complete access to the distribution underlying the data. While the

[4] For example, reconstruction of features like gender can be carried out simply from knowing population statistics rather than breaking anonymity [51].

intuition behind this work—that an attack is much more concerning if it reveals more than what could be learned from a detailed knowledge of the distributional properties—applies to many attack paradigms, the baseline considered in their work is specific to reconstruction attacks. Reconstruction attacks are not always possible and conducting a reconstruction attack assumes malicious intent. Our framework which extends this intuition to the evaluation of a more general class of attacks.

Another place where the efficacy of specific attacks is measured via comparison to baselines is the literature on auditing differentially private algorithms (*e.g.,* [38, 37, 46, 54]). Here, attacks are carried out on existing systems, and the efficacy of the attack is used to measure the maximum level of "effective privacy" that the system confers.

Identifying realistic attacks. Research into conducting privacy attacks makes a variety of assumptions about the setting in which those attacks could be conducted, including the goal of the attacker, the power of the attacker, the type of system attacked, etc... These assumptions can radically change the extent to which an attack should be considered a realistic threat against real-world data releases; attacks that require unrealistic assumptions may not be concrete threats. The works of Rigaki & Garcia [49], Salem et al. [52], and Cummings et al. [17] classify existing attack strategies by adversarial resources and goals in order to provide a structure for evaluating privacy risks. In addition to this, Cohen [13] and Giomi et al. [29] take a different approach, appealing to the law to determine the goals of a realistic attacker. Specifically, they contextualize the attacks they consider by tying them to existing privacy law. Still, individual attacks, even if successful and realistic, don't provide a clear path forward in terms of designing protections.

Identifying necessary conditions. Some prior work has started to identify *necessary* conditions for achieving privacy. Cohen & Nissim [15] introduce a necessary condition, called "predicate singling out," inspired by the GDPR notion of singling out. Balle et. al. [6] introduce an alternative necessary condition called "reconstruction robustness," which is closely related to reconstruction attacks. Cummings et. al. [17] build on the notion of reconstruction robustness, extending it to a weaker adversarial setting. Our framework extends this general approach but applies to a much broader class of attacks—namely, any attacks from which a confidence rated predictor could be distilled.

Recent work by Cohen et al. [16] also recognizes the need to bridge the gap between formal privacy guarantees and practical attacks. Building on definitions in prior work [6, 15, 17] they introduce "narcissus resiliency," a framework for establishing precise conditions under which an algorithm prevents various classes of existing attacks, including reconstruction attacks, singling out attacks, and membership inference attacks. Our definition defines invulnerability against a different type of privacy loss, providing complementary insights in the form of necessary conditions that can be considered alongside their definitions. Specifically, we believe that it is important to consider demographic coherence alongside their notion of narcissus singling out; the latter captures an important property that the former does not. (An algorithm that chooses a small subset of the data to publish in the clear does not meet the definition of *singling out security* even though it may be *demographic coherence enforcing* if the subset is small enough.) Another key difference between our works is that the narcissus framework does not naturally lend itself to concrete experimental evaluation, whereas demographic coherence is intentionally designed with this use case in mind.

Finally, most of the works discussed above measure the success of an attack via its accuracy (*i.e.,* is the information extracted about the data subject *true*?). We observe that harm is not necessarily predicated on accuracy, and we design demographic coherence to be intentionally independent of accuracy. One impact of this choice is that demographic coherence is a more natural fit for settings in which ground truth is difficult or impossible to measure.

2 Overview of Technical Results

In Section 6, we show parameter conversions under which any pure-DP algorithm, and any approx-DP algorithm enforces demographic coherence. Here, we present informal statements of these technical results.

We start by presenting a simplified definition of *coherence enforcement* (Definition 3, 4). (This presentation is meant to allow the informal statements of our technical results. For a formal presentation, see Section 4. Additionally, the concept of enforcing demographic coherence emerges from careful consideration of several key principles, which are discussed in detail in Section 3.) Informally, a coherence-enforcing $\mathcal{A}$ guarantees that predictors trained using its private reports will be demographically coherent.[5]

Definition 1 (Informal Version of Definitions 3 and 4)**.** *Consider a data universe $\mathcal{X}$ and a data-curation algorithm $\mathcal{A} : \mathcal{X}^* \to \mathcal{Y}$. We say that $\mathcal{A}$ enforces (α, β)-demographic coherence, if for all algorithms $\mathcal{L} : \mathcal{Y} \to (\mathcal{X} \to [-1, 1])$ that use the report produced by the curator to create a confidence-rated predictor $h : \mathcal{X} \to [-1, 1]$, the following condition is satisfied. For all datasets X,*

$$\Pr_{\substack{X_a, X_b \xleftarrow{\$} X, \\ R_a \leftarrow \mathcal{A}(X_a), h \leftarrow \mathcal{L}(R_a)}} [\mathsf{dist}(h(X_a), h(X_b)) \geq \alpha] \leq \beta,$$

where X_a, X_b represent a random split of the dataset X into halves, report R_a is produced by the data-curator using only X_a, and h is created by running algorithm $\mathcal{L}$ on the report, $h(X_a)$ represents the empirical distribution of predictions made on X_a, and $\mathsf{dist}(\cdot, \cdot)$ represents a metric distance between distributions. Here, β is the probability that h is not demographically coherent, and α represents how close the distributions of $h(X_a)$ and $h(X_b)$ are required to be.

The formal definition of *coherence enforcement* is more intricate than the one above. One key technical distinction is that the restriction on predictor h applies not only to the full sets X_a and X_b, but also across different subpopulations in those sets. For the remainder of this section, we specify the distance metric $\mathsf{dist}(\cdot, \cdot)$ as Wasserstein-1 distance between distributions. In this context, we say that and algorithm $\mathcal{A}$ *enforces Wasserstein-coherence.*

Theorem 1 is an informal statement of Theorem 3, which argues that any data-curation algorithm with bounded max-information [24] (a notion that mathematically captures the dependence of algorithms' outputs to their inputs) also enforces Wasserstein coherence.

Theorem 1 (Informal Version of Theorem 3)**.** *Let $n \in \mathbb{N}, \zeta > 0$, $\beta \in (0, 1)$, $\alpha \in (0, 2]$.*

Consider a data curation algorithm $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$ with bounded max-information

$$I_{\infty}^{\beta/2}(\mathcal{A}, n/2) < \zeta.$$

Then, $\mathcal{A}$ enforces (α, β)-demographic coherence provided that $n \geq k \cdot \frac{\zeta + \ln(1/\beta)}{\alpha^2}$ for some constant k.

We leverage the connection between differential privacy and max-information to show the exact parameter conversion under which differentially private algorithms enforce demographic coherence. Theorem 2 is an informal statement of Theorem 7, the result for pure-DP. For the approximate-DP result, we point the reader to Theorem 8 in Section 6.

[5]In reality, the property of demographic coherence applies to algorithms $\mathcal{L}$ that use private reports to design predictors.

Theorem 2 (Informal Version of Theorem 7). *Let $n \in \mathbb{N}$, $\beta, \varepsilon \in (0,1)$, $\alpha \in (0,2]$.*

Consider an ε-DP data curation algorithm $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$. Then, $\mathcal{A}$ enforces (α, β)-demographic coherence provided that $\varepsilon \leq k \cdot \frac{\alpha}{\ln(1/\beta)}$ for some constant k.

This theorem should be understood as follows: a data curator identifies (possibly experimentally) regimes for α and β that they find to be "too risky" for a data release (with respect to demographic coherence). That curator can then use this theorem to suggest a value of ε such that, if they were to use differential privacy as their privatization mechanism, the resulting data release would achieve their desired constraints. While the parameter conversion in Theorem 7 would likely result in a value of ε that is too small for most use cases, we expect this to be inherent to a black-box conversion of differential privacy to the enforcement of demographic coherence. An important open question is identifying other ways of achieving our definition, including non-black-box uses of DP-algorithms for obtaining better coherence enforcement guarantees.

3 A Walk Through Our Definition

We now slowly build up our definition in order to clearly explain the choices embedded within it.

Notation and Conventions. Assume that the data curator has collected a sample X from the overall population of interest. We make no requirements on the relative sizes of X and the population such that our framework can be used broadly—even when the goal is to sample the entire population, as in censuses. Our ultimate goal is to reason about a data curator $\mathcal{A}$ who uses X to generate a privacy-preserving release R.[6]

3.1 Predictive Harms

Within our framework, we characterize the adversary as a party interested in making predictions about individuals, *e.g.,* if people have some particular stigmatized feature or are going to buy a product if they are targeted with an advertisement. We formalize this conceptual approach by considering an arbitrary algorithm $\mathcal{L}$ used by the adversary to design the predictor h.[7] We choose to measure privacy risk in terms of predictive harms for the following reasons:

- *Predictors are commonplace:* The predictions made by machine learning models increasingly have direct impacts on people's daily lives. Diagnostic models are being tested as potential aides for medical experts [3], and increasingly complex and opaque models are used to "match" job candidates with prospective employees [36, 2, 47] in order to increase the odds that an individual ends up with a lucrative job. Even complex infrastructures, like those used in digital advertising, can be seen as predictive models that are attempting to classify individuals into target audiences. The fact that predictive algorithms are increasingly commonplace—and the decisions they make concretely impact our daily lives—makes them a very *believable* source of harm.

- *Harmful predictors need not be maliciously produced:* By considering the impact of predictors, we free ourselves from needing to see the adversary as *intentionally* trying to cause harm and instead can refocus on the (perhaps accidental) harms that a data release has the potential to cause.

[6] In our formal experiment, we actually suppress the formal object of the report. Specifically, we reason directly about the composition of some data processing algorithm $\mathcal{A}$ with an arbitrary algorithm $\mathcal{L}$, rather than making the report an explicit object that is then passed to $\mathcal{L}$.

[7] A more complex interpretation could also cover an approach to prediction that takes into account a social decision-making process. While this interpretation is beyond the technical scope of our work, it may be interesting to consider in future work.

A conceptual concern about considering predictive harms is whether we are explicitly ruling out particular, important types of adversarial behavior that are attempting to extract information (*e.g.,* reconstruction attacks). However, we note that by discussing predictors we are only limiting the input/output behavior of the adversary's product, and not how the predictor is produced. For example, our framework could capture an adversary that runs a known reconstruction algorithm (*e.g.,* [20]) and then makes predictions about individuals based on the produced table. In this way, our approach highlights the ways in which existing approaches could be *used* when applied in decision-making contexts. Looking ahead, in theoretically analyzing our framework we will *universally quantify* over algorithms to preserve generality, which means that reconstruction-based approaches—or other known malicious approaches to data extraction—are naturally captured. Still, in choosing to concretize the type of our adversary, we do risk failing to consider a *different* type of attacker with inconsistent goals. Definitions created with this philosophy can be extremely helpful at establishing *necessary* conditions for ensuring privacy, but do not claim to be *sufficient.* On the other hand, our approach helps highlight a specific way in which data is likely to be weaponized in the real world.

Incoherent predictions. Within this work we focus on capturing predictive harms that occur specifically by virtue of a data subject appearing in a dataset. To give a concrete example, consider the now classic case of Narayanan and Shmatikov's re-identification of Netflix users within an anonymized data release using public IMDB data [45]. In this setting, we might consider an adversary interested in learning a predictor which predicts queerness (*e.g.,* imagine the adversary is operating in a regime in which queerness is criminalized or highly stigmatized). Now, imagine two similar[8] individuals Asahi and Blair; each intentionally avoids being perceived as queer, and in particular does not provide ratings on movies with queer themes on their public IMDB profiles. Assume that based on a random sample, one of them (*e.g.,* Asahi) has their movie ratings released by Netflix and the other (*e.g.,* Blair) does not. A predictor that is likely to guess that Asahi is queer when they are present in the dataset but would not have guessed they are queer otherwise indicates that the predictor was able to extract some information about Asahi from the data release. Given that Asahi and Blair are similar, this would also be true of a predictor that guessed that Asahi is queer while Blair is not.

Importantly, this is true even if it's not clear exactly what form leakage takes or if the prediction as to their queerness is inaccurate. We call predictors that act in this way "demographically incoherent." There are two important (if unintuitive) subtleties that immediately emerge from this description of incoherent predictions:

(1) *Harmful predictors need not be accurate:* Incoherent predictions focus on the behavior of the predictor *independent of accuracy.* Within the example above, it is not important if Asahi is actually queer, it is enough that the predictor guesses that Asahi is queer because of their presence in the data. This is because we envision our predictor being used to make some real-world decision, *e.g.,* limiting the opportunities available to Asahi due to their perceived queerness. As such, the prediction's accuracy is a secondary concern.

(2) *Measuring confidence is critical:* When considering the ways in which data releases can be translated into real-world harms, it's important to recognize that enabling an adversary to make high confidence predictions about private attributes is a problem. Importantly, this means that

[8] The notion of similarity is obviously a loaded one, as the ways in which two individuals are similar or different depend on the types of predictions being made about them. We eventually handle this by quantifying over many notions of similarity. For the sake of this motivation, it is enough to assume that the similarity of these individuals is meaningful with respect to the characteristic being predicted about them.

we should not require that the adversary can predict private attributes with 100% certainty in order for it to be considered harmful. Indeed, there is no particular cut-off threshold for certainty at which point it is natural to consider a harm occurring for all contexts. In turning to predictors as our adversarial strategy, we naturally arrive in a context within which notions of confidence have been extensively explored. Specifically, our approach considers confidence-rated predictors h, allowing us to directly reckon with predictive uncertainty.

We note that there are other pathological predictors which do not indicate privacy loss, and are therefore not considered demographically incoherent. For example, a predictor may make guesses that are entirely random or guess that everyone in the population has some feature, *i.e.,* make predictions that do not depend on the characteristics of individuals. The challenge then is to detect demographically incoherent predictors, whose behavior indicates privacy leakage, without depending on accuracy and without accidentally measuring variance in behavior that is not dataset dependent.

3.2 An Experiment to Detect Demographically Incoherent Predictors

With intuition about our class of "bad" predictors in hand, we now turn our attention to designing an experiment for detecting algorithms that produce them. In this discussion, we will defer on the concrete ways in which we will measure the demographic incoherence, and first focus on the experiment itself—that is, first we will decide what values we should measure, and then proceed to deciding how to do that measurement.

Because a symptom of incoherent predictors is differing performance on in-sample and out-of-sample individuals, it is clear that a *comparison* is required. However, it is not immediately obvious what the "right" comparison should actually be. In fact, some natural approaches fall short of our goals. As such, we walk through two seemingly natural, but flawed, experiments before discussing our final choice. Recall that the data curator has a dataset X and will be releasing a privacy-preserving report R.

(1) *Comparing before and after a data release:* A very natural approach would be to compare the performance of a predictor created *before* a data release with one created *after*, *i.e.,* comparing the performance (on individuals in X) of h_0 produced by an algorithm $\mathcal{L}$ with access to the adversary's pre-existing, auxiliary information Aux to a predictor h_1 produced by $\mathcal{L}$ with access to both Aux and the report R.[9] Such a comparison, intuitively, should isolate exactly the predictive changes associated with releasing R.

    Where this approach fails is that it does not recognize that there *should* be a difference between the predictors h_0 and h_1 over the inputs in X. After all, if there was no difference between h_0 and h_1, there would be no value whatsoever in releasing R! As such, this comparison is necessarily conflating potential "bad" types of predictions that releasing R enables with the "good" types of predictions that motivated the release of R in the first place.

(2) *Comparing to the base population:* The next most natural approach would be to compare the behavior of a single predictor h on individuals in X with that on individuals in the rest of the population. For example, by comparing its behavior on another similarly sampled dataset Y. This improves on our previous approach because we might expect that a "good" learning algorithm uses the dataset to learn about the population at large instead of revealing specifics about individuals.

[9] As this approach sketch is mainly to motivate our final approach below, we gloss over some formalities in this description. For example, how do we know that $\mathcal{L}$ does not act differently when provided one input (Aux) and two inputs (Aux and R)?

While this approach gets to the core of our interests, it has an important flaw. Technically speaking, we cannot assert that a real world sampling procedure has access to the base population distribution, *i.e.,* one cannot assume that two real world datasets are i.i.d samples from the same distribution. Also, we could conceivably be in a situation where the *entire* population of interest is contained in X, leaving no one in $\bar{X}$ against which we could compare. Therefore, keeping in mind the ergonomics of our definition in a concrete deployment scenario, this approach also falls short of our goals.

Our approach. We build off the second approach above by taking control of the randomness used to separate the two comparison populations. Specifically, we split the dataset X into two uniformly selected halves, X_a and X_b. We then use the data curation algorithm to generate the report R using only X_a, holding X_b in reserve as our "test" data set. We then test the behavior of a predictor h, designed based on the report R. Specifically, we compare the predictions of h on individuals in X_a and X_b. This approach "fixes" our second failed attempt by moving the assumptions about the randomness used in sampling X—something over which we have no control—into the randomness we use to split X into X_a and X_b—something over which we do have control. We say that a data release is *demographically incoherent* with respect to X if its predictions on members of X_a are noticeably different than the predictions it makes on members of X_b (who necessarily have similar demographic distributions, given the uniform split.)

3.3 Measuring the Incoherence of a Predictor

Finally, we discuss how to compare the behavior of h on X_a and X_b without relying on accuracy. Formally, we consider real-valued confidence-rated predictors $h : \mathcal{X} \to [-1, 1]$ which predict something about individuals. To capture the fact that these predictions are confidence rated, h outputs values in $[0, 1]$ when it predicts the attribute is likely true, and values in $[-1, 0]$ when it predicts the attribute is likely false, with a higher absolute value representing higher confidence.

For any such predictor, we will denote by $h(X_a)$ and $h(X_b)$ the uniform distributions over the predictions of h on X_a and X_b respectively. Comparing these distributions allows us to reason about the general behavior of the predictor h on X_a and X_b without considering accuracy of predictions on individuals. In order to get a more granular understanding of the behavior of h we further formalize the intuition of making comparisons over "similar" individuals in X_a and X_b below.

Measuring a difference with respect to "similar" individuals. In our motivating discussion of incoherent predictions, our representative individuals Asahi and Blair were assumed to be "similar" to one another. To formalise this intuition, we ask that a predictor is demographically coherent not only on the population as a whole, but also on recognizable subgroups from the population, *e.g.,* men, women, college freshmen, middle-school teachers etc... [10] For each of these subgroups of the population, the things that bind them together make them similar, in some particular sense. By operationalizing our earlier intuition in this way, we ask that the demographic coherence property holds not only over some particular notion of similarity, but rather over many notions of similarity at the same time. It also has the following technical and social benefits: (1) From a technical perspective, considering only the full population might hide incoherent decisions within sub-populations that effectively "cancel out." That is, there might be a right-shift in one group that masks a left-shift in a different group, each shift effectively "disappearing" in the collective distribution over all individuals. (2) From a social perspective, there may be particularly important groups within the population for whom we want to ensure coherent predictors for normative reasons. For example, if

[10] We borrow this conceptual approach from [32].

X is a Census-like dataset, we may want to ensure that there are not sub-geographies on which incoherent predictors are allowed. Similarly, we may want to ensure that there aren't legally protected categories (*e.g.,* race, sex, religion, etc. . .) on whom incoherent predictions are allowed.

The *lens* of a predictor. Consider the adversary using the Netflix dataset to learn a predictor for queerness. We assume that at the time of making predictions, the adversary sees a public user profile (their IMDB ratings) which contains only some of the user information that was contained in the dataset. To formalize this intuition we introduce the *lens* ρ of the predictor, which indicates the attributes contained in the dataset which the predictor "sees." We then compare the behavior of h on members of X_a and X_b as seen through the lens ρ.[11]

Choosing a metric. In this work, we recommend instantiating the demographic coherence experiment with the distance metric of *Wasserstein distance* (Definition 5), also known as earth-mover's distance, when measuring demographic coherence (or lack thereof). Intuitively, this metric measures the minimum amount of work that it requires to deform one probability distribution into another. If one visualizes a probability distribution as a mound of dirt, Wasserstein distance measures the effort required to move enough dirt to make one mound appear identical to the other (thus, earth-mover's distance). Unlike total variation distance, which disregards the spatial displacement needed to deform one distribution to another, Wasserstein distance is greater with a higher shift in confidence; recall that the importance of measuring confidence is one of the insights we highlight in Section 3.1. Another advantage of the Wasserstein distance is that it has been widely studied and used in theoretical and empirical statistics, and so there is a rich mathematical toolkit that one can borrow from when reasoning about it.

We recognize that there may be other measurement metrics that could be applied to the demographic coherence experiment that might highlight risk in different ways, and encourage this as important follow-up work.

4 Formal Definition

This section presents the formal definitions corresponding to our framework. For a discussion about the various choices made here, see Section 3.

Section 4 contains a glossary of the notation we use, Section 4.1 formally defines the notion of *incoherence* that we measure, Section 4.2 defines what it means for an algorithm to be *demographically incoherent*, and Section 4.3 defines what it means for a data curation algorithm to be *coherence enforcing.*

Notation. We fix the following notation:

- We define a data universe $\mathcal{X} = (\{0,1\}^* \cup \perp)^*$, where each feature of the data can also take the value $\perp$ (a.k.a. null).
- We denote a dataset consisting of n records from $\mathcal{X}$ by dataset $X \in \mathcal{X}^n$. [12]
- We consider a collection $\mathcal{C}$ of sub-populations $C \subseteq \mathcal{X}$.
- For any dataset X and sub-population $C \in \mathcal{C}$, we define $X|_C \stackrel{\text{def}}{=} X \cap C$ to be the restriction of X to the sub-population C, *i.e.,* the members of the dataset X that belong to sub-population C.
- We define a lens ρ as a set of features from $\mathcal{X}$.

[11]The predictor may also have side information about individuals not contained in the dataset, which can be formally included in the description of the adversarial algorithm $\mathcal{L}$.

[12]In a real-world scenario, X might be sampled from some underlying population, but our definition does not require this and makes no assumptions about how it might be done.

- For a lens ρ, we define $\boldsymbol{\pi}_\rho(X)$ to be the data in X restricted to the features in the lens. That is, for every feature represented by ρ, the sets $\boldsymbol{\pi}_\rho(X)$ and X are exactly the same, and for features not represented by ρ, the entries in $\boldsymbol{\pi}_\rho(X)$ always have the value $\perp$.
- For any fixed predictor $h : \mathcal{X} \to [-1, 1]$, define the distribution $h(X)$ as the uniform distribution over the predictions of h on X. That is, the distribution $h(X)$ has cumulative distribution function

$$cdf_{h(X)}(p) = \Pr_{x \xleftarrow{\$} X}[h(x) \le p].$$

4.1 Measuring Demographic Incoherence of Predictors

In this section we define the notion of *incoherence* that we measure over predictors. This measurement is used informally in the demographic coherence experiment DemCoh (Figure 1).

Definition 2 (Demographically Incoherent Predictor)**.** *Let $\mathcal{X}$ be a data universe, let $X_a, X_b \in \mathcal{X}^{n/2}$ be datasets consisting of $n/2$ data entries each, let $\mathcal{C}$ be a collection of sub-populations $C \subseteq \mathcal{X}$, let the lens ρ be a set of features from $\mathcal{X}$, and let $h : \mathcal{X} \to [-1, 1]$ be a confidence rated predictor. Let $d(\cdot, \cdot)$ represent some metric distance between probability distributions.*

We say that the h has α-incoherent predictions *with respect to X_a, X_b, ρ, and $\mathcal{C}$, if for some $C \in \mathcal{C}$,*

$$d\Big(h(\boldsymbol{\pi}_\rho(X_a|_C)),\, h(\boldsymbol{\pi}_\rho(X_b|_C))\Big) > \alpha$$

(In contrast, h has α-coherent predictions *with respect to X_a, X_b, ρ, and $\mathcal{C}$, if for all $C \in \mathcal{C}$, it satisfies $d\Big(h(\boldsymbol{\pi}_\rho(X_a|_C)),\, h(\boldsymbol{\pi}_\rho(X_b|_C))\Big) \le \alpha$.)*

In other words, the predictions of h are incoherent if there is some sub-population C that witnesses a big distance between its predictions on two sets X_a, and X_b. Note that this definition distills a notion of "incoherence" only once we fix some assumptions on X_a, X_b, perhaps that they are drawn from similar distributions.

4.2 Demographic Coherence

Consider a data universe $\mathcal{X}$, an algorithm $\mathcal{L} : \mathcal{X}^{n/2} \to (\mathcal{X} \to [-1, 1])$ which uses a dataset of size $n/2$ to produce a confidence-rated predictor $h : \mathcal{X} \to [-1, 1]$. Let $\mathcal{C}$ be a collection of sub-populations $C \subseteq \mathcal{X}$, let the lens ρ be a set of features from $\mathcal{X}$, Let $\mathsf{dist}(\cdot, \cdot)$ represent some distance metric between probability distributions.

In this section, we formally define when the algorithm $\mathcal{L}$ is *demographically coherent*. We start by defining the demographic coherence experiment DemCoh which checks the demographic coherence of $\mathcal{L}$ with respect to on a specific dataset X, a collection $\mathcal{C}$, a lens ρ, and a distance metric $\mathsf{dist}(\cdot, \cdot)$. This experiment works similarly to the description in Section 3.2, except we are testing the coherence of an algorithm that gets the dataset in the clear. In Section 4.3, we will use the notion of demographic coherence to define when a data curator is *coherence enforcing*.

In the DemCoh experiment, the input dataset X is split into sets X_a, X_b where X_b is held in reserve as a "test" set. Then the algorithm $\mathcal{L}$, with input X_a, is used to produce a predictor h. Finally, the predictor h is checked for *demographic incoherence* (Def 2) with respect to X_a, X_b, ρ, and $\mathcal{C}$.

$\mathsf{DemCoh}_{\mathcal{L},X,\mathcal{C},\rho,\mathsf{dist}(\cdot,\cdot)}(\alpha)$

Input:
An algorithm $\mathcal{L} : \mathcal{X}^{n/2} \to (\mathcal{X} \to [-1,1])$, A dataset $X \in \mathcal{X}^n$, a collection $\mathcal{C}$ of subgroups $C \subseteq \mathcal{X}$, a lens ρ, and a distance metric $\mathsf{dist}(\cdot,\cdot)$.

Split Data:
$I \xleftarrow{\$} \{S \subseteq [n] : |S| = n/2\}$
$X_a = (x_i)_{i \in I}$
$X_b = (x_i)_{i \in [n] \setminus I}$

Compute Predictor(X_a)**:**
$h \leftarrow \mathcal{L}(X_a)$

Incoherence Condition:
Set $b = 0$ if there exists $C \in \mathcal{C}$ such that:
$$\mathsf{dist}\Big(h(\boldsymbol{\pi}_\rho(X_a|_C)),\, h(\boldsymbol{\pi}_\rho(X_b|_C))\Big) > \alpha$$
Else set $b = 1$

Figure 1: Demographic Coherence Experiment

A natural formalization of the intuition we developed in Section 3 would say that an algorithm $\mathcal{L}$ produces (α,β)-demographically coherent predictions with respect to collections $\mathcal{C}$ and lens ρ if the following holds for all datasets X:

$$\Pr[\mathsf{DemCoh}_{\mathcal{L},X,\mathcal{C},\rho}(\alpha) = 0] \leq \beta.$$

However, this definition cannot be realized with respect to arbitrary categories C and all datasets, because the sampling experiment itself introduces some incoherence. For example, consider a category C that is men over 60, and a dataset that contains only two people in this category. There exists a predictor that predicts -1 on one of them and 1 on the other. With probability 1/2, these two men end up in X_a and X_b respectively, and the Wasserstein distance between the predictors' distributions on $X_a|_C$ and $X_b|_C$ is 2. Hence, for our definition to be meaningfully achievable, we require the datasets considered to be sufficiently representative so that the incoherence due to sampling does not dominate. Thus, we define demographic coherence with respect to a size constraint.

To remove the need for the parameter γ, one could redefine $X_a|_C$ by "zeroing out" members of X_a that are not in C instead of taking the intersection $X_a \cap C$. This would effectively result in asking the predictor h in Definition 2, and Figure 1 to make "dummy" predictions, with no information, for every member of X_a and X_b not belonging to C. While such a definition may be more mathematically elegant, we believe that the explicit failure point represented by γ in our defintion is important for interpretability and ease of use—especially by non-experts. For the same reason, it is important to have an explicit collection $\mathcal{C}$, and lens ρ, even though the eventual theorems may be quite general, holding for large ranges of γ, $\mathcal{C}$, and ρ.

Definition 3. (Demographic Coherence). *Consider a data universe $\mathcal{X}$, and an algorithm $\mathcal{L} : \mathcal{X}^{n/2} \to (\mathcal{X} \to [-1,1])$ which uses a dataset of size $n/2$ to produce a fixed confidence-rated predictor $h : \mathcal{X} \to [-1,1]$. Let $\mathcal{C}$ be a collection of sub-populations $C \subseteq \mathcal{X}$, let the lens ρ be a set of features from $\mathcal{X}$, Let $\mathsf{dist}(\cdot,\cdot)$ represent some metric distance between probability distributions. We say that* $\mathcal{L}$ produces (α,β)-demographically coherent predictions *with respect to collection $\mathcal{C}$, size-constraint γ, and lens ρ if the following holds:*

For all $X \in \mathcal{X}^n$, $\mathcal{C}^* = \{C \in \mathcal{C} \mid |C \cap X| \geq \gamma\}$

$$\Pr[\mathsf{DemCoh}_{\mathcal{L},X,\mathcal{C}^*,\rho}(\alpha) = 0] \leq \beta.$$

4.3 Coherence Enforcing Algorithms

We finally define *coherence-enforcing* algorithms by reference to the definition of *demographic coherence* (Definition 3). Specifically, a data curator $\mathcal{A}$ is *coherence enforcing* if, any algorithm $\mathcal{L}$ can be rendered *demographically coherent* simply by filtering its inputs through the data curator $\mathcal{A}$, without any changes to $\mathcal{L}$.

Definition 4. (Coherence Enforcing Algorithms)**.** *Consider data universes* $\mathcal{X}, \mathcal{Y}$*, a collection* $\mathcal{C}$ *of sub-populations* $C \subseteq \mathcal{X}$*, a lens* ρ*, and an algorithm* $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$*. Let* $\mathsf{dist}(\cdot, \cdot)$ *represent some metric distance between probability distributions.*

We say $\mathcal{A}$ *enforces* (α, β)*-demographic coherence with respect to collection* $\mathcal{C}$*, size-constraint* γ*, and lens* ρ *if:*

For any algorithm $\mathcal{L} : \mathcal{Y} \to (\mathcal{X} \to [-1, 1])$*, the combined algorithm* $\mathcal{L} \circ \mathcal{A} : \mathcal{X}^{n/2} \to (\mathcal{X} \to [-1, 1])$ *satisfies* (α, β)*-demographic coherence with respect to the collection* $\mathcal{C}$*, size-constraint* γ*, and lens* ρ*.*

4.4 Instantiating Demographic Coherence with a Metric

We will instantiate these definitions in the following sections using the Wasserstein-1 metric defined below:

Definition 5. *Let* P, Q *represent distributions over a discrete subset* $S \subseteq \mathbb{R}$*. Then, the* 1*-Wasserstein distance between* P, Q *is defined as*

$$\mathsf{dist}_{\mathcal{W}_1}(P, Q) = \inf_{\pi} \sum_{i \in S} \sum_{j \in S} \|x_i - x_j\|_1 \pi(x_i, x_j),$$

where the infimum is over all joint distributions π *on the product space* $S \times S$ *with marginals* P *and* Q *respectively.*

If an algorithm is (α, β)-demographically-coherent as per Definition 3 with this Wasserstein metric instantiation, we say that it is (α, β)*-Wasserstein-demographically-coherent* (or (α, β)-Wasserstein-coherent for short.) Similarly, if an algorithm enforces (α, β)-demographic-coherence as per Definition 4 with this Wasserstein metric, then we say it *enforces* (α, β)*-Wasserstein-demographic-coherence* (or it enforces (α, β)-Wasserstein-coherence for short.)

5 Capturing Real-Life Vulnerabilities from Releasing Data in the Clear

In this section, we construct a simple and realistic experiment to demonstrate how our framework can be used to capture real-life privacy leaks.

Set up: A governmental entity is interested in facilitating academic study on disability with the goal of providing better support, and so releases a dataset consisting of data about people—excluding their names, but including age, sex, racial features, type of work, as well as cognitive disability status. The governmental entity is debating whether to release this data in the clear reckoning with the following considerations:

- The privacy harm of releasing such data in the clear is not immediately apparent—disability status is a sensitive feature, but the remaining attributes may not uniquely identify most individuals and hence may not provide a tool to systematically deduce the disability status of individuals.

- Releasing such data in the clear would maximize the utility associated with the release.

The adversary in our setup is a corporation that is screening job applicants, and due to ableism, wants to avoid hiring employees with cognitive disabilities. However, regulations prevent companies from requiring job applicants to provide this information. Hence, the adversary decides to use the data release to train a classifier predicting cognitive disability status, with the intent of using it to predict disability status of individuals who apply for jobs. We note that machine learning methods are already used to screen resumes , and so this flow captures a realistic decision-making process.

Our experiment aims to demonstrate the utility of our framework in this scenario by exploring whether we can use our framework with the above set up to expose the privacy harms associated with releasing such data.

Experimental details: Next, we describe the details of the experiment we run using the above set up.

Dataset: We simulate the governmental entity in the setup using the NIST American Community Survey (ACS) Data Excerpts benchmark [56], which consists of curated subsets of publicly released datasets containing info about real households in the US. This benchmark is intended to investigate the performance of data synthesizers and other privacy enhancing technologies. We specifically use the national version of the dataset from 2019, which consists of 27254 records with 24 features each drawn from 20 different geographical areas in the United States. We use a subset of 9 features for each record—age, population density of the geographical region, sex, race, hispanic origin, highest educational attainment, general and specific categories of work, and cognitive disability status (a binary attribute). We note that 1521 people in the dataset (5.5%) have a cognitive disability, which is a small but significant fraction.

Algorithm used by the adversary: Our adversary uses a supervised learning approach: training a random forest classifier with 50 sub-classifiers as the algorithm to generate a predictor for cognitive disabilities. We use the scikit-learn [48] implementation of this algorithm.

Experimenting with our framework: Our framework involves randomly splitting the dataset into two halves, training the random forest classifier on one half of the dataset and using the other half as a held-out set for assessing the privacy loss. Our framework requires specifying vulnerable groups- and we consider 2 overlapping groups: the entire dataset, and the set of individuals who have a cognitive disability (and hence are at particular risk of not being hired if their disability status is disclosed or inferred). After training the classifier, we compute the prediction distributions on the two halves (restricted to the subgroups of interest), and then consider the Wasserstein distance between these prediction distributions. We repeat this experiment a number of times and consider the average Wasserstein distance. Note once again that cognitive disability is a binary attribute in the dataset; we use -1 to represent the presence of a cognitive disability and 1 to represent the absence of a cognitive disability.

Results: Figure 2 shows the Wasserstein distance between the training and test set restricted to the cognitive disability subgroup. The corresponding average Wasserstein distance is 1.561. We note that for the same set of iterations, the average Wasserstein distance when considering the whole training and test sets was a significantly smaller value of 0.0773.

To interpret the source of the Wasserstein distance difference and explain the associated intuition, we also plot the entire prediction distributions over the training and test sets restricted to the

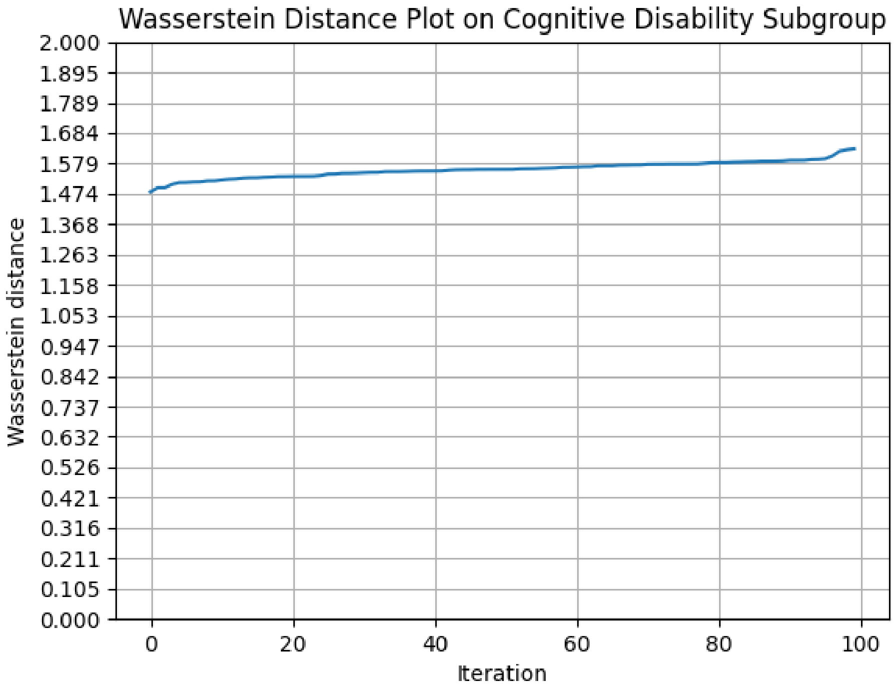

Figure 2: This plot shows the Wasserstein distances in different iterations between the prediction distributions on the training and test set restricted to the cognitive disability subgroup. (Note that the distances have been sorted).

Cognitive Disability subgroup for a single run of the algorithm. Note that the ground truth values in both cases are all -1s.

Discussion: Next, we discuss and interpret the results of the experiment.

Our framework captures privacy harms of releasing data in the clear in this setup: From the results of the experiment, we see that for individuals with a cognitive disability, there is a very large Wasserstein distance between the prediction distributions on the training set and the test set (1.561, for a value ranging between 0 and 2). As seen in Fig 3, this is because $+1$ is predicted on a much more significant fraction of this subgroup in the test set than the training set. In particular, in our setup this would mean that people with cognitive disabilities whose data was disclosed in the release have a much higher chance of being predicted as having a cognitive disability by an adversary than people with cognitive disabilities whose data was not disclosed in the release. As a result, people with cognitive disabilities whose data is in the release face a much higher risk of discrimination (e.g., in hiring). The large Wasserstein distance can be used to flag this privacy harm, supporting its use as a metric in our framework.

Our explanation of this is as follows: since a small percentage (5.5%) of the entire dataset has a cognitive disability, and there may not be enough information to systematically identify this subset. As a result. the random forest classifier simply memorizes and regurgitates many members of the training dataset that have the cognitive disability in order to minimize training loss. For most other members, the classifer simply resorts to predicting $+1$, including for most members of the test set restricted to the cognitive disability subgroup.

Focusing on subgroups is important: Note that the average Wasserstein distance between

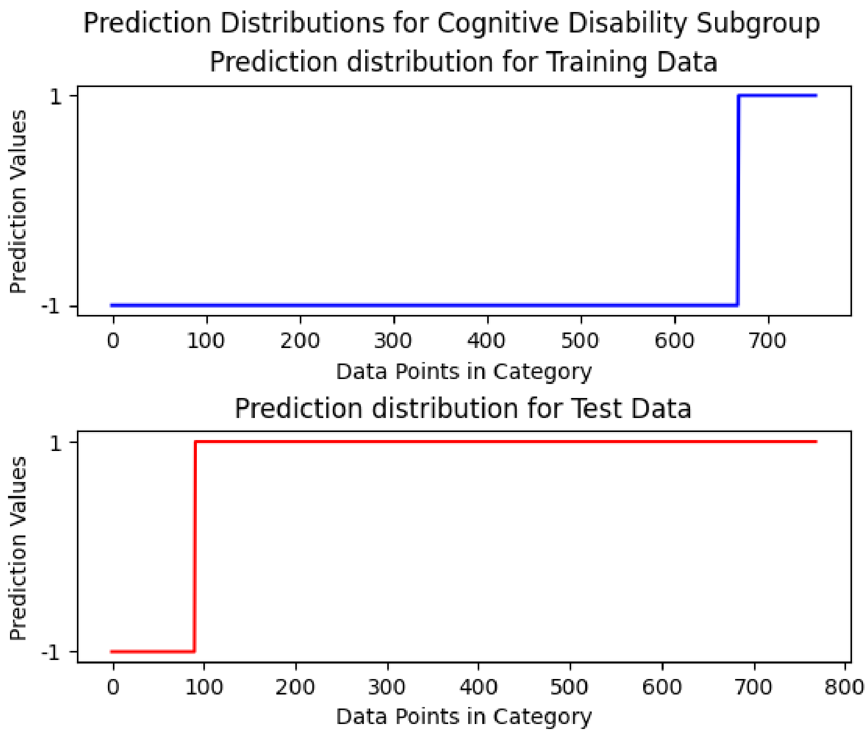

Figure 3: These plots show the predictions (sorted) on the training and test set restricted to the subgroup of people with a cognitive disability for a single iteration of the experiment.

the prediction distributions on the entire training and test set is relatively small - it is roughly 0.073. On the other hand, when we zoom into the subgroup containing individuals with a cognitive disability, the Wasserstein distance blows up to 1.561, exposing a more significant privacy harm. This is why we consider subgroups in our framework- because aggregating over the entire dataset can result in privacy harms on subgroups being diminished to a point where they are not noticeable.

Privacy harms can accrue even without an explicit attack: A common critique of the literature on privacy attacks is the question of who has the incentives to launch particular classes of attacks. Our results show that privacy harms can accrue even without such attacks being explicitly carried out. In our experiment, the adversary does not seek to infer membership in the training dataset or to reconstruct the data of individuals, but is attempting to learn a good classifier of a sensitive attribute. It turns out that the classifer memorizes information in the training dataset that can be leveraged to cause discrimination that affects members of the training set disproportionately, thereby precipitating a privacy harm. Hence, our framework systematically captures a set of privacy violations not directly addressed by the literature on privacy attacks.

6 Differentially Private Algorithms Enforce Wasserstein Demographic Coherence

In this section, we show that differentially private algorithms enforce Wasserstein demographic coherence. In fact, we prove the stronger statement Theorem 3 (in Section 6.1) which says that algorithms with *bounded max-information* are coherence enforcing (for all possible lenses ρ). Since differentially private algorithms have bounded max-information (Theorems 14 and 15) , this also proves that all differentially private algorithms are coherence enforcing regardless of the choice of lens ρ (Theorems 7 and 8).

Let $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$ denote a data curator operating on datasets of size $n/2$. For parameters $\beta, \zeta > 0$, algorithm $\mathcal{A}$ has β-approximate max-information at most γ if for all sets of outcomes $T \subseteq (\mathcal{X} \times \mathcal{Y})$,

$$\Pr[(S, \mathcal{A}(S)) \in T] \leq e^{\gamma} \Pr[(S, \mathcal{A}(S')) \in T] + \beta,$$

where S, S' denote independent samples each of size $n/2$ from $\mathcal{X}$. (See Definition 7 for the full definition.) Intuitively, this condition guarantees that output of the algorithm $\mathcal{A}$ does not reveal much more information about the input dataset S than it would about an independently sampled dataset S' drawn from the same distribution. The main technical result in this section shows that every algorithm with bounded max-information enforces demographic coherence with respect to the collection $\mathcal{C}$ of all sufficiently large groups.

Recalling the experiment $\mathsf{DemCoh}_{\mathcal{L}\circ\mathcal{A},X,\mathcal{C},\rho}(\alpha)$ described in Figure 1, our goal is to show that for all algorithms $\mathcal{L} : \mathcal{Y} \to \{\mathcal{X} \to [-1, 1]\}$, all lenses ρ, all datasets $X \in \mathcal{X}^n$, and all sufficiently large $C \subseteq \mathcal{X}$, that with high probability over a choice of split $X_a, X_b \xleftarrow{\$} X$, and predictor $h \leftarrow \mathcal{L} \circ \mathcal{A}(X_a)$, the following holds:

$$\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(X_a|_C)), h(\boldsymbol{\pi}_\rho\,(X_b|_C))) < \alpha.$$

Here, α is a parameter that we relate to γ, β and the minimum subpopulation size in $\mathcal{C}$. A conceptual challenge in analyzing this Wasserstein distance is that the predictor h depends on the split X_a, X_b. If the draws being fed into h could instead be modeled as independent samples, we could hope to use concentration inequalities to control this distance. Fortunately, this is where we can leverage bounded max-information of $\mathcal{A}$. In Claim 1 we show that it is sufficient to reason about the case where the predictor h is independent of the samples with which the prediction distributions are defined, that is,

$$\begin{aligned}\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(X_a|_C)), h(\boldsymbol{\pi}_\rho\,(X_b|_C)))&\\ &\approx \mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(S|_C)), h(\boldsymbol{\pi}_\rho\,(\overline{S}|_C))),\end{aligned}$$

where S denotes a fresh (independent of X_a, X_b) random subset of X.

It now suffices to show that $\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(S|_C)), h(\boldsymbol{\pi}_\rho\,(\overline{S}|_C))) < \alpha$ with high probability. To do this, we use an equivalent characterization of the Wasserstein distance between two probability distributions μ, ν over $[-1, 1]$ as

$$\mathsf{dist}_{\mathcal{W}_1}(\mu, \nu) = \int_{-1}^{1} |\mathsf{cdf}_\mu(\ell) - \mathsf{cdf}_\nu(\ell)|\; d\ell,$$

where $\mathsf{cdf}_\mu, \mathsf{cdf}_\nu$ denote the cumulative distribution functions of μ and ν. The upshot of this formulation is that for each fixed value of the parameter ℓ, the value of $\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(S|_C))}(\ell)$ concentrates tightly around its expectation, and is hence close to that of $\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(\overline{S}|_C))}(\ell)$ with high probability. We formalize this argument in Claim 2, where we use the fact that these quantities can be explicitly modeled as hypergeometric random variables. We then argue that pointwise closeness of these CDFs implies closeness of the integral of their difference, and hence of the desired Wasserstein distance.

6.1 Bounded Max-Information Algorithms are Coherence Enforcing

Definition 6 (Max-information of random variables [24])**.** *Let X and Y be jointly distributed random variables over the domain $(\mathcal{X}, \mathcal{Y})$. The β-approximate max-information between X and Y,* *denoted by* $I_\infty^\beta(X;Y)$ *is*

$$I_\infty^\beta(X;Y) = \ln\left(\sup_{\substack{T\subseteq(\mathcal{X}\times\mathcal{Y})\\ \Pr[(X,Y)\in T]>\beta}} \frac{\Pr[(X,Y)\in T]-\beta}{\Pr[X\otimes Y\in T]}\right)$$

Definition 7 (Max-information of algorithms [24])**.** [13] *Fix* $n\in\mathbb{N}$, $\beta>0$. *Let* $\mathcal{X}$ *be a finite data universe of size* m. *Let* S *be a sample of size* n *chosen without replacement from* $\mathcal{X}$. *Let* $\mathcal{A}:\mathcal{X}^n\to\mathcal{Y}$ *be an algorithm.*

Then we define the max-information of the algorithm as follows:

$$I_\infty^\beta(\mathcal{A},n) = I_\infty^\beta(S,\mathcal{A}(S)) \tag{1}$$

The main technical result we prove is the following.

Theorem 3. *Let* $n\in\mathbb{N}$,$\zeta>0$, $\beta\in(0,2)$, $\alpha\in(0,2]$. *Let* $\mathsf{dist}_{\mathcal{W}_1}(\cdot,\cdot)$ *denote the Wasserstein-1 distance metric. Consider a collection* $\mathcal{C}$ *of sub-populations* $C\subseteq\mathcal{X}$, *a lens* ρ, *and an algorithm* $\mathcal{A}:\mathcal{X}^{n/2}\to\mathcal{Y}$ *with bounded max-information*

$$I_\infty^{\beta/2|\mathcal{C}|}(\mathcal{A},n/2)<\zeta.$$

Then $\mathcal{A}$ *enforces* (α,β)*-Wasserstein-coherence with respect to collection* $\mathcal{C}$, *lens* ρ, *and size constraint* γ, *where*

$$\begin{aligned}\gamma = \max\Big\{&\frac{33.2\cdot(\zeta+\ln(16|\mathcal{C}|/\beta))}{\alpha^2}, \frac{144\ln(6/\alpha)}{\alpha^2},\\ &16.6\cdot(\zeta+\ln(16|\mathcal{C}|/\beta)), \frac{10.6}{\alpha}, 80\Big\}.\end{aligned} \tag{2}$$

The intuition behind the result in Theorem 3 is that the output of an algorithm with bounded max-information does not contain too much specific information about the input dataset. This intuition is leveraged to prove Lemma 1, which is the main lemma underlying this result. Theorem 3 follows from this lemma by an appropriate setting of parameters.

Lemma 1. *Let* $\eta,\zeta>0$, $\alpha\in(0,1)$. *Let* $\mathsf{dist}_{\mathcal{W}_1}(\cdot,\cdot)$ *represent the Wasserstein-1 distance metric. Consider a sub-population* $C\subseteq\mathcal{X}$, *a lens* ρ, *and an algorithm* $\mathcal{A}:\mathcal{X}^{n/2}\to\mathcal{Y}$ *with bounded max-information*

$$I_\infty^\eta(\mathcal{A},n/2)<\zeta.$$

For all algorithms $\mathcal{L}:\mathcal{Y}\to\{\mathcal{X}\to[-1,1]\}$, *datasets* $X\in\mathcal{X}^n$, $\mu>0$, *as long as*

$$|X\cap C|\geq\max\left\{\frac{4.15\cdot\ln(4/\mu)}{\alpha^2}, \frac{16/3}{\alpha}, 8.3\cdot\ln(4/\mu), 40\right\},$$

we have

$$\Pr_{X_a\leftarrow X, h\leftarrow\mathcal{L}\circ\mathcal{A}(X_a)}\left[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho(X_a|_C)), h(\boldsymbol{\pi}_\rho(X_b|_C)))>2\alpha\right] \quad\leq\ 2\mu(|X\cap C|+1)\cdot e^\zeta+\eta.$$

Here X_a *and* X_b *denote a random split of the dataset* X *as in the* $\mathsf{DemCoh}_{\mathcal{L}\circ\mathcal{A},X,\mathcal{C}^*,\rho}(\alpha)$ *experiment in Figure 1.*

[13] This definition, for sampling without replacement, is slightly different than the original one.

Proof of Lemma 1. Fix any arbitrary lens ρ. The proof proceeds in two claims. First, in Claim 1, we use the definition of max-information to replace X_a, X_b with an independently chosen half-sample S and its complement $\overline{S} = X \setminus S$.

Claim 1. *Consider $\eta, \alpha, \zeta > 0$ and a fixed dataset $X \in \mathcal{X}^n$. Consider a data-curation algorithm $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$ with bounded max-information, $I^{\eta}_{\infty,P}(\mathcal{A}, n/2) \leq \zeta$, and an algorithm $\mathcal{L} : \mathcal{Y} \to \{\mathcal{X} \to [-1,1]\}$ that uses the data report to create a predictor. Independently choose two random half samples $X_a, S \leftarrow X$, and let sets $X_b = X \setminus X_a, \overline{S} = X \setminus S$. Finally let $h \leftarrow \mathcal{L}(X_a)$. Then, we have that*

$$\begin{aligned}&\Pr_{X_a,h}[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho(X_a|_C)), h(\boldsymbol{\pi}_\rho(X_b|_C))) > \alpha] \\ &\leq \Pr_{S,X_a,h}[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho(S|_C)), h(\boldsymbol{\pi}_\rho(\overline{S}|_C))) > \alpha] \cdot e^{\zeta} + \eta\end{aligned}$$

Then, in Claim 2, we bound $\Pr[\mathsf{dist}_{\mathcal{W}_1}\left(g(\boldsymbol{\pi}_\rho(S|_C)), g\left(\boldsymbol{\pi}_\rho\left(\overline{S}|_C\right)\right)\right) > \alpha]$ for any confidence rated predictor $g : \mathcal{X} \to [-1,1]$ that is produced independently of S.

Claim 2. *Let $\alpha \in (0,1)$, let S be a sample of size $n/2$ drawn uniformly without replacement from X, let $\overline{S} = X \setminus S$, and let $g : \mathcal{X} \to [-1,1]$ be any confidence rated predictor. For any $\mu > 0$, when $|X \cap C| \geq \max\{\frac{4.15}{\alpha^2}\ln(4/\mu), \frac{5.3}{\alpha}, 8.3\ln(4/\mu), 40\}$, we have that*

$$\Pr\left[\mathsf{dist}_{\mathcal{W}_1}\left(g(\boldsymbol{\pi}_\rho(S|_C)), g(\boldsymbol{\pi}_\rho\left(\overline{S}|_C\right))\right) > \alpha\right] \leq 2(1 + |X \cap C|)\mu.$$

Putting these claims together, we get that

$$\Pr_{X_a,h}[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho(X_a|_C)), h(\boldsymbol{\pi}_\rho(X_b|_C))) > \alpha] \;\leq\; 2\mu(|X \cap C| + 1) \cdot e^{\zeta} + \eta.$$

Now we proceed to prove Claims 1 and 2.

Proof of Claim 1. First, note that since the algorithm $\mathcal{L}$ postprocesses the report output by the data curator, by the fact that max-information is preserved under postprocessing, it inherits its max-information. Let $\mathcal{A}^*$ be the combined algorithm $\mathcal{L} \circ \mathcal{A}$. Then by the definition of max-information, and since $(X_a, \mathcal{A}^*(S))$ is distributed exactly the same as $(S, \mathcal{A}^*(X_a))$,

$$I^{\eta}_{\infty}(X_a; \mathcal{A}^*(X_a)) = \mathsf{dist}^{\eta}_{\infty}\Big(\big(X_a, \mathcal{A}^*(X_a)\big)||\big(X_a, \mathcal{A}^*(S)\big)\Big) = \mathsf{dist}^{\eta}_{\infty}\Big(\big(X_a, \mathcal{A}^*(X_a)\big)||\big(S, \mathcal{A}^*(X_a)\big)\Big).$$

we have that for all T such that $\Pr[(X_a, \mathcal{A}^*(X_a) \in T] > \eta$,

$$\log\left(\frac{\Pr[(X_a, \mathcal{A}^*(X_a)) \in T] - \eta}{\Pr[(S, \mathcal{A}^*(X_a))) \in T]}\right) \leq \zeta.$$

Given C, we can post-process a pair $(X_a, \mathcal{A}^*(X_a))$ to compute $(h(\boldsymbol{\pi}_\rho(X_a|_C)))$ and $(h(\boldsymbol{\pi}_\rho(X_b|_C)))$. This is because $h \leftarrow \mathcal{A}^*(X_a)$ and $X_b = X \setminus X_a$. Applying the same post-processing to $(S, \mathcal{A}^*(X_a))$ yields $(h(\boldsymbol{\pi}_\rho(S|_C)))$ and $(h(\boldsymbol{\pi}_\rho\left(\overline{S}|_C\right)))$.

Let $T = \{(S,h) \mid \mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho(S|_C)), h(\boldsymbol{\pi}_\rho\left(\overline{S}|_C\right))) > \alpha\}$.Then if $\Pr[(X_a, \mathcal{A}^*(X_a)) \in T] > \eta$,

$$\log\left(\frac{\Pr[(X_a, \mathcal{A}^*(X_a)) \in T] - \eta}{\Pr[(S, \mathcal{A}^*(X_a))) \in T]}\right) \leq \zeta.$$

This means that if $\Pr[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(X_a|_C)), h(\boldsymbol{\pi}_\rho\,(X_b|_C))) > \alpha] > \eta$,

$$\log\left(\frac{\Pr[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(X_a|_C)), h(\boldsymbol{\pi}_\rho\,(X_b|_C))) > \alpha] - \eta}{\Pr[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(S|_C)), h(\boldsymbol{\pi}_\rho\,(\overline{S}|_C))) > \alpha]}\right) \leq \zeta.$$

We can rearrange the above equation to get that

$$\begin{aligned}&\Pr[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(X_a|_C)), h(\boldsymbol{\pi}_\rho\,(X_b|_C))) > \alpha] \\ \leq\, &\Pr[\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho\,(S|_C)), h(\boldsymbol{\pi}_\rho\,(\overline{S}|_C))) > \alpha] \cdot e^{\zeta} + \eta,\end{aligned}$$

as required. ■

Proof of Claim 2. Let $\alpha, \mu \in (0,1)$ (the statement holds trivially for $\mu > 1$), and suppose $|X \cap C| \geq \max\{\frac{4.15}{\alpha^2}\ln(4/\mu), \frac{5.3}{\alpha}, 8.3\ln(4/\mu), 40\}$. By the definition of the distance metric we have the following:

$$\mathsf{dist}_{\mathcal{W}_1}\left(g(\boldsymbol{\pi}_\rho\,(S|_C)), g(\boldsymbol{\pi}_\rho\,(\overline{S}|_C))\right) = \int_{-1}^{1} |\mathsf{cdf}_{g(\boldsymbol{\pi}_\rho(S|_C))}(g) - \mathsf{cdf}_{g(\boldsymbol{\pi}_\rho(\overline{S}|_C))}(g)|\, dg. \tag{3}$$

To bound this value, we first prove the following for a fixed $y \in [-1,1]$.

$$\Pr\left[|\mathsf{cdf}_{g(\boldsymbol{\pi}_\rho(S|_C))}(y) - \mathsf{cdf}_{g(\boldsymbol{\pi}_\rho(\overline{S}|_C))}(y)| \geq \alpha\right] \leq \mu. \tag{4}$$

Then, we observe that there are at most $|X \cap C| + 1$ effectively different values of y we need to consider with respect to any fixed g and C. (For every realization of $S, \overline{S}$, $\mathsf{cdf}_{g(\pi_\rho(S|_C))}$ can only change for values of y on which $\mathsf{cdf}_{g(\pi_\rho(X|_C))}$ changes. These values correspond to the partitioning of $[-1,1]$ into intervals induced by applying $g \circ \pi_\rho(\cdot)$ to the elements in $X \cap C$.) By union bounding over these $|X \cap C| + 1$ effectively different values of y, Equation (4) gives us the following.

$$\Pr\left[\sup_{y\in[-1,1]} |\mathsf{cdf}_{g(\boldsymbol{\pi}_\rho(S|_C))}(y) - \mathsf{cdf}_{g(\boldsymbol{\pi}_\rho(\overline{S}|_C))}(y)| \geq \alpha\right] \leq (1 + |X \cap C|)\mu. \tag{5}$$

Finally, substituting the bound from Equation (5) in Equation (3) proves the lemma.

To show Equation (4), we reinterpret the sampling process in terms of hypergeometric distributions, defined as follows.

Definition 8 (Hypergeometric distribution)**.** *Fix $0 < a, s \leq b$. Consider a population of b items of which a items have a special property P. Consider s items sampled without replacement from b. The distribution of the number of items in s with property P is called the hypergeometric distribution parameterized by b, a, s (denoted by $H(b,a,s)$).*

This will in turn allow us to invoke Theorem Theorem 4, a concentration result for the hypergeometric distribution.

Theorem 4 ([34])**.** *Let K have a hypergeometric distribution $H(b,a,s)$. Then for every $\gamma \geq 2$,*

$$\begin{aligned}\Pr[K > s\frac{a}{b} + \gamma] < e^{-2c(\gamma^2-1)}\\ \Pr[K < s\frac{a}{b} - \gamma] < e^{-2c(\gamma^2-1)}\end{aligned}$$

where

$$c = \max\left\{\frac{1}{s+1} + \frac{1}{b-s+1}, \frac{1}{a+1} + \frac{1}{b-a+1}\right\}.$$

Consider an urn consisting of n balls. Among those n balls, m are marked with a red stripe, representing membership in $C \cap X$. Among the m red-striped balls, t are further marked with a blue stripe, representing $x \in C \cap X$ such that $g(\boldsymbol{\pi}_\rho(x)) \le y$ for the value of y being considered. Consider the experiment where we sample $n/2$ balls without replacement, and define the joint pair of random variables (V, W) where V counts the number of red-striped balls in the sample, (i.e., the number of sampled points that are in $X \cap C$) and W counts the number of (red and) blue-striped balls in the sample, (i.e., the number of sampled points x that are in $X \cap C$ and satisfy $g(\boldsymbol{\pi}_\rho(x)) \le y$. The random variables W and V follow hypergeometric distributions as follows:

$$\begin{aligned} V &\sim H(n, m, n/2) \\ (W|V = v) &\sim H(m, t, v). \end{aligned}$$

Observe that the absolute value of the CDF difference we are trying to bound is equal to $\left|\frac{W}{V} - \frac{t-W}{m-V}\right|$ by definition.

Let E_1 be the event that the number of red-striped balls in the sample is close to its expected value (i.e., $|V - m/2| < m/4$). Then applying Theorem 4 and using $m > 40$ and $m > 8.3\ln(4/\mu)$ we have that

$$\begin{aligned} \Pr[\overline{E_1}] &< 2\exp\left(\frac{-2}{m+1} \cdot \left((m/4)^2 - 1\right)\right) \\ &\le 2\exp\left(\frac{-2}{1.025m} \cdot \left(0.99(m/4)^2\right)\right) \\ &= 2\exp\left(\frac{1.98}{16.4}m\right) < \mu/2. \end{aligned}$$

Now let us condition on a realization $V = v$. Given this, let E_2 be the event that the number of blue-striped balls in the sample is close to its expected value (i.e., $\left|W - \frac{tv}{m}\right| \le \zeta$). Then applying Theorem 4 for $\zeta \ge 2$ and c as in the theorem:

$$\Pr[\overline{E_2} \;\Big|\; V = v] < 2\exp\left(-2c \cdot \left(\zeta^2 - 1\right)\right).$$

Observe that

$$c = \max\left\{\frac{1}{v+1} + \frac{1}{m-v+1}, \frac{1}{t+1} + \frac{1}{m-t+1}\right\} \ge \frac{1}{t+1} + \frac{1}{m-t+1} \ge \frac{2}{\frac{m}{2}+1}.$$

Therefore,

$$\Pr[\overline{E_2} \;\Big|\; V = v] < 2\exp\left(\frac{-4}{\frac{m}{2}+1} \cdot \left(\zeta^2 - 1\right)\right). \tag{6}$$

Assume that both events E_1 and E_2 hold. Then in this case we will argue that:

$$\left|\frac{W}{V} - \frac{t-W}{m-V}\right| < \frac{m\zeta}{m^2/4 - \gamma^2} = \alpha. \tag{7}$$

We will then substitute the derived value of ζ into Equation 6 to show that $\Pr[\overline{E_2} \;\Big|\; V = v] < \mu/2$.

To this end, observe that if the number of blue-striped balls in the sample is within ζ of the expected value, $\frac{tV}{m}$, then the number of blue-striped balls in the sample is also within ζ of its own expected value, $\frac{t(m-V)}{m}$. This is because the balls can only be in one of these two sets. Therefore, when both events E_1 and E_2 hold, we have that:

$$\left|(t-W) - \frac{t(m-V)}{m}\right| < \zeta.$$

Therefore,

$$\begin{aligned}
&\frac{W}{V} - \frac{t-W}{m-V} \\
&\quad < \big(\mathbb{E}[W] + \zeta\big) \cdot \frac{1}{V} - \big(\mathbb{E}[t-W] - \zeta\big)\frac{1}{m-V} && \left(\text{because } |W - \frac{tV}{m}| < \zeta\right) \\
&\quad = \left(\frac{tV}{m} + \zeta\right) \cdot \frac{1}{V} - \left(\frac{t(m-V)}{m} - \zeta\right) \cdot \frac{1}{m-V} \\
&\quad = \frac{m\zeta}{V(m-V)} \\
&\quad < \frac{m\zeta}{m^2/4 - \gamma^2} && (\text{because } |V - m/2| < \gamma)\,.
\end{aligned}$$

Similarly,

$$\frac{t-W}{m-V} - \frac{W}{V} < \frac{m\zeta}{m^2/4 - \gamma^2}.$$

This gives us Equation 7. We can now set $\alpha = \frac{m\zeta}{m^2/4-\gamma^2}$ to get that $\zeta = \frac{3m\alpha}{16}$. Now, substituting this ζ value back into Equation 6 and using $m > 40$, $m > 16/3\alpha$, and $m > \frac{4.15}{\alpha^2}\ln(4/\mu)$ we have the following:

$$\begin{aligned}
\Pr[\overline{E_2} \;\big|\; V = v] &< \exp\left(\frac{-4}{\frac{m}{2}+1} \cdot \left(\zeta^2 - 1\right)\right) \\
&= 2\exp\left(\frac{-4}{\frac{m}{2}+1} \cdot \left(\frac{9m^2\alpha^2}{16^2} - 1\right)\right) \\
&\leq 2\exp\left(\frac{-8}{1.05m} \cdot \left(\frac{0.9 \cdot 9m^2\alpha^2}{16^2}\right)\right) \\
&= 2\exp\left(\frac{8.1\alpha^2 m}{33.6}\right) < \mu/2.
\end{aligned}$$

Finally, we get the following:

$$\Pr\left[\left|\frac{W}{V} - \frac{t-W}{m-V}\right| > \alpha\right] \leq \Pr[\overline{E_1} \vee \overline{E_2}] \leq \mu.$$

■

With Claims 1 and 2 now proved, this concludes the proof of Theorem 3. ■

Proof of Theorem 3. In the remaining part of this section, we use Lemma 1 to prove that any data curation algorithm $\mathcal{A}$ with bounded max-information also enforces Wasserstein-coherence.

Proof Of Theorem 3. Fix $\eta > 0$. Fix any subpopulation $C \in \mathcal{C}$. Consider any dataset X. Then for any algorithm $\mathcal{L} : \mathcal{Y} \to \{\mathcal{X} \to [-1,1]\}$, dataset $X \in \mathcal{X}^n$, $\alpha^* \in (0,1)$, and $\mu > 0$, such that

$$|X \cap C| \geq \max\left\{\frac{4.15 \cdot \ln(4/\mu)}{\alpha^{*2}}, \frac{5.3}{\alpha^*}, 8.3 \cdot \ln(4/\mu), 40\right\}, \tag{8}$$

we have the following (by Lemma 1)

$$\Pr_{X_a, X_b \leftarrow X, h \leftarrow \mathcal{L}\circ\mathcal{A}(X_a)}\left[\mathsf{dist}_{\mathcal{W}_1}\Big(h(\boldsymbol{\pi}_\rho\,(X_a|_C)),\ h(\boldsymbol{\pi}_\rho\,(X_b|_C))\Big) > 2\alpha^*\right] \leq 2(|X \cap C| + 1)\mu \cdot e^{\zeta} + \eta \tag{9}$$

where the choice of split $X_a, X_b \overset{\$}{\leftarrow} X$ and the computed predictor $h \leftarrow \mathcal{L}(X_a)$ are as in the $\mathsf{DemCoh}_{\mathcal{L}\circ\mathcal{A},X,\mathcal{C}^*,\rho}(\alpha)$ experiment in Figure 1.

We show, instead, that[14]

$$\gamma = \max\Big\{\frac{8.3 \cdot (\zeta + \ln(16|\mathcal{C}|/\beta))}{\alpha^{*2}}, \frac{36\ln((3/\alpha^*)}{\alpha^{*2}} \qquad , 16.6 \cdot (\zeta + \ln(16|\mathcal{C}|/\beta)), \frac{5.3}{\alpha^*}, 80\Big\},$$

and all algorithms $\mathcal{L} : \mathcal{Y} \to (\mathcal{X} \to [-1,1])$, all datasets $X \in \mathcal{X}^n$, the probability the following holds for all $C \in \mathcal{C}$ (such that $|X|_C| \geq \gamma$,) is low:

$$\mathsf{dist}_{\mathcal{W}_1}\left(h\big(\boldsymbol{\pi}_\rho\,(X_a|_C)\big),\ h\big(\boldsymbol{\pi}_\rho\,(X_b|_C)\big)\right) > 2\alpha^*,$$

where the split $X_a, X_b \overset{\$}{\leftarrow} X$ and predictor $h \leftarrow \mathcal{L}(X_a)$ are as in Figure 1.

To that end, we start by considering the fixed subpopulation C and setting $\mu = \eta/\left(2(|X \cap C| + 1) \cdot e^{\zeta}\right)$ (ensuring that the RHS of Equation (9) is 2 η).

The main content in the proof will be arguing that the following lower bound on the size of $|X \cap C|$ implies the condition in Equation (8). We can then union bound over all sub-populations in $\mathcal{C}$ to get the theorem.

$$|X \cap C| \geq \max\Big\{\frac{8.3 \cdot (\zeta + \ln(16/\eta))}{\alpha^{*2}}, \frac{36\ln(3/\alpha^*)}{\alpha^{*2}} , 16.6 \cdot (\zeta + \ln(16/\eta)), \frac{5.3}{\alpha^*}, 80\Big\}. \tag{10}$$

Substituting the value of μ back in Equation (8), the first term in the max corresponds to the condition

$$|X \cap C| \geq \frac{4.15 \cdot \ln(8(|X \cap C| + 1)e^{\zeta}/\eta)}{\alpha^{*2}}$$

which can also be written as:

$$|X \cap C| \geq \frac{4.15 \cdot \ln((|X \cap C| + 1) + g}{\alpha^{*2}}$$

where $g = 4.15 \cdot (\zeta + \ln(8/\eta))$.

Assume $\frac{|X \cap C|}{2} \geq \frac{4.15 \cdot \ln((|X \cap C|+1)}{\alpha^{*2}}$. Then, as long as $|X \cap C| \geq \frac{2g}{\alpha^{*2}}$, the condition is satisfied. Now, using the fact that $|X \cap C| \geq 80$, we have that $\ln((|X \cap C| + 1) \leq 1.01\ln((|X \cap C| + 1)$, which

[14] Looking ahead, we set $\alpha^* = \alpha/2$ to get the final theorem.

implies that it suffices for $|X \cap C| \geq \frac{9 \cdot \ln((|X \cap C|)}{\alpha^{*2}}$. Consider the inequality $\frac{|X \cap C|}{\ln((|X \cap C|)} \geq c$. Note that the left hand side is an increasing function of $|X \cap C|$. Let $|X \cap C| \geq 2c \ln c$. Then, we get that $\frac{|X \cap C|}{\ln((|X \cap C|)} \geq \frac{2c \ln c}{\ln(2c \ln c)}$, and some arithmetic shows that for $c \geq 9$ (which is true whenever $\alpha^* \leq 1$), the right hand side is indeed larger than c. Hence, it is additionally sufficient that $|X \cap C| \geq \frac{36 \ln((3/\alpha^*)}{\alpha^{*2}}$.

Similarly, to be larger than the third term in the max in Equation (8), we need

$$|X \cap C| \geq 8.3 \cdot \ln(8(|X \cap C| + 1)e^{\zeta}/\eta)$$

which can also be written as

$$|X \cap C| \geq 8.3 \cdot \ln((|X \cap C| + 1) + g$$

where $g = 8.3 \cdot (\zeta + \ln(8/\eta))$.

Assume $\frac{|X \cap C|}{2} \geq 8.3 \cdot \ln((|X \cap C| + 1)$. Then, as long as $|X \cap C| \geq 2g$, the condition is satisfied. Now, using the fact that $|X \cap C| \geq 80$, we have that $\ln((|X \cap C| + 1) \leq 1.01 \ln((|X \cap C| + 1)$, which implies that it suffices for $|X \cap C| \geq 18 \cdot \ln((|X \cap C|)$,which is true as long as $|X \cap C| \geq 80$, hence this case is taken care of.

Hence, for a single subpopulation C we have argued that it is sufficient that

$$|X \cap C| \geq \max\Big\{\frac{8.3 \cdot (\zeta + \ln(8/\eta))}{\alpha^{*2}}, \frac{36 \ln(3/\alpha^*)}{\alpha^{*2}} , 16.6 \cdot (\zeta + \ln(8/\eta)), \frac{5.3}{\alpha^*}, 80\Big\}. \tag{11}$$

Setting $\alpha^* = \alpha/2$, $\eta = \beta/2|\mathcal{C}|$, and applying a union bound on Equation (9) (over all subpopulations in the class) then gives the theorem. ■

6.2 Differentially Private Algorithms Enforce Demographic Coherence

In this section, we state formal connections between differential privacy and demographic coherence. We do this by adapting known connections between differential privacy and max-information and Theorem 3 connecting max-information and demographic coherence.

Definition 9 (Differential Privacy [25])**.** *Let $n \in \mathbb{N}$. A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ is (ϵ, δ)-*differentially private *if for all subsets $Y \subseteq \mathcal{Y}$ of the output space, and for all neighboring datasets $X, X' \in \mathcal{X}^n$ (i.e. $\|X - X'\|_0 \leq 1$), we have that*

$$\Pr[\mathcal{A}(X) \in Y] \leq e^{\varepsilon} \Pr[\mathcal{A}(X') \in Y] + \delta$$

If $\delta = 0$, we refer to the algorithm as satisfying pure differential privacy (pure-DP), whereas $\delta > 0$ corresponds to approximate DP (approx-DP).

The proofs of the following theorems connecting differential privacy and max-information can be found in Appendix B.

Theorem 5. (Pure-DP $\implies$ Bounded Max-Information) *Fix $n \in \mathbb{N}$, $\varepsilon > 0$ and let $\mathcal{X}$ be a data universe of size at least n. Let $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$ be an order-invariant ε-DP algorithm. Then for any $\gamma > 0$,*

$$I_{\infty}^{\gamma}(\mathcal{A}, n/2) \leq \varepsilon^2 n/4 + \varepsilon\sqrt{n \ln(2/\gamma)/4}.$$

The following theorem is a generalized version of that in [50]. The proof follows theirs, with the following key distinctions: (1) it applies to sampling without replacement (2) It carefully tracks constants and (3) It maintains flexibility in setting parameters. We anticipate that this version of the result might be independently useful.

Theorem 6. (Approx-DP $\implies$ Bounded Max-Information, Generalised) *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an (order-invariant) (ε, δ)-differentially private algorithm for $\varepsilon \in (0, 1/2]$, $\delta \in (0, \varepsilon)$. For $\hat{\delta} \in (0, \varepsilon/15]$, $t > 0$, and $\beta(t, \hat{\delta}) = e^{-t^2/2} + n\left(\frac{2\delta}{\hat{\delta}} + \frac{2\hat{\delta}+2\delta}{1-e^{-3\varepsilon}}\right)$ we have*

$$I_\infty^\beta(\mathcal{A}, n) \leq n\left(347\hat{\delta} + 75\left(\frac{\hat{\delta}}{\varepsilon}\right)^2 + 24\frac{\hat{\delta}^2}{\varepsilon} + 240\varepsilon^2\right) + 6t\varepsilon\sqrt{n}.$$

Corollary 1. (Approx-DP $\implies$ Bounded Max-Information, Specific) *Fix $n \in \mathbb{N}$, for $\varepsilon \in (0, 1/2]$, $\gamma \in (0, 1]$, $\delta \in (0, \frac{\varepsilon^2\gamma^2}{(120n)^2}]$ and let $\mathcal{X}$ be a data universe of size at least n. Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an (order-invariant) (ε, δ)-differentially private algorithm. We have that*

$$I_\infty^\gamma(\mathcal{A}, n) \leq 265\varepsilon^2 n + 12\varepsilon\sqrt{n\ln(2/\gamma)}.$$

The proofs for pure differential privacy and approximate differential privacy follow a similar flavor. Firstly, we adapt known connections between (pure and approximate) differential privacy and max-information to the setting of sampling without replacement. Then, we use Theorem 3 (connecting bounded max-information to demographic coherence) to argue that differential privacy implies demographic coherence.

Theorem 7. *[Pure-DP Enforces Wasserstein Coherence] Fix any $\varepsilon, \beta \in (0, 1]$, $\alpha \in (0, 2], n \in \mathbb{N}$. Let $\mathcal{C}$ be a collection of subpopulations $C \in \mathcal{X}^*$. Consider an order-invariant ε-DP algorithm $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$. Fix any lens ρ. Then, $\mathcal{A}$ enforces (α, β)-Wasserstein-coherence with respect to collection $\mathcal{C}$, lens ρ, and size constraint γ, where γ is the maximum of the following terms*

$$\begin{aligned}\gamma = \max\Big\{&\frac{33.2 \cdot (\varepsilon^2 n/4 + \varepsilon\sqrt{n\ln(4|\mathcal{C}|/\beta)}/2 + \ln(16|\mathcal{C}|/\beta))}{\alpha^2}, \frac{144\ln((6/\alpha)}{\alpha^2} \\ &, 16.6 \cdot (\varepsilon^2 n/4 + \varepsilon\sqrt{n\ln(4|\mathcal{C}|/\beta)}/2 + \ln(16|\mathcal{C}|/\beta)), \frac{10.6}{\alpha}, 80\Big\}.\end{aligned} \tag{12}$$

Proof. Fix $\beta > 0$. By Theorem 14 connecting differential privacy and max-information, we have that we have that,

$$I_\infty^{\beta/2|\mathcal{C}|}(\mathcal{A}, n/2) \leq \varepsilon^2 n/4 + \varepsilon\sqrt{n\ln(4|\mathcal{C}|/\beta)/4}.$$

Applying Theorem 3 and substituting the bound on max-information then completes the proof. ■

Theorem 8. *[Approx-DP Enforces Wasserstein Coherence] Fix any $\beta \in (0, 1]$, $\alpha \in (0, 2)$, $n \in N$. Let $\varepsilon \in (0, \frac{1}{2}]$, and $\delta \in (0, \frac{\varepsilon^2\beta^2}{(120n)^2|\mathcal{C}|^2}]$ Let $\mathcal{C}$ be a collection of subpopulations $C \in \mathcal{X}^*$. Consider an order-invariant [15] (ε, δ)-DP algorithm $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$. Fix any lens ρ. Then, $\mathcal{A}$ enforces (α, β)-Wasserstein-coherence with respect to collection $\mathcal{C}$, lens ρ, and size constraint γ, where γ is the maximum of the following terms*

$$\gamma = \max\Big\{\frac{33.2 \cdot (133\varepsilon^2 n + 12\varepsilon\sqrt{n\ln(4|\mathcal{C}|/\beta)} + \ln(32|\mathcal{C}|/\beta))}{\alpha^2}, \frac{144\ln(6/\alpha)}{\alpha^2}$$

[15] Order-invariance can be relaxed by multiplying ε by 2 in the γ value, and dividing by 2 in the range of δ.

$$, 16.6 \cdot (133\varepsilon^2 n + 12\varepsilon\sqrt{n \ln(4|\mathcal{C}|/\beta)} + \ln(32|\mathcal{C}|/\beta)), \frac{10.6}{\alpha}, 80\Big\}. \tag{13}$$

Proof. Fix $\beta > 0$ and $\gamma = \frac{\beta}{2|\mathcal{C}|}$. By Corollary 2 connecting differential privacy and max-information, we have that we have that, as long as $\delta \in (0, \frac{\varepsilon^2\beta^2}{(120n)^2|\mathcal{C}|^2}]$

$$I_\infty^{\beta/2|\mathcal{C}|}(\mathcal{A}, n/2) \leq \frac{265}{2}\varepsilon^2 n + 12\varepsilon\sqrt{\frac{n}{2}\ln(4|\mathcal{C}|/\beta)}.$$

Applying Theorem 3 and substituting the bound on max-information then completes the proof.

■

6.3 Using our Theorems for Setting DP Parameters

Recall that one of our goals in designing Demographic Coherence was to support data curators as they go through the process of parameter selection, as described in Section 1. Namely, by providing a *necessary* privacy condition that is more concrete, there is a hope that data curators can make parameter decisions in a contextualized way. In this section, we leverage Theorem 7 to illustrate how this might work for a data curator interested in using DP, using the American Community Survey (ACS) as a motivating example. The ACS is an randomized annual survey that provides in-depth demographic and housing information on respondents, including demographics, employment, income, taxes, living conditions, insurance, etc. . . The sample includes ≈ 3.5 million households each year, resulting in data about ≈ 5 million individuals included in the final data set. The collected data is crucial input for many, including local governments, community organizers, researchers, and businesses, but the in-depth nature of the collected data comes with significant privacy concerns. Moreover, there are ongoing conversations about how DP should be used with ACS data products [19].

Using our framework requires a collaborative effort between socio-technical researchers, privacy experts, legal experts, and data curators to identify concrete privacy risks that may accrue from a data release. For example, imagine that the data curators for the ACS are concerned that the data they release might be used by hiring employers to identify job candidates who have dependents (and therefore might be more likely to miss work to care for those dependents). Directly asking candidates for this information may be illegal in some circumstances, but can also be predictive of the subset of job candidates the employer might prefer. The ACS gathers detailed information about household composition, which could leak information about the dependent responsibilities of candidates within the sample.

Having selected a concrete predictive harm, demographic coherence asks the data curator to perform parameter selection for this harm within the demographic coherence experiment. Specifically, we need to select values for α and β. For example, the curator might decide that predictive algorithms should not "implicitly" reconstruct more than 10% of the data subjects' dependents status. This can be seen as having 10% of the survey population for whom the predictor would have predicted -1 based on statistical aggregate information, but the predictor outputs 1 instead.[16] If a predictive algorithm were able to conduct such an implicit reconstruction, it would result in a Wasserstein distance of at least 0.2 between the prediction distributions on the data subjects and

[16]Because demographic coherence is independent of accuracy, the model's baseline assumption could equivalently be -1. Alternatively, a model might have a good estimate ($\pm p$ for $p \in [0, 1]$) for the population-level base-rate due to previous ACS survey results.

α	β	$\lvert\mathcal{C}\rvert$	γ	ε
0.1	e^{-30}	100	$.05n$	.003
0.2	e^{-30}	100	$.05n$	.01
0.3	e^{-30}	100	$.05n$	.02
0.4	e^{-30}	100	$.05n$	.025
0.8	e^{-30}	100	$.05n$	0.056
0.9	e^{-30}	100	$.05n$	0.065
1.3	e^{-30}	100	$.05n$	0.09
1.4	e^{-30}	100	$.05n$	0.1

Table 1: Parameters examples that satisfy Theorem 7 when $n = 5 \times 10^6$ (the ACS dataset size).

another similarly distributed set of individuals,[17] as .1 of the total mass of the probability distribution would be moved 2 units (from -1 to 1). framework, for example a leakage of whether more than 10 percent of people have dependents would correspond to a Wasserstein distance larger than 0.2 between the predictive distributions in our framework, which means one may aim to set this value (α) less than 0.2 to eliminate this risk. Thus, a data curator concerned about such a harm may want to set α to .2 (or less) to eliminate such a risk. The curator can then set β to be a small constant, like e^{-30}, much as δ is set to a small constant in approximate DP.

Finally, demographic coherence asks the data curator to fix a category class $\mathcal{C}$ (with associated minimum size of a category γ) for which this privacy guarantee would hold. These might be demographic communities against whom there is a long history of discrimination, e.g., black women. Alternatively, they might be tailored to the specific set of harms about with the data curator is concerned. For example, men between the ages of 18 and 25 might be particularly unlikely to have dependents (ie., a good model might default to predictions of nearly -1 for this population) and therefore might have their job prospects particularly impact by disclosure of their dependent status. This later cohort constitutes about 5% of the US population (and, therefore, should also constitute about 5% of the ACS sample). The data curator team could collaboratively identify many such group of interest, attempting to be as comprehensive as possible. For the sake of this example, let us assume there are about 100 potential categories within $\mathcal{C}$, each constitution 5% of the dataset ($\gamma \geq 250,000$).

Given these choices of α, β, and $|\mathcal{C}|$ (with associated γ), we can leverage Theorem 7 to identify a value of ϵ needed such that an ε-DP algorithm ensures (α, β)-demographically coherent predictions with respect to $\mathcal{C}$ and size constraint γ. Plugging in the value above, we find that $\varepsilon = 0.01$ is sufficient. We note, however, that Theorem 7 does not imply that $\varepsilon = 0.01$ is *necessary*, as our theorem is not optimal; future work is likely to show that a higher value of ε may also be sufficient. We give other sets of satisfying parameters for Theorem 7 in Table 1. We note that this value of ε also corresponds to a certain level of utility (for various data analysis tasks such as linear regressions, clustering, parameter estimation etc.). Hence our framework precipitates concrete tradeoffs between privacy concerns and desired utility guarantees for the intended applications, which can then be the subject of robust conversation in order to choose a value of ε that best balances these considerations.

The above illustration shows that the process of parameter selection can be made more concrete

[17] Formally, we consider a random dataset split. Informally, however, it is reasonable to think about demographic coherence as comparing the performance of a predictor on individuals in the dataset and unobserved individuals.

and tangible using demographic coherence. Challenges, however, remain. For example, there is nothing inherently good or bad about 10%, meaning that ambiguity will remain in the process. But, this is significantly more concrete than existing approaches in the literature.

Moreover, the experiments in Section 5 suggest natural experimental tools that can help data curators get a clearer sense for what any given value of α *feels* like - they can choose and train natural learning algorithms on data releases, and plot the corresponding prediction distributions of the associated predictor as well as compute the Wasserstein distance. Observing the value of Wasserstein distance at which the prediction distribution seem disconcertingly different visually can help build intuition about setting α.

7 Non-Differentially Private Algorithms that Enforce Wasserstein Demographic Coherence

In this section, we analyze an algorithm that is not differentially private ($\varepsilon = \infty$), but satisfies excellent demographic coherence parameters, thereby showing that our framework is provably weaker than differential privacy. This also lends support to the encouraging possibility that it is possible to achieve a better privacy-utility tradeoff for many tasks with demographic coherence (in conjunction with other necessary conditions) than with differential privacy.

We consider an algorithm that releases a single (shuffled) binary column of a dataset (with possibly many columns) in the clear. Such a curator can be thought of as publishing the outcome of a vote without adding noise, thereby allowing for the accurate evaluation of the outcomes of such votes even when they are very close (which would be impossible with differential privacy).

This algorithm is clearly not differentially private since the algorithm reproduces its input exactly, but we will argue that it enforces Wasserstein demographic coherence with good parameters, owing to the fact that a single binary column with no additional context reveals essentially no information about the split (since the 1s and 0s in the column are partitioned roughly equally between the two halves of the split and since the information released is uncorrelated with other attributes due to shuffling). In particular, the data released does not have enough complexity to allow for a violation of demographic coherence.

Fix a universe $\mathcal{X}$. Consider a data curator $\mathcal{A} : (\mathcal{X}\times\{0,1\})^{n/2} \to \{0,1\}^{n/2}$ that (deterministically) publishes a binary column of its input in the clear.

We will start by considering the lens of the adversary at prediction time being just the column released, hence ensuring there is an absence of information correlated to the bit (*e.g.,* personal identifiers) that the adversary can use.

Theorem 9. *Fix any $\beta \in (0,1]$, $\alpha \in (0,2], n \in \mathbb{N}$. Let $\mathcal{X}$ be an arbitrary space, and let $\mathcal{C}$ be a collection of subpopulations $C \in \{\mathcal{X} \times \{0,1\}\}^*$. Consider algorithm $\mathcal{A} : (\mathcal{X} \times \{0,1\})^{n/2} \to \{0,1\}^{n/2}$ that publishes one binary column of the input in the clear. Let the lens ρ be this one column.*

Then, $\mathcal{A}$ enforces (α, β)-Wasserstein-coherence with respect to collection $\mathcal{C}$, lens ρ, and size constraint $\gamma = \frac{4\sqrt{n\log(|\mathcal{C}|/\beta)}}{\alpha}$.

Proof. Fix any category $C \in \mathcal{C}$. Fix any dataset $X \in \{0,1\}^n$ (we will restrict to the one column that is being released since the others will not matter). Let $\beta' = \frac{\beta}{|\mathcal{C}|}$.

We start by stating a concentration inequality we will use in the proof.

Theorem 10 (Hoeffding's Inequality Without Replacement)**.** *Let $R = (r_1, \ldots, r_n) \in \{0,1\}^N$ be a finite population, and let $R_1, \ldots, R_m$ be a random sample of size m drawn without replacement from* *it. Let* $\mu = \frac{1}{n}\sum_{i=1}^{n} R_i$*. Then,*

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} R_i - \mu \geq \alpha\right] \leq e^{-2m\alpha^2}.$$

Now, consider a random split $X_a, X_b \leftarrow X$ of the dataset. Let N_a be the number of ones in category C in X_a, and N_b be the number of ones in category C in X_b and M_a and M_b be the number of zeros in category C in X_a and X_b respectively. Setting $m = n/2$, $\alpha = \sqrt{\frac{\log(1/\beta')}{n}}$, and using Theorem 10 where we let $R_i = 1[X_i \in C, X_i = 1]$, and consider the dataset of R_i values and apply the random split to this.

$$\Pr\left[\sum_{i \in X_a} R_i - \frac{n}{2}\mu \geq \sqrt{n \log(1/\beta')}\right] \leq \beta'.$$

Observe that this implies that N_a is at most $\sqrt{n\log(1/\beta')}$ from half the number of ones in the category C in the dataset, which implies that $|N_a - N_b| \leq 2\sqrt{n\log(1/\beta')}$ with probability at least $1-\beta'$. Also, note that $|N_a - N_b| = |M_a - M_b|$.

Now, consider any predictor h on the elements of the dataset- which can be captured by the values $h(0)$ and $h(1)$.

Then, with probability at least $1-\beta'$ over the split X_a, X_b, (and also using the fact that $|N_a - N_b| = |M_a - M_b|$),

$$\mathsf{dist}_{\mathcal{W}_1}\Big(h(\boldsymbol{\pi}_\rho\,(X_a|_C)),\, h(\boldsymbol{\pi}_\rho\,(X_b|_C))\Big) = |h(1) - h(0)| \cdot \frac{|N_a - N_b|}{|C \cap X|} \leq 2\frac{|N_a - N_b|}{|C \cap X|} \leq \frac{4\sqrt{n\log(1/\beta')}}{|C \cap X|}$$

Since this applies for all predictors, it applies in particular for any predictor produced by the algorithm $\mathcal{L}$.

Now, taking a union bound over all categories $C \in \mathcal{C}$, we get that with probability at least $1-\beta$ over the split X_a, X_b and the randomness of the algorithm $\mathcal{L}$ generating the predictor h,

$$\mathsf{dist}_{\mathcal{W}_1}\Big(h(\boldsymbol{\pi}_\rho\,(X_a|_C)),\, h(\boldsymbol{\pi}_\rho\,(X_b|_C))\Big) \leq \frac{4\sqrt{n\log(|\mathcal{C}|/\beta)}}{|C \cap X|}$$

Hence, we get that $\mathcal{A}$ enforces (α, β)-demographic coherence with respect to collection $\mathcal{C}$, lens ρ, and size constraint γ, where $\gamma = \frac{4\sqrt{n\log(|\mathcal{C}|/\beta)}}{\alpha}$.

■

Now, we will handle the more general case where the lens of the adversary is the entire input space- that is, the adversary sees all attributes of the input when making a prediction.

The sanitization algorithm will involve shuffling the column before outputting it as the report (or equivalently outputting the number of 1s in the column- as we will think about it in our analysis). Intuitively, the report should still not give the adversary much information about the split, since it only tells the adversary how many records with bit 1 in a specific column there are in X_a, but not *which* records they are.

However, the analysis in this case is significantly trickier and involves careful conditioning arguments and analysis of hypergeometric distributions. Consequently, the size constraint is slightly worse (by a logarithmic factor in the dataset size), and also has worse constants.

Theorem 11. *Fix any $\beta' \in (0,1]$, $\alpha \in (0,2]$, sufficiently large $n \in \mathbb{N}$. Let $\mathcal{X}$ be an arbitrary space, and let $\mathcal{C}$ be a collection of subpopulations $C \in \{\mathcal{X} \times \{0,1\}\}^*$. Consider algorithm $\mathcal{A} : (\mathcal{X} \times \{0,1\})^{n/2} \to \mathbb{N}$ that publishes the number of 1s in the binary column in the input. Let the lens ρ be the entire input space $\mathcal{X} \times \{0,1\}$.*

Then, $\mathcal{A}$ enforces (α, β)-Wasserstein-coherence with respect to collection $\mathcal{C}$, lens ρ, and size constraint $\gamma = \frac{112\sqrt{n\log(10n|\mathcal{C}|/\beta')}}{\alpha}$.

Proof. Let the binary column whose number of 1s is released in the clear be column i of the dataset. Fix a subpopulation $C \in \mathcal{C}$. Let the number of 1s in this column in the population dataset be L_1 and the number of 0s be L_2. Let N_a be the number of ones sampled into X_a and N_b be the number of zeros sampled into X_a. By Hoeffding's inequality without replacement (Theorem 10), we have that with probability at least $1-\beta$, $|N_a - L_1/2| \le \sqrt{n\log(1/\beta)}$ and $|N_b - L_2/2| \le \sqrt{n\log(1/\beta)}$ (where we have also used that $N_a + N_b = n/2$).

Now, condition on a specific report N_a that satisfies these inequalities $|N_a - L_1/2| \le \sqrt{n\log(1/\beta)}$ and $|N_b - L_2/2| \le \sqrt{n\log(1/\beta)}$. Conditioned on this, the split (X_a, X_b) can be thought of in the following way: X_a is generated as follows: N_a records with column i value being one are sampled without replacement from the L records with this property in the dataset and $N_b = n/2 - N_a$ records with column i value being zero are sampled without replacement from L_2 records satisfying this property in the dataset. X_b is then simply $X \setminus X_a$.

Now, consider an adversary producing a predictor h based on the report N_a. Condition on the coins of the adversary. Observe that this fixes the corresponding predictor h - since all the randomness generating h has been handled via conditioning (we have already conditioned on the report N_a and have also now conditioned on the random coins of the algorithm).

We will now reason about the demographic coherence properties of h.

$$
\begin{aligned}
&\mathsf{dist}_{\mathcal{W}_1}\left(h(\boldsymbol{\pi}_\rho\left(X_a|_C\right)), h(\boldsymbol{\pi}_\rho\left(X_b|_C\right))\right) \\
&\quad = \int_{-1}^{1} |\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_a|_C))}(\ell) - \mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_b|_C))}(\ell)|\, d\ell \qquad (14) \\
&\quad \le \int_{-1}^{1} \Big[|\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_a|_C))}(\ell) - \mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X|_C))}(\ell)| \\
&\quad + |\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_b|_C))}(\ell) - \mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X|_C))}(\ell)|\Big]\, d\ell \qquad (15)
\end{aligned}
$$

Fix a specific $\ell \in [-1,1]$. We will leverage the properties of the hypergeometric distribution in order to reason about the CDF difference in question, and then use this to bound the Wasserstein difference of interest.

We redefine the sampling process in a way that will allow us to use Theorem 4, a concentration result for the hypergeometric distribution (See Definition 8 for a definition of this distribution): Consider an urn consisting of L_1 balls representing records whose *ith* column is 1. Among those L_1 balls, m_1 are marked with a red stripe, representing membership in $C \cap X$. Among the m_1 red-striped balls, t_1 are further marked with a blue stripe, representing $x \in C \cap X$ such that $h(\boldsymbol{\pi}_\rho(x)) \le \ell$ for the value of ℓ being considered. Consider the experiment where we sample N_a balls without replacement, and define the joint pair of random variables (V_1, W_1) where V_1 counts the number of red-striped balls in the sample, (i.e., the number of sampled points that are in $X \cap C$) and W_1

counts the number of (red and) blue-striped balls in the sample, (i.e., the number of sampled points x that are in $X \cap C$ and satisfy $h(\boldsymbol{\pi}_\rho(x)) \leq y$. Similarly, consider an analogous case with another urn with L_2 balls (representing records with i^{th} column equal to 0), m_2 balls marked with a red stripe (indicating membership in subgroup), t_2 balls further marked with a blue stripe (indicating hypothesis value smaller than ℓ), and N_b balls sampled without replacement, with (W_2, V_2) defined analagously.

The random variables W_1, V_1, W_2, V_2 follow hypergeometric distributions as follows:

$$
\begin{aligned}
V_1 &\sim H(L_1, m_1, N_a) \\
(W_1 | V_1 = v_1) &\sim H(m_1, t_1, v_1). \\
V_2 &\sim H(L_2, m_2, N_b) \\
(W_2 | V_2 = v_2) &\sim H(m_2, t_2, v_2).
\end{aligned}
$$

Observe that the absolute value of the CDF difference we are trying to bound is equal to $\left|\frac{W_1+W_2}{V_1+V_2} - \frac{t_1+t_2}{m_1+m_2}\right|$ by definition.

Let E_1 be the event that the number of red-striped balls in the sample is close to its expected value for Urn 1. (i.e., $|V_1 - m_1 N_a / L_1| < \sqrt{n \log(1/\beta)}$). Then applying Theorem 4 and using the fact that n is sufficiently large,

$$
\begin{aligned}
\Pr[\overline{E_1}] &< 2 \exp\left(\frac{-2}{N_a + 1} \cdot \left(n \log(1/\beta) - 1\right)\right) \\
&\leq \beta.
\end{aligned}
$$

Let E_2 be the same event for Urn 2. Hence, by the same argument, we get that with probability at least $1 - \beta$, $|V_2 - m_2 N_b / L_2| < \sqrt{n \log(1/\beta)}$.

Now let us condition on a realization $V_1 = v_1$. Given this, let G_1 be the event that the number of blue-striped balls in the sample for Urn 1 is close to its expected value (i.e., $\left|W_1 - \frac{t_1 v_1}{m_1}\right| \leq \zeta$). Then applying Theorem 4 for $\zeta \geq 2 = \sqrt{n \log 1/\beta}$ and c as in the theorem, and using the fact that n is sufficiently large:

$$
\Pr[\overline{E_2} \, \Big| V = v] < 2 \exp\left(-2c \cdot \left(\zeta^2 - 1\right)\right).
$$

Observe that

$$
\begin{aligned}
c &= \max\left\{\frac{1}{v_1 + 1} + \frac{1}{m_1 - v_1 + 1}, \frac{1}{t_1 + 1} + \frac{1}{m_1 - t_1 + 1}\right\} \\
&\geq \frac{1}{t_1 + 1} + \frac{1}{m_1 - t_1 + 1} \geq \frac{2}{\frac{m_1}{2} + 1}.
\end{aligned}
$$

Therefore,

$$
\Pr[\overline{G_1} \, \Big| V = v] < 2 \exp\left(\frac{-4}{\frac{m_1}{2} + 1} \cdot \left(\zeta^2 - 1\right)\right) \leq \beta. \tag{16}
$$

Similarly, if we define G_2 similarly for Urn 2 and condition on $V_2 = v_2$, we get that

$$\Pr[\overline{G_2} \mid V = v] < 2\exp\left(\frac{-4}{\frac{m_2}{2}+1}\cdot\left(\zeta^2-1\right)\right) \le \beta. \tag{17}$$

Now, substituting these values into the CDF difference, we can write that with probability at least $1-5\beta$,

$$\begin{aligned}
&\frac{W_1+W_2}{V_1+V_2}\\
&\le \frac{W_1+W_2}{m_1N_a/L_1+m_2N_b/L_2-2\sqrt{n\log 1/\beta}}\\
&\le \frac{W_1+W_2}{m_1/2+m_2/2-4\sqrt{n\log(1/\beta)}}\\
&\le \frac{t_1V_1/m_1+t_2V_2/m_2+2\sqrt{n\log 1/\beta}}{m_1/2+m_2/2-4\sqrt{n\log(1/\beta)}}\\
&\le \frac{t_1V_1/m_1+t_2V_2/m_2+2\sqrt{n\log 1/\beta}}{m_1/2+m_2/2-4\sqrt{n\log(1/\beta)}}\\
&\le \frac{t_1N_a/L_1+t_2N_b/L_2+4\sqrt{n\log 1/\beta}}{m_1/2+m_2/2-4\sqrt{n\log(1/\beta)}}\\
&\le \frac{t_1/2+t_2/2+6\sqrt{n\log 1/\beta}}{m_1/2+m_2/2-4\sqrt{n\log(1/\beta)}}
\end{aligned}$$

Substituting in the CDF difference, we get that

$$\begin{aligned}
&\frac{W_1+W_2}{V_1+V_2}-\frac{t_1+t_2}{m_1+m_2}\\
&\le \frac{t_1+t_2+12\sqrt{n\log 1/\beta}}{m_1+m_2-8\sqrt{n\log(1/\beta)}}-\frac{t_1+t_2}{m_1+m_2}\\
&\le \frac{t_1+t_2+12\sqrt{n\log 1/\beta}}{m_1+m_2-8\sqrt{n\log(1/\beta)}}-\frac{t_1+t_2-8\sqrt{n\log(1/\beta)}}{m_1+m_2-8\sqrt{n\log(1/\beta)}}\\
&\le \frac{20\sqrt{n\log 1/\beta}}{m_1+m_2-8\sqrt{n\log(1/\beta)}} \le \alpha/4
\end{aligned}$$

when $|C\cap X| = m_1+m_2 \ge \frac{112\sqrt{n\log(1/\beta)}}{\alpha}$.

Similarly, with probability at least $1-5\beta$,

$$\begin{aligned}
&\frac{W_1+W_2}{V_1+V_2}\\
&\ge \frac{W_1+W_2}{m_1N_a/L_1+m_2N_b/L_2+2\sqrt{n\log 1/\beta}}\\
&\ge \frac{W_1+W_2}{m_1/2+m_2/2+4\sqrt{n\log(1/\beta)}}
\end{aligned}$$

$$
\begin{aligned}
&\geq \frac{t_1 V_1/m_1 + t_2 V_2/m_2 - 2\sqrt{n\log 1/\beta}}{m_1/2 + m_2/2 - 4\sqrt{n\log(1/\beta)}}\\
&\geq \frac{t_1 V_1/m_1 + t_2 V_2/m_2 - 2\sqrt{n\log 1/\beta}}{m_1/2 + m_2/2 - 4\sqrt{n\log(1/\beta)}}\\
&\geq \frac{t_1 N_a/L_1 + t_2 N_b/L_2 - 4\sqrt{n\log 1/\beta}}{m_1/2 + m_2/2 - 4\sqrt{n\log(1/\beta)}}\\
&\geq \frac{t_1/2 + t_2/2 - 6\sqrt{n\log 1/\beta}}{m_1/2 + m_2/2 + 4\sqrt{n\log(1/\beta)}}
\end{aligned}
$$

Substituting in the CDF difference, we get that

$$
\begin{aligned}
&\frac{t_1+t_2}{m_1+m_2} - \frac{W_1+W_2}{V_1+V_2}\\
&\leq \frac{t_1+t_2}{m_1+m_2} - \frac{t_1+t_2-12\sqrt{n\log 1/\beta}}{m_1+m_2+8\sqrt{n\log(1/\beta)}}\\
&\leq \frac{t_1+t_2+8\sqrt{n\log 1/\beta}}{m_1+m_2+8\sqrt{n\log(1/\beta)}}\\
&- \frac{t_1+t_2-12\sqrt{n\log(1/\beta)}}{m_1+m_2+8\sqrt{n\log(1/\beta)}}\\
&\leq \frac{20\sqrt{n\log 1/\beta}}{m_1+m_2-8\sqrt{n\log(1/\beta)}} \leq \alpha/4
\end{aligned}
$$

when $|C \cap X| = m_1 + m_2 \geq \frac{112\sqrt{n\log(1/\beta)}}{\alpha}$.

Hence, we get that with probability at least $1-5\beta$, $|\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_a|_C))}(\ell) - \mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X|_C))}(\ell)| \leq \frac{\alpha}{4}$. By a symmetric argument, with probability at least $1-5\beta$, $|\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_b|_C))}(\ell) - \mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X|_C))}(\ell)| \leq \frac{\alpha}{4}$, giving that with probability at least $1-10\beta$, $|\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_a|_C))}(\ell) - \mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_b|_C))}(\ell)| \leq \frac{\alpha}{2}$. Now, observe that the CDF difference can change at at most n points, since datasets X_a, X_b each contain $n/2$ points. Hence, setting $\beta = \beta'/10n|\mathcal{C}|$, and using a union bound, we get that with probability at least $1-\beta'/|\mathcal{C}|$, for all $\ell \in [-1,1]$, $|\mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_a|_C))}(\ell) - \mathsf{cdf}_{h(\boldsymbol{\pi}_\rho(X_b|_C))}(\ell)| \leq \frac{\alpha}{2}$. Hence, we get that with probability at least $1-\beta'/|\mathcal{C}|$ over the randomness of the algorithm and the split, $\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho(X_a|_C)), h(\boldsymbol{\pi}_\rho(X_b|_C))) \leq \alpha$. Hence, by a union bound, we get that with probability at least $1-\beta'$ over the randomness of the algorithm and the split, for all categories $C \in \mathcal{C}$, $\mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho(X_a|_C)), h(\boldsymbol{\pi}_\rho(X_b|_C))) \leq \alpha$. Further substituting the value of β into the restriction on $|C \cap X|$, we get that $|C \cap X| \geq \frac{112\sqrt{n\log(10n|\mathcal{C}|/\beta)}}{\alpha}$ is sufficient as the size constraint, completing the proof. ■

Given our results, one might be tempted to conjecture that the above results can be extended to releasing multiple columns in the clear. However, the amount of information released about the split grows rapidly with the number of columns released in the clear, and so this intuition breaks down. In fact, in Section 5, we took a real dataset with many integer-valued columns and demonstrated a predictor that witnesses very large Wasserstein distance between the two predictive distributions restricted to a specific relevant subgroup, thereby indicating that releasing entire datasets in the clear is not coherence enforcing. When curators want to release more complex statistics than a single vote, our results in Section 6 show that differentially private algorithms give us an avenue to do this. We leave it as an interesting open question to investigate other ways to achieve demographic coherence.

8 Acknowledgments

We would like to thank Adam Smith for insightful comments and feedback that helped improve this work. We thank Christina Xu for her contributions to shaping the early phases of this project.

References

[1] John M. Abowd. The U.S. Census Bureau adopts Differential Privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2867, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Adrian Dixon. 3 Resume Screening Tools That Every Recruiter Should Know About. `https://ideal.com/resume-screening-tools/`., November 2016. Accessed on 21 January 2025.

[3] Md Manjurul Ahsan, Shahana Akter Luna, and Zahed Siddique. Machine-learning-based disease diagnosis: A comprehensive review. *Healthcare*, 10(3), 2022.

[4] Kareem Amin, Alex Kulesza, and Sergei Vassilvitskii. Practical considerations for differential privacy, 2024.

[5] Apple Differential Privacy Team. Learning with privacy at scale. 2017.

[6] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022*, pages 1138–1156. IEEE, 2022.

[7] Raef Bassily and Yoav Freund. Typicality-based stability and privacy. *CoRR*, abs/1604.03336, 2016.

[8] Mark Bun, Marco Gaboardi, Max Hopkins, Russell Impagliazzo, Rex Lei, Toniann Pitassi, Satchit Sivakumar, and Jessica Sorrell. Stability is stable: Connections between replicability, privacy, and adaptive generalization. In Barna Saha and Rocco A. Servedio, editors, *Proceedings of the 55th Annual ACM Symposium on Theory of Computing, STOC 2023, Orlando, FL, USA, June 20-23, 2023*, pages 520–527. ACM, 2023.

[9] The U.S Census Bureau. Appendix b — 2010 reconstruction-abetted re-identification simulated attack. `https://www2.census.gov/about/policies/foia/records/disclosure-avoidance/appendix-b-summary-of-simulated-reconstruction-abetted-re-identification-attack.pdf`.

[10] The U.S Census Bureau. The census bureau's simulated reconstruction-abetted re-identification attack on the 2010 census. `https://www.census.gov/data/academy/webinars/2021/disclosure-avoidance-series/simulated-reconstruction-abetted-re-identification-attack-on-the-2010-census.html`, May 2021.

[11] Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. Is private learning possible with instance encoding?, 2021.

[12] Aloni Cohen. Attacks on deidentification's defenses. pages 1469–1486. USENIX Association, 2022.

[13] Aloni Cohen. Attacks on deidentification's defenses. In Kevin R. B. Butler and Kurt Thomas, editors, *31st USENIX Security Symposium, USENIX Security 2022, Boston, MA, USA, August 10-12, 2022*, pages 1469–1486. USENIX Association, 2022.

[14] Aloni Cohen and Kobbi Nissim. Linear program reconstruction in practice. *J. Priv. Confidentiality*, 10(1), 2020.

[15] Aloni Cohen and Kobbi Nissim. Towards formalizing the gdpr's notion of singling out. *Proc. Natl. Acad. Sci. USA*, 117(15):8344–8352, 2020.

[16] Edith Cohen, Haim Kaplan, Yishay Mansour, Shay Moran, Kobbi Nissim, Uri Stemmer, and Eliad Tsfadia. Data reconstruction: When you see it and when you don't. *arXiv preprint arXiv:2405.15753*, 2024.

[17] Rachel Cummings, Shlomi Hod, Jayshree Sarathy, and Marika Swanberg. ATTAXONOMY: unpacking differential privacy guarantees against practical adversaries. *CoRR*, abs/2405.01716, 2024.

[18] Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 772–814. JMLR.org, 2016.

[19] Donna Daily. Disclosure avoidance protections for the american community survey, Dec 2022.

[20] Travis Dick, Cynthia Dwork, Michael Kearns, Terrance Liu, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu. Confidence-ranked reconstruction of census microdata from published statistics. *CoRR*, abs/2211.03128, 2022.

[21] Travis Dick, Cynthia Dwork, Michael Kearns, Terrance Liu, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu. Confidence-ranked reconstruction of census microdata from published statistics. *Proceedings of the National Academy of Sciences*, 120(8):e2218605120, 2023.

[22] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

[23] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In Frank Neven, Catriel Beeri, and Tova Milo, editors, *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*, pages 202–210. ACM, 2003.

[24] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In Corinna Cortes, Neil D.

Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2350–2358, 2015.

[25] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. *J. Priv. Confidentiality*, 7(3):17–51, 2016.

[26] Cynthia Dwork, Adam Smith, Thomas Steinke, and Jonathan Ullman. Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application (2017)*, 2017.

[27] Cynthia Dwork, Adam D. Smith, Thomas Steinke, Jonathan R. Ullman, and Salil P. Vadhan. Robust traceability from trace amounts. In Venkatesan Guruswami, editor, *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 650–669. IEEE Computer Society, 2015.

[28] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In Gail-Joon Ahn, Moti Yung, and Ninghui Li, editors, *ACM CCS 2014*, pages 1054–1067. ACM Press, November 2014.

[29] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data. *arXiv preprint arXiv:2211.10459*, 2022.

[30] Iftach Haitner, Daniel Nukrai, and Eylon Yogev. Lower bound on SNARGs in the random oracle model. In *CRYPTO 2022, Part III*, LNCS, pages 97–127, August 2022.

[31] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018.

[32] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1944–1953. PMLR, 2018.

[33] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLOS Genetics*, 4(8):1–9, 08 2008.

[34] Don Hush and Clint Scovel. Concentration of the hypergeometric distribution. *Statistics & Probability Letters*, 75:127–132, 11 2005.

[35] Russell Impagliazzo, Rex Lei, Toniann Pitassi, and Jessica Sorrell. Reproducibility in learning. In Stefano Leonardi and Anupam Gupta, editors, *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pages 818–831. ACM, 2022.

[36] Indeed Employer Content Team. Machine Learning in Recruitment: Revolutionize Your Hiring Process. https://www.indeed.com/hire/c/info/machine-learning-recruitment. Accessed on 21 January 2025.

[37] Matthew Jagielski, Jonathan R. Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[38] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In Nadia Heninger and Patrick Traynor, editors, *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*, pages 1895–1912. USENIX Association, 2019.

[39] Shiva P. Kasiviswanathan and Adam Smith. On the 'semantics' of differential privacy: A bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1), Jun. 2014.

[40] Sallie Ann Keller and John M. Abowd. Database reconstruction does compromise confidentiality. *Proceedings of the National Academy of Sciences*, 120(12):e2300976120, 2023.

[41] Os Keyes and Abraham D. Flaxman. How census data put trans children at risk. https://www.scientificamerican.com/article/how-census-data-put-trans-children-at-risk/, Sep 2022.

[42] Michael Khavkin and Eran Toch. Differential privacy configurations in the real world: A comparative analysis. *IEEE Transactions on Knowledge and Data Engineering*, 37(11):6334–6351, 2025.

[43] Michael P. Kim, Amirata Ghorbani, and James Y. Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor, editors, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 247–254. ACM, 2019.

[44] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (SP 2008), 18-21 May 2008, Oakland, California, USA*, pages 111–125. IEEE Computer Society, 2008.

[45] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy*, pages 111–125. IEEE Computer Society Press, May 2008.

[46] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In Joseph A. Calandrino and Carmela Troncoso, editors, *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 1631–1648. USENIX Association, 2023.

[47] Neelie Verlinden. Top 40+ Pre-Employment Assessment Tools. https://www.aihr.com/blog/top-pre-employment-assessment-tools/., July 2020. Accessed on 21 January 2025.

[48] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830, November 2011.

[49] Maria Rigaki and Sebastián García. A survey of privacy attacks in machine learning. *ACM Comput. Surv.*, 56(4):101:1–101:34, 2024.

[50] Ryan M. Rogers, Aaron Roth, Adam D. Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In Irit Dinur, editor, *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pages 487–494. IEEE Computer Society, 2016.

[51] Steven Ruggles and David Van Riper. The role of chance in the census bureau database reconstruction experiment. *Population Research and Policy Review*, 41(3):781–788, 2022.

[52] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella Béguelin. Sok: Let the privacy games begin! A unified treatment of data inference privacy in machine learning. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pages 327–345. IEEE, 2023.

[53] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society, 2017.

[54] Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[55] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.

[56] Christine Task, Karan Bhagat, Streat Damon, and Gary Howarth. Nist excerpts benchmark data. 2025.

[57] Katie Tezapsidis. Uber releases open source project for differential privacy, 2017.

[58] I. O. Tolstikhin. Concentration inequalities for samples without replacement. *Theory of Probability & Its Applications*, 61(3):462–481, 2017.

[59] V.N. Vapnik and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 1971.

[60] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018*, pages 268–282. IEEE Computer Society, 2018.

A Comparison to Perfect Generalization

In this section we include a comparison between *demographic coherence* (Definition 3) and the notion of *perfect generalization* introduced by Cummings et al. [18]. (See also the work of Bassily and Freund [7] that independently introduced a generalization thereof.) This notion was originally meant to capture generalization under post-processing but has since been shown to be closely related to other desirable properties as well (*e.g.,* replicability [8]). Our framework shares conceptual similarities with this definition, but the technical details differ in important ways.

The following comparison uses the definition of *sample perfect generalization* (Definition 10) from [8] which is roughly equivalent to the original definition from [18]. Intuitively, a mechanism M running on i.i.d. samples from some distribution (sample) perfectly generalizes if the distribution of its output does not depend "too much" on specific realization of its sampled input. That is, its output distributions when run on two i.i.d samples from any distribution are indistinguishable.

Definition 10 (Sample perfect generalization [8, Def 3.4]). *An algorithm $\mathcal{A} : \mathcal{X}^m \to \mathcal{Y}$ is said to be $(\beta, \epsilon, \delta)$-sample perfectly generalizing if, for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability at least $1 - \beta$ over the draw of two i.i.d. samples $X_a, X_b \sim \mathcal{D}^m$,*

$$\mathcal{A}(X_a) \approx_{\epsilon,\delta} \mathcal{A}(X_b),$$

where $\approx_{\epsilon,\delta}$ denotes ϵ, δ indistinguishability.

Definition 3 has several noticeable syntactic differences when compared to Definition 10. First, demographic coherence is defined within a specific framework that explicitly lays out the entire data release pipeline, a design choice that intentionally lends itself to concrete intuition (and experimental evaluation) of data release. However, this still leaves open the possibility that the core statistical guarantee of demographic coherence is roughly equivalent to perfect generalization. In other words, it may still be the case that demographic coherence is simply a different way to describe the protections offered by perfect generalization; as we see below, this is not the case.

Second, while "closeness" in the definition of perfect generalization is required for distributions over the entire sets X_a, X_b, the "closeness" in the definition of demographic coherence is required for distributions over the sets $X_a|_C, X_b|_C$ for subpopulations $C \subseteq \mathcal{X}$ from some collection $\mathcal{C}$. For the sake of drawing a more direct comparison here, we collapse this difference by comparing sample perfect generalization to demographic coherence with $\mathcal{C} = \{\mathcal{X}\}$.

A third difference in the definitions is is the choice of sets X_a, X_b that the comparison is made with respect to, *i.e.,* a random partition of a fixed dataset in demographic coherence vs. i.i.d. draws from a distribution in the case of perfect generalization. The choice of random partitioning in our framework is made to ensure concreteness and applicability in census-like settings but it is chosen intentionally to maintain both intuitive and quantitative similarities to i.i.d. sampling. Thus, we view this as more of a difference in interpretability and applicability of the definitions, rather than one about their underlying guarantees.

The main difference between the two definitions, thus, is in how how "closeness" is measured—as spelled out in Figure 4.

Perfect generalization asks that w.h.p. over independent samples, $\mathcal{A}$ produces indistinguishable distributions over reports $R_a \leftarrow \mathcal{A}(X_a)$ and $R_b \leftarrow \mathcal{A}(X_b)$. Meanwhile, coherence enforcement asks that w.h.p. $\mathcal{L} \circ \mathcal{A}(X_a)$ produces a confidence rated predictor $h_a : \mathcal{X} \to [-1, 1]$ which has "similar" predictions on X_a and X_b (a property enforced by $\mathcal{A}$). That is, the comparison in perfect generalization is on the behavior of the algorithm $\mathcal{A}$ itself, while the comparison in demographic coherence is on the likely behavior of a realized hypothesis h_a that is produced only over the report $R_a \leftarrow \mathcal{A}(X_a)$.

An algorithm $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$ is:

1. $(\beta, \varepsilon, \delta)$-*sample perfectly generalizing if*
   $\forall$ distributions $\mathcal{D}$ over $\mathcal{X}$, with probability at least $1 - \beta$ over $X_a, X_b \sim \mathcal{D}^{n/2}$:
   $$\mathcal{A}(X_a) \approx_{\varepsilon,\delta} \mathcal{A}(X_b)$$
2. (α, β)-*coherence enforcing if*
   $\forall$ datasets $X \in \mathcal{X}^n$, learners $\mathcal{L} : \mathcal{Y} \to (\mathcal{X} \to [-1, 1])$, with probability at least $1 - \beta$ over the random split $X_a \cup X_b = X$ and the coins of $\mathcal{A}$, $\mathcal{L}$:
   $$\mathsf{dist}_W(h_a(X_a), h_a(X_b)) \leq \alpha$$
   where $h_a \leftarrow \mathcal{L} \circ \mathcal{A}(X_a)$ is a confidence rated predictor, and $h_a(X_i)$ is the distribution induced by randomly choosing $x \sim X_i$ and computing $h_a(x)$.

Figure 4: Comparing the definition of sample perfect generalization to a simplified definition of demographic coherence.

Since coherence enforcement limits the set of algorithms against which indistinguishability applies, one might expect that perfectly generalizing algorithms also enforce demographic coherence, and indeed, Theorem 3 proves this to be true. However, the converse need not be true—implying that demographic coherence is a relaxation of perfect generalization. In particular, the example below shows a set X such that no confidence rated predictor $h : \mathcal{X} \to [-1, 1]$ violates this property. In this case, all data curators are vacuously coherence enforcing.

Consider a data curator $\mathcal{A} : \{0, 1\}^{n/2} \to \{0, 1\}^{n/2}$ that (deterministically) publishes its input in the clear. This is clearly not perfectly generalizing as the distribution of the report X_a is a point mass that (for reasonable choices of X, with high probability) is distinct from the distribution of the report X_b.

Meanwhile, considering a dataset $X \in \{0, 1\}^n$ and a random split $X_a, X_b \leftarrow X$ of the dataset, there are two possible predictors $h : \{0, 1\} \to [-1, 1]$ that witnesses the highest possible Wasserstein distance when run on X_a vs. X_b. That is, without loss of generality h is either $h(0) = -1, h(1) = 1$ or $h(0) = 1, h(1) = -1$. In either case, h cannot be improved even by seeing X_a in the clear. So, any data curation algorithm in this scenario, is coherence enforcing since the data itself doesn't have enough complexity to allow for a violation of demographic coherence. Note that the absence of information correlated with the bits contained in the dataset X (*e.g.,* time, location, computer system) is crucial to this example.

B Differential Privacy implies Bounded Max-Information: Sampling without Replacement

Prior work shows that for datasets sampled i.i.d., differentially private algorithms have bounded max-information [24, 50]. In this appendix we prove Theorem 14 and Theorem 15, which are analogs of those theorems for sampling without replacement.

B.1 Preliminaries

First we state Theorem 12, a version of McDiarmid's inequality that applies to the case of sampling without replacement. This result follows from Lemma 2 in [58]. This result in used in the proof of Theorem 14, which says that pure-DP algorithms have bounded max-information (even in the case of sampling without replacement.)

The following definition of order-invariant algorithms appears as a technical assumption in some of our theorem statements.[18] This is a minimal assumption because any non-order-invariant algorithm can be made order-invariant by simply pre-processing the dataset with a sorting or shuffling operation.

Definition 11. *A function $f : \mathcal{X}^m \to \mathcal{Y}$, is called* order invariant *if for all $X \in \mathcal{X}^m$, the value of the function $f(X)$ does not depend on the order of the elements of X.*

Theorem 12 (McDiarmid's for sampling without replacement [58])**.** *Let $f : \mathcal{X}^n \to \mathcal{Y}$ be an an order invariant function with global sensitivity $\Delta > 0$. Let $\mathcal{X}$ be a data universe of size n, let S be a sample of size m chosen without replacement from $\mathcal{X}$. Then for $t \geq 0$,*

$$\Pr_S[f(S) - \mathbb{E}[f(S)] \geq t] \leq \exp\left(-\frac{2t^2}{m\Delta^2} \cdot \left(\frac{n-1/2}{n-m}\right) \cdot \left(1 - \frac{1}{2\max(m, n-m)}\right)\right)$$

In particular, for $m = n/2$ and $n \geq 3$,

$$\Pr[f(S) - \mathbb{E}[f(S)] \geq t] \leq \exp\left(-\frac{4t^2}{n\Delta^2}\right)$$

Next we state some lemmas that are used in the proof of Theorem 15 and Corollary 2 which say that approximate-DP algorithms have bounded max-information (even in the case of sampling without replacement.)

Definition 12 (Point-wise indistinguishibility [39])**.** *Two random variables A, B are point-wise (ε, δ)-indistinguishable if with probability at least $1 - \delta$ over $a \sim P(A)$:*

$$e^{-\epsilon} \Pr[B = a] \leq \Pr[A = a] \leq e^{\epsilon} \Pr[B = a].$$

Lemma 2 (Indistinguishability implies Pointwise Indistinguishability [39])**.** *Let A, B be two random variables. If $A \approx_{\varepsilon,\delta} B$ then A and B are pointwise $\left(2\varepsilon, \frac{2\delta}{1-e^{-\varepsilon}}\right)$-indistinguishable.*

Lemma 3 (Conditioning Lemma [39])**.** *Suppose that $(A, B) \approx_{\varepsilon,\delta} (A', B')$. Then for every $\hat{\delta} > 0$, the following holds:*

$$\Pr_{t \sim P(B)}\left[A|_{B=t} \approx_{3\epsilon,\hat{\delta}} A'|_{B'=t}\right] \geq 1 - \frac{2\delta}{\hat{\delta}} - \frac{2\delta}{1-e^{-\varepsilon}}.$$

Theorem 13 (Azuma's Inequality)**.** *Let $C_1, \cdots, C_n$ be a sequence of random variables such that for every $i \in [n]$, we have*

$$\Pr[|C_i| \leq \alpha] = 1$$

and for every fixed prefix $\mathbf{C}_1^{i-1} = \mathbf{c}_1^{i-1}$, we have

$$\mathbb{E}\left[C_i | \mathbf{c}_1^{i-1}\right] \leq \gamma,$$

[18] Since it is an assumption in the version of McDiarmid's Inequality for sampling without replacement (Theorem 12) that we use.

then for all $t \geq 0$, we have

$$\Pr\left[\sum_{i=1}^{n} C_i > n\gamma + t\sqrt{n}\alpha\right] \leq e^{-t^2/2}.$$

B.2 Pure-DP $\implies$ Bounded Max-Information

In this appendix we state Theorem 14, which is an analog of Theorem from [24]. The proof of this theorem works exactly as in [24], except replacing the application of McDiarmid's Lemma with a version of McDiarmid's for sampling without replacement (Theorem 12) which we state in Appendix B.1.

Theorem 14. (Pure-DP $\implies$ Bounded Max-Information) *Fix $n \in \mathbb{N}$, $\varepsilon > 0$ and let $\mathcal{X}$ be a data universe of size at least n. Let $\mathcal{A} : \mathcal{X}^{n/2} \to \mathcal{Y}$ be an order-invariant ε-DP algorithm. Then for any $\gamma > 0$,*

$$I_{\infty}^{\gamma}(\mathcal{A}, n/2) \leq \varepsilon^2 n/4 + \varepsilon\sqrt{n\ln(2/\gamma)/4}.$$

B.3 (ε, δ)-DP $\implies$ Bounded Max-Information

In this appendix we prove Theorem 15, which is an analog of Theorem 1 from Rogers et al.[50]. In fact, the following proof is an adaptation of the proof of them from [50] with the following differences: (1) we compute all the constants exactly and avoid using asymptotic notation, and (2) we keep the tunable parameters in the final version of the theorem to obtain the most flexible result that we can. Finally, we set the parameters in Theorem 15 to get Corollary 2, which is used in the proof of Theorem 8 (3) We attempt to make the proof more readable by simplifying the notation. In particular, we do not define "good sets" $\mathcal{E}, \mathcal{F}$.

Theorem 15. (Approx-DP $\implies$ Bounded Max-Information, Generalised) *Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an (order-invariant) (ε, δ)-differentially private algorithm for $\varepsilon \in (0, 1/2]$, $\delta \in (0, \varepsilon)$. For $\hat{\delta} \in (0, \varepsilon/15]$, $t > 0$, and $\beta(t, \hat{\delta}) = e^{-t^2/2} + n\left(\frac{2\delta}{\hat{\delta}} + \frac{2\hat{\delta}+2\delta}{1-e^{-3\varepsilon}}\right)$ we have*

$$I_{\infty}^{\beta}(\mathcal{A}, n) \leq n\left(347\hat{\delta} + 75\left(\frac{\hat{\delta}}{\varepsilon}\right)^2 + 24\frac{\hat{\delta}^2}{\varepsilon} + 240\varepsilon^2\right) + 6t\varepsilon\sqrt{n}.$$

Corollary 2. (Approx-DP $\implies$ Bounded Max-Information, Specific) *Fix $n \in \mathbb{N}$, for $\varepsilon \in (0, 1/2]$, $\gamma \in (0, 1]$, $\delta \in (0, \frac{\varepsilon^2\gamma^2}{(120n)^2}]$ and let $\mathcal{X}$ be a data universe of size at least n. Let $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ be an (order-invariant) (ε, δ)-differentially private algorithm. We have that*

$$I_{\infty}^{\gamma}(\mathcal{A}, n) \leq 265\varepsilon^2 n + 12\varepsilon\sqrt{n\ln(2/\gamma)}.$$

We will sometimes abbreviate conditional probabilities of the form $\Pr[\mathbf{X} = \mathbf{x} \mid \mathcal{A} = a]$ as $\Pr[\mathbf{X} = \mathbf{x} \mid a]$ when the random variables are clear from context. Further, for any $\mathbf{x} \in \mathcal{X}^n$ and $a \in \mathcal{Y}$, we define

$$Z_i(a, \mathbf{x}_{[i]}) \stackrel{\text{def}}{=} \log\frac{\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]}. \tag{18}$$

$$Z(a, \mathbf{x}) \stackrel{\text{def}}{=} \log\left(\frac{\Pr_{\mathbf{x}}[\mathcal{A}(x) = a, \mathbf{X} = \mathbf{x}]}{\Pr[\mathcal{A} = a] \cdot \Pr[\mathbf{X} = \mathbf{x}]}\right)$$

$$= \sum_{i=1}^{n} Z_i(a, \mathbf{x}_{[i]}) \tag{19}$$

If we can bound $Z(a, \mathbf{x})$ with high probability over $(a, \mathbf{x}) \sim p(\mathcal{A}(\mathbf{X}), \mathbf{X})$, then we can bound the approximate max-information by using the following lemma:

Lemma 4 ([24, Lemma 18])**.**

$$\Pr\left[\log\left(\frac{\Pr_{\mathbf{x}}\left[\mathcal{A}(x) = a, \mathbf{X} = \mathbf{x}\right]}{\Pr\left[\mathcal{A} = a\right] \cdot \Pr\left[\mathbf{X} = \mathbf{x}\right]}\right) \geq k\right] \leq \beta \implies I_{\infty}^{\beta}(\mathcal{A}(\mathbf{X}); \mathbf{X}) \leq k.$$

To bound $Z(a, \mathbf{x})$ with high probability over $(a, \mathbf{x}) \sim p(\mathcal{A}(\mathbf{X}), \mathbf{X})$ we will apply Azuma's inequality (Theorem 13) to the sum of the $Z_i(a, \mathbf{x}_{[i]})$'s. For this we must first argue that each $Z_i(a, \mathbf{x}_{[i]})$ term is bounded with high probability:

Claim 3. *Let $\hat{\delta} > 0$, and $\delta'' \stackrel{def}{=} \frac{2\hat{\delta}}{1-e^{-3\varepsilon}}$. If $\mathcal{A}$ is (ε, δ)-differentially private and, $\mathbf{X} \in \mathcal{X}^n$ is sampled without replacement from a finite universe $\mathcal{X}$, then for each $i \in [n]$, and each prefix $\mathbf{x}_{[i-1]} \in \mathcal{X}^{i-1}$ and answer a, we have:*

$$\Pr_{x_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}})}\left[\log \frac{\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]} \leq 6\varepsilon\right] \geq 1 - \delta''$$

Proof. Whenever $X_i|_{\mathbf{x}_{[i-1]}}$ and $X_i|_{a,\mathbf{x}_{[i-1]}}$ are $\left(3\epsilon, \hat{\delta}\right)$-indistinguishable, Lemma 2 tells us that $X_i|_{\mathbf{x}_{[i-1]}}$ and $X_i|_{a,\mathbf{x}_{[i-1]}}$ are point-wise $(6\epsilon, \delta'')$-indistinguishable. *i.e.*, given that $X_i|_{\mathbf{x}_{[i-1]}}$ and $X_i|_{a,\mathbf{x}_{[i-1]}}$ are $\left(3\epsilon, \hat{\delta}\right)$-indistinguishable, we have that

$$\Pr_{x_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}})}\left[\log \frac{\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]} \leq 6\varepsilon\right] \geq 1 - \delta''$$

■

Claim 4. *Let $\hat{\delta} > 0$, $\delta' \stackrel{def}{=} \frac{2\delta}{\hat{\delta}} + \frac{2\delta}{1-e^{-\varepsilon}}$, and $\delta'' \stackrel{def}{=} \frac{2\hat{\delta}}{1-e^{-3\varepsilon}}$. If $\mathcal{A}$ is (ε, δ)-differentially private and, $\mathbf{X} \in \mathcal{X}^n$ is sampled without replacement from a finite universe $\mathcal{X}$, then for each $i \in [n]$, and each prefix $\mathbf{x}_{[i-1]} \in \mathcal{X}^{i-1}$ we have:*

$$\Pr_{\substack{x_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}}) \\ a \sim \mathcal{A}|_{\mathbf{x}_{[i-1]}}}}\left[\log \frac{\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]} \leq 6\varepsilon\right] \geq 1 - \delta' - \delta''$$

Proof. For this proof, we use Claim 3 and then show for each $i \in [n]$, and prefix $\mathbf{x}_{[i-1]} \in \mathcal{X}^{i-1}$,

$$\Pr_{a \sim p\left(\mathcal{A}|_{\mathbf{x}_{[i-1]}}\right)}\left[X_i|_{\mathbf{x}_{[i-1]}} \approx_{3\varepsilon, \hat{\delta}} X_i|_{a,\mathbf{x}_{[i-1]}}\right] \geq 1 - \delta'.$$

We use the differential privacy guarantee on $\mathcal{A}$ to show that $(\mathcal{A}, X_i)|_{\mathbf{x}_{[i-1]}} \approx_{\varepsilon,\delta} \mathcal{A}|_{\mathbf{x}_{[i-1]}} \otimes X_i|_{\mathbf{x}_{[i-1]}}$. The above equation then follows directly from the conditioning lemma (Lemma 3).

To show that $(\mathcal{A}, X_i)|_{\mathbf{x}_{[i-1]}} \approx_{\varepsilon,\delta} \mathcal{A}|_{\mathbf{x}_{[i-1]}} \otimes X_i|_{\mathbf{x}_{[i-1]}}$, we will show the following:

$$\Pr\left[(\mathcal{A}(\mathbf{X}), X_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] \leq e^{\varepsilon} \cdot \Pr\left[\mathcal{A}(\mathbf{X}) \otimes X_i \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] + \delta.$$

($\Pr\left[\mathcal{A}(\mathbf{X}) \otimes X_i \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] \leq e^{\varepsilon} \cdot \Pr\left[(\mathcal{A}(\mathbf{X}), X_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] + \delta$ follows from a very similar argument.)

Fix any set $\mathcal{O} \subseteq \mathcal{Y} \times \mathcal{X}$ and prefix $\mathbf{x}_{[i-1]} \in \mathcal{X}^{i-1}$. From the differential privacy of $\mathcal{A}$, and the order-invariance of the algorithm, we get the following:

$$\begin{aligned}
&\Pr\left[(\mathcal{A}(\mathbf{X}), X_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] \\
&= \sum_{x_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}})} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}, x_i\right] \\
&\leq \sum_{x_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}})} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \left(e^{\varepsilon} \Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}, t_i\right] + \delta\right) \qquad \forall t_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}}) \\
&= \sum_{x_i, t_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}})} \Pr\left[X_i = t_i \mid \mathbf{x}_{[i-1]}\right] \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \left(e^{\varepsilon} \Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}, t_i\right] + \delta\right) \\
&= \sum_{x_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}})} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \left(e^{\varepsilon} \Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] + \delta\right) \\
&\leq e^{\varepsilon} \left(\sum_{x_i \in \mathsf{Supp}(X_i|_{\mathbf{x}_{[i-1]}})} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \Pr\left[\mathcal{A}(\mathbf{X}), X_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] \right) + \delta \\
&= e^{\varepsilon} \Pr\left[\mathcal{A}(\mathbf{X}) \otimes X_i \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] + \delta
\end{aligned}$$

Where, the first inequality is true by the following coupling argument where we separately consider the cases where $t_i \notin \mathbf{x}_{[i-1]}$ and those where $t_i \in \mathbf{x}_{[i-1]}$. First, whenever $t_i \notin \mathbf{x}_{[i-1]}$, The following is true by differential privacy:

$$\Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}, x_i\right] \leq e^{\varepsilon} \Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}, t_i\right] + \delta$$

Otherwise, if $t_i \in \mathbf{x}_{[i-1]}$, then $\mathbf{x}_{[i-1]} = \mathbf{x}_{[j-1]}, t_i, \mathbf{x}_{[j+1\,:\,i-1]}$. In these cases, couple $\mathbf{x}_{[i-1]}, x_i$ with $\mathbf{x}_{[j-1]}, x_i, \mathbf{x}_{[j+1\,:\,i-1]}, t_i$. Notice that the following is true because of the order invariance of the algorithm:[19]

$$\Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[j-1]}, t_i, \mathbf{x}_{[j+1\,:\,i-1]}, x_i\right] = \Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[j-1]}, x_i, \mathbf{x}_{[j+1\,:\,i-1]}, t_i\right]$$

■

Having shown a high probability bound on the terms Z_i, our next step is to bound their expectation so that we can continue towards our goal of applying Azuma's inequality.

[19] If the algorithm were not order invariant we could instead double the parameters to get $\Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}, x_i\right] \leq e^{2\varepsilon} \Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}, t_i\right] + 2\delta$ by invoking differential privacy twice in the latter case.

We will use the following shorthand notation for conditional expectation:

$$\begin{aligned}&\mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_{[i]}) \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] \\ &\stackrel{def}{=} \mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_{[i]}) \mid \mathcal{A} = a, \mathbf{X}_{[i-1]} = \mathbf{x}_{[i-1]}, |Z_i(\mathcal{A}, \mathbf{X}_{[i]})| \leq 6\varepsilon\right],\end{aligned}$$

Lemma 5. *Let $\mathcal{A}$ be (ε, δ)-differentially private and, $\mathbf{X} \in \mathcal{X}^n$ be sampled without replacement from a finite universe $\mathcal{X}$. Let $\varepsilon \in (0, 1/2]$ and $\hat{\delta} \in (0, \varepsilon/15]$,*

$$X_i|_{\mathbf{x}_{[i-1]}} \approx_{3\varepsilon,\hat{\delta}} X_i|_{a,\mathbf{x}_{[i-1]}} \quad \Longrightarrow \quad \mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_{[i]}) \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] = O(\varepsilon^2 + \hat{\delta}).$$

More precisely, $\mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_{[i]}) \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] \leq \nu(\hat{\delta})$, where $\nu(\hat{\delta})$ is defined in (20).

Proof. Let $S \stackrel{\text{def}}{=} \{x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| < 6\varepsilon\}$. Given an outcome and prefix $(a, \mathbf{x}_{[i-1]})$ such that $X_i|_{\mathbf{x}_{[i-1]}} \approx_{3\varepsilon,\hat{\delta}} X_i|_{a,\mathbf{x}_{[i-1]}}$, we have the following by definition:

$$\begin{aligned}&\mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_{[i]}) \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] \\ &= \sum_{x_i \in S} \Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] Z_i(a, x_{[i]}) \\ &= \sum_{x_i \in S} \Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right)\end{aligned}$$

Claim 5.

$$\sum_{x_i \in S} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right) \leq \log\left(\frac{1 - \Pr\left[X_i \notin S | a, \mathbf{x}_{[i-1]}\right]}{1 - \Pr\left[X_i \notin S | \mathbf{x}_{[i-1]}\right]}\right)$$

Proof.

$$\begin{aligned}&\sum_{x_i \in S} \Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}\right] \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right) \\ &= \Pr\left[X_i \in S \mid \mathbf{x}_{[i-1]}\right] \sum_{x_i \in S} \frac{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i \in S | \mathbf{x}_{[i-1]}\right]} \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right) \\ &\leq \sum_{x_i \in S} \frac{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i \in S | \mathbf{x}_{[i-1]}\right]} \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right) \\ &\leq \log\left(\sum_{x_i \in S} \frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i \in S | \mathbf{x}_{[i-1]}\right]}\right) \\ &\leq \log\left(\frac{\Pr\left[X_i \in S | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i \in S | \mathbf{x}_{[i-1]}\right]}\right) = \log\left(\frac{1 - \Pr\left[X_i \notin S | a, \mathbf{x}_{[i-1]}\right]}{1 - \Pr\left[X_i \notin S | \mathbf{x}_{[i-1]}\right]}\right)\end{aligned}$$

The first inequality follows form the fact that all probabilities are less than one. The second inequality follows from noticing that $\sum_{x_i \in S} \frac{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i \in S | \mathbf{x}_{[i-1]}\right]} = 1$ and applying Jensen's inequality. ■

Let $\Pr\left[X_i \notin \{x_i \mid a, \mathbf{x}_{[i-1], |Z_i|<6\varepsilon}\} \mid a, \mathbf{x}_{[i-1]}\right] = \Pr\left[X_i \notin S \mid a, \mathbf{x}_{[i-1]}\right] \stackrel{\text{def}}{=} q$. Note that, because $X_i|_{\mathbf{x}_{[i-1]}} \approx_{3\varepsilon,\hat{\delta}} X_i|_{a,\mathbf{x}_{[i-1]}}$, we have for $\hat{\delta} > 0$:

$$\Pr\left[X_i \notin S \mid \mathbf{x}_{[i-1]}\right] \leq e^{3\varepsilon}\Pr\left[X_i \notin S \mid a, \mathbf{x}_{[i-1]}\right] + \hat{\delta} = e^{3\varepsilon}q + \hat{\delta}$$

Note that $q \leq \delta''$ by Claim 3. Now, we can bound the following:

$$\begin{aligned}
&\sum_{x_i \in S} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right) \\
&\quad \leq \log\left(\frac{1-\Pr\left[X_i \notin S | a, \mathbf{x}_{[i-1]}\right]}{1-\Pr\left[X_i \notin S | \mathbf{x}_{[i-1]}\right]}\right) \\
&\quad \leq \log(1-q) - \log(1-(e^{3\varepsilon}q + \hat{\delta})) \\
&\quad \leq \log(e) \cdot (-q + e^{3\varepsilon}q + \hat{\delta} + 2(e^{3\varepsilon}q + \hat{\delta})^2) \\
&\quad = \log(e) \cdot ((e^{3\varepsilon} - 1)q + \hat{\delta} + 2(e^{3\varepsilon}q + \hat{\delta})^2) \\
&\quad \stackrel{\text{def}}{=} \tau(\hat{\delta})
\end{aligned}$$

where the second inequality follows by using the inequality $(-x - 2x^2)\log(e) \leq \log(1-x) \leq -x\log(e)$ for $0 < x \leq 1/2$, and as $(e^{3\varepsilon}q + \hat{\delta}) \leq 1/2$ for ε and $\hat{\delta}$ bounded as in the lemma statement.

We use the results above to to upper bound the expectation we wanted:

$$\begin{aligned}
&\mathbb{E}\left[Z_i(\mathcal{A}, \mathbf{X}_{[i]}) \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] \\
&\leq \sum_{x_i \in S} \Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right) \\
&\quad - \sum_{x_i \in S} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right) + \tau(\hat{\delta}) \\
&= \sum_{x_i \in S} \left(\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] - \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]\right) \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right) + \tau(\hat{\delta}) \\
&\leq_{|Z_i| \leq 6\varepsilon} 6\varepsilon \sum_{x_i \in S} \left|\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] - \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]\right| + \tau(\hat{\delta}) \\
&\leq_{\text{Def of } S \text{, Claim 3}} 6\varepsilon \sum_{x_i \in S} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \max\left\{\frac{e^{6\varepsilon}}{\Pr\left[|Z_i| \leq 6\varepsilon \mid a, \mathbf{x}_{[i-1]}\right]} - 1, 1 - \frac{e^{-6\varepsilon}}{\Pr\left[|Z_i| \leq 6\varepsilon \mid a, \mathbf{x}_{[i-1]}\right]}\right\} + \tau(\hat{\delta}) \\
&\leq_{\text{Claim 3}} 6\varepsilon\left(\frac{e^{6\varepsilon}}{1 - \frac{2\hat{\delta}}{1-e^{-3\varepsilon}}} - 1\right) + \tau(\hat{\delta}) \\
&\leq_{\text{Substituting for } \tau(\hat{\delta})} 6\varepsilon\left(e^{6\epsilon}\left(1 + \frac{4\hat{\delta}}{1-e^{-3\epsilon}}\right) - 1\right) + \log(e) \cdot ((e^{3\varepsilon} - 1)q + \hat{\delta} + 2(e^{3\varepsilon}q + \hat{\delta})^2) \\
&\leq_{\text{Upper bound for } q} 6\varepsilon\left(e^{6\epsilon}\left(1 + \frac{4\hat{\delta}}{1-e^{-3\epsilon}}\right) - 1\right) \\
&\quad + \log(e) \cdot \left((e^{3\varepsilon} - 1)\frac{2\hat{\delta}}{1 - e^{-3\epsilon}} + \hat{\delta} + 2\hat{\delta}^2 + 8\frac{\hat{\delta}^2 e^{6\varepsilon}}{(1 - e^{-3\varepsilon})^2} + 8\frac{\hat{\delta}^2 e^{3\varepsilon}}{1 - e^{-3\varepsilon}}\right) \\
&=_{b = \frac{\hat{\delta}}{1-e^{-3\varepsilon}}} 6\varepsilon\left(e^{6\epsilon}(1 + 4b) - 1\right) + \log(e) \cdot \left(b\left(2e^{3\varepsilon} - 2 + 8\frac{\hat{\delta}e^{6\varepsilon}}{(1 - e^{-3\varepsilon})} + 8\hat{\delta}e^{3\varepsilon}\right) + \hat{\delta} + 2\hat{\delta}^2\right) \\
&= b\left(24\varepsilon e^{6\varepsilon} + 2e^{3\varepsilon} - 2 + 8\frac{\hat{\delta}e^{6\varepsilon}}{(1 - e^{-3\varepsilon})} + 8\hat{\delta}e^{3\varepsilon}\right) + \hat{\delta} + 2\hat{\delta}^2 + 6\varepsilon(e^{6\varepsilon} - 1) \\
&= \frac{\hat{\delta}}{1 - e^{-3\varepsilon}}\left(2e^{3\varepsilon}(4e^{3\varepsilon}(3\varepsilon + \frac{\hat{\delta}}{1 - e^{-3\varepsilon}})) + 4\hat{\delta} + 1) - 2\right) + \hat{\delta} + 2\hat{\delta}^2 + 6\varepsilon(e^{6\varepsilon} - 1)
\end{aligned}$$

$$
\begin{aligned}
&\leq_{e^{-3\varepsilon}\leq 1-1.5\varepsilon \text{ for } \varepsilon\in[0,0.5]} 2\frac{\hat{\delta}}{1.5\varepsilon}e^{3\varepsilon}\left(4e^{3\varepsilon}(3\varepsilon+\frac{\hat{\delta}}{1.5\varepsilon}))+4\hat{\delta}+1\right)+\hat{\delta}\left(\frac{-2}{1.5\varepsilon}+2\hat{\delta}+1\right)+6\varepsilon(e^{6\varepsilon}-1)\\
&\leq 8\frac{e^{6\varepsilon}\hat{\delta}}{1.5\varepsilon}\left(3\varepsilon+\frac{\hat{\delta}}{1.5\varepsilon}\right)+2\frac{\hat{\delta}}{1.5\varepsilon}e^{3\varepsilon}\left(4\hat{\delta}+1\right)+\hat{\delta}\left(\frac{-2}{1.5\varepsilon}+2\hat{\delta}+1\right)+6\varepsilon(e^{6\varepsilon}-1)\\
&\leq_{e^{3\varepsilon}\leq 1+7\varepsilon, e^{6\varepsilon}\leq 1+40\varepsilon \text{ for } \varepsilon\in[0,0.5]} 8\frac{(1+40\varepsilon)\hat{\delta}}{1.5\varepsilon}\left(3\varepsilon+\frac{\hat{\delta}}{1.5\varepsilon}\right)+2\frac{(1+7\varepsilon)\hat{\delta}}{1.5\varepsilon}\left(4\hat{\delta}+1\right)\\
&\quad+\hat{\delta}\left(\frac{-2}{1.5\varepsilon}+2\hat{\delta}+1\right)+6\varepsilon(40\varepsilon)\\
&\leq \frac{(8+320\varepsilon)\hat{\delta}}{1.5\varepsilon}\left(3\varepsilon+\frac{\hat{\delta}}{1.5\varepsilon}\right)+\frac{(2+14\varepsilon)}{1.5\varepsilon}\left(4\hat{\delta}^2+\hat{\delta}\right)-\frac{2\hat{\delta}}{1.5\varepsilon}+2\hat{\delta}^2+\hat{\delta}+240\varepsilon^2\\
&\leq_{\varepsilon<0.5} \frac{168\hat{\delta}}{1.5\varepsilon}(3\varepsilon+\frac{\hat{\delta}}{1.5\varepsilon})+\frac{8}{1.5}\frac{\hat{\delta}^2}{\varepsilon}+\frac{56}{1.5}\hat{\delta}^2+\frac{14}{1.5}\hat{\delta}+2\hat{\delta}^2+\hat{\delta}+240\varepsilon^2\\
&\leq_{\varepsilon\leq 0.5} 347\hat{\delta}+75\left(\frac{\hat{\delta}}{\varepsilon}\right)^2+24\frac{\hat{\delta}^2}{\varepsilon}+240\varepsilon^2\\
&\stackrel{\text{def}}{=} \nu(\hat{\delta}) \qquad (20)
\end{aligned}
$$

■

Finally, we need to apply Azuma's inequality (stated in Theorem 13) to a set of variables that are bounded with probability 1, not just with high probability. Towards this end, we now define (1) the sets $\mathcal{G}_i(\hat{\delta})$ and $\mathcal{G}_{\leq i}(\hat{\delta})$ of "good" tuples of outcomes and databases, and (2) a variable T_i that will match Z_i for "good events", and will be zero otherwise—and hence, is always bounded:

$$\mathcal{G}_i(\hat{\delta}) = \left\{(a, \mathbf{x}_{[i]}) \;\middle|\; |Z_i(a, \mathbf{x}_{[i]})| \leq 6\varepsilon \quad \& \quad X_i\,|_{\mathbf{x}_{[i-1]}} \approx_{3\varepsilon,\hat{\delta}} X_i|_{a,\mathbf{x}_{[i-1]}}\right\}, \tag{21}$$

$$\mathcal{G}_{\leq i}(\hat{\delta}) = \left\{(a, \mathbf{x}_{[i]}) : (a, x_1) \in \mathcal{G}_1(\hat{\delta}), \cdots, (a, \mathbf{x}_{[i]}) \in \mathcal{G}_i(\hat{\delta})\right\} \tag{22}$$

$$T_i(a, \mathbf{x}_{[i]}) = \begin{cases} Z_i(a, \mathbf{x}_{[i]}) & \text{if } (a, \mathbf{x}_{[i]}) \in \mathcal{G}_{\leq i}(\hat{\delta}) \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

Note that the variables T_i indeed satisfy the requirements of Azuma's inequality. The first condition, $\Pr\left[|T_i(\mathcal{A}, \mathbf{X}_{[i]})| \leq 6\varepsilon\right] = 1$ holds by definition, and the second holds because of Lemma 5.

We are now ready to prove our main theorem.

Proof of Theorem 15. For any constant ν, we have:

$$
\begin{aligned}
&\Pr\left[\sum_{i=1}^{n} Z_i(\mathcal{A}, \mathbf{X}_{[i]}) > n\nu + 6t\varepsilon\sqrt{n}\right]\\
&\leq \Pr\left[\sum_{i=1}^{n} Z_i(\mathcal{A}, \mathbf{X}_{[i]}) > n\nu + 6t\varepsilon\sqrt{n} \cap (\mathcal{A}, \mathbf{X}) \in \mathcal{G}_{\leq n}(\hat{\delta})\right] + \Pr\left[(\mathcal{A}, \mathbf{X}) \notin \mathcal{G}_{\leq n}(\hat{\delta})\right]\\
&= \Pr\left[\sum_{i=1}^{n} T_i(\mathcal{A}, \mathbf{X}_{[i]}) > n\nu + 6t\varepsilon\sqrt{n} \cap (\mathcal{A}, \mathbf{X}) \in \mathcal{G}_{\leq n}(\hat{\delta})\right] + \Pr\left[(\mathcal{A}, \mathbf{X}) \notin \mathcal{G}_{\leq n}(\hat{\delta})\right]
\end{aligned}
$$

We then substitute ν by $\nu(\hat{\delta})$ as defined in Equation (20), and apply a union bound on $\Pr\left[(\mathcal{A}, \mathbf{X}) \notin \mathcal{G}_{\leq n}(\hat{\delta})\right]$ using Claim 4 to get

$$\Pr\left[\sum_{i=1}^{n} Z_i(\mathcal{A}, \mathbf{X}_{[i]}) > n\nu(\hat{\delta}) + 6t\varepsilon\sqrt{n}\right] \leq \Pr\left[\sum_{i=1}^{n} T_i(\mathcal{A}, \mathbf{X}_{[i]}) > n\nu(\hat{\delta}) + 6t\varepsilon\sqrt{n}\right] + n(\delta' + \delta'')$$

$$\leq e^{-t^2/2} + n(\delta' + \delta'')$$

where the two inequalities follow from Claim 4 and Theorem 13, respectively. Therefore,

$$\Pr\left[Z(\mathcal{A}(\mathbf{X}), \mathbf{X}) > n\nu(\hat{\delta}) + 6t\varepsilon\sqrt{n}\right] \leq e^{-t^2/2} + n(\delta' + \delta'') \quad \stackrel{def}{=} \beta(t, \hat{\delta})$$

From Lemma 4, we have $I_\infty^{\beta(t,\hat{\delta})}(\mathbf{X}; \mathcal{A}(\mathbf{X})) \leq n\nu(\hat{\delta}) + 6t\varepsilon\sqrt{n}$.

■

We now prove the Corollary 2, which we use in Section 6.2 to prove Theorem 8.

Proof of Corollary 2. Setting $t = \sqrt{2\ln(2/\gamma)}$, and $\hat{\delta} = \frac{\sqrt{\varepsilon\delta}}{15}$, in Theorem 15, we get that $\beta(t, \hat{\delta}) \leq \gamma/2 + 30n\sqrt{\delta/\varepsilon} + n\frac{2\sqrt{\varepsilon\delta}+2\delta}{1.5\varepsilon}$, where we've used that $1 - e^{-3\varepsilon} \geq 1.5\varepsilon$ for $\varepsilon \in (0, 1/2]$. We note that for $\delta \leq \frac{\varepsilon^2\gamma^2}{(120n)^2}$, we have that $n\frac{2\sqrt{\varepsilon\delta}+2\delta}{1.5\varepsilon} \leq \frac{\gamma}{2}$, ensuring that $\beta(t, \hat{\delta}) \leq \gamma$. Also, the same bound on δ ensures that $n\left(347\hat{\delta} + 75\left(\frac{\hat{\delta}}{\varepsilon}\right)^2 + 24\frac{\hat{\delta}^2}{\varepsilon} + 240\varepsilon^2\right) \leq 265\varepsilon^2 n$. Substituting for t directly in the max-information bound then completes the proof. ■