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

Mark Bun, Marco Carmosino, Palak Jain, Gabriel Kaptchuk, Satchit Sivakumar

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

The collection of data and dissemination of aggregated statistics is a key function of government and civil society, driving critical data-driven decision making processes, *e.g.,* democratic apportionment, collective resource allocation, and documenting ongoing social ills. Indeed, data has become an indispensable modern tool for producing knowledge. However, the collection of personal data—particularly mass scale collection—introduces the potential for the inappropriate disclosure of information that individuals might prefer to remain private. Thus, data curators must carefully apply privacy protection mechanisms to their data, ideally without compromising the utility of the eventual data release. The study of privacy preserving data releases started with *attacks* that compellingly demonstrated that data releases which had not taken steps to ensure privacy could be weaponized for harm, *e.g.,* Latanya Sweeney's infamous re-identification of the Massachusetts Governor Bill Weld's medical records [52]. Since then, there has been a robust literature describing increasingly sophisticated attacks which continue to motivate efforts towards privacy preserving data release [40, 10, 9, 12]. While these attacks have proved convincing enough to shift data protection practices in many fields, attack demonstrations do not provide a clear path towards designing data protection mechanisms themselves, even in the form of "prevent all attacks like this one"; the attack demonstrations do not take on the task of distilling a set of agreed upon properties that make the attack convincing. In the aftermath of these attacks, the research community has developed a set of formal approaches that aim to provide robust privacy guarantees. While early attempts, like k-anonymity, proved inadequate, differential privacy [24] has recently emerged as an accepted standard, seeing deployment in both industry [27, 5, 21, 53] and government [1]. These formal approaches are often seen 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 can be significant barriers to applying differential privacy, which stem from the need to strike a delicate balance between the benefits of privacy preservation and its cost to utility [4]. Moreover, the generality of sufficient conditions means that they naturally lend themselves to being very abstract, which can make it far too easy to lose sight of the concrete privacy harms they are intended to prevent [17].

The deficiencies inherent in each of the existing approaches compels us to explore an intermediary design philosophy: *necessary* conditions. Within this approach, we can formally define (possibly many) properties that any private data release should 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 harm. Specifying necessary conditions promises to be an approach that simultaneously embraces the formality of sufficient conditions, while being just as concrete and convincing as attacks. Thinking in terms of necessary conditions has always been implicit in the practice of differential privacy (albeit, usually informally), where selecting the "best" privacy parameters ε, δ, for a deployment requires a trade-off with other metrics, such as accuracy. This makes it necessary to understand how small the parameters *must be* for the prevention of concrete privacy harms. Therefore, necessary conditions can provide a concrete methodology for justifying parameter choices by identifying parameter regimes that could enable specific

harms.

A new necessary condition: Demographic Coherence. In this work, we design a novel necessary privacy notion rooted in three key insights: (1) privacy harms are increasingly going to come in the form of inferences at the hands of predictive algorithms.[1] 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.

Our resulting notion, which we call *demographic coherence*, is intentionally designed to be *ergonomic*[2] 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.

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:

  – *Captures predictive harms.* Demographic coherence builds on conceptual tools from generalization [55, 23, 18, 34, 8] and multicalibration [30, 41] 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

---

[1]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.

[2]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.

of the concrete risk. In this work we study an instantiation of demographic coherence measured using 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. For an overview of these theorems, see Section 2.

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.,* [32, 26, 50, 56]), reconstruction (*e.g.,* [22, 14, 29, 11]), and reidentification (*e.g.,* [52, 42]), as shown by Dwork et. al.[25]. 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.[3] In recent work, Dick et al. implemented a reconstruction attack [20] 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" [39]. 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 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 to carry out, and, furthermore, conducting a reconstruction attack assumes malicious intent in a way that may or may not be convincing

---

[3]*e.g.,* reconstruction of features like gender can be carried out simply from knowing population statistics rather than breaking anonymity [48].

to all stakeholders. We introduce the demographic coherence 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.,* [37, 36, 44, 51]). 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 [46], Salem et al. [49], 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. [28] 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 5, 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.[4]

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,*

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

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.*

The following theorem is an informal statement of Theorem 6, which argues that any data-curation algorithm with bounded max-information [23] (a notion that mathematically captures the dependence of algorithms' outputs to their inputs) also enforces Wasserstein coherence.

Theorem 1 (Informal Version of Theorem 6). *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*

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

---

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

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 5.

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. We leave it as an important open question to identify 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

In order to clearly motivate and explain the choices embedded within our definition, we incrementally build up our approach in this section; for the formal definition see Section 4.

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 in Census-like circumstances, in which the goal is to sample the entire population. Our ultimate goal is to reason about a data curator $\mathcal{A}$ who uses X to generate a privacy-preserving release R.[5]

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.[6] We choose to measure privacy risk in terms of predictive harms for the following

---

[5]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 $\mathcal{L}$.

[6]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.

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 [35, 2, 45] 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.

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.,* [19]) 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 [43]. 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[7] individuals Asahi and Blair; each intentionally avoids being perceived as queer, and in particular does not provide ratings on movies with

---

[7]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.

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 the assumption that we claim 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 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 to 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.[8] 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.

   While this approach gets to the core of our interests, it has an important flaw. Technically speaking, we can not 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

---

[8]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)?

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 to be true, and values in $[-1, 0]$ when it predicts the attribute likely to be false, with a higher absolute value representing higher confidence.

For any such predictor, we will consider $h(X_a)$ and $h(X_b)$ as representing the uniform distribution 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 as explained 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...[9] 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 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

---

[9]We borrow this conceptual approach from [31].

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 can "see." We then compare the behavior of h on members of X_a and X_b as seen through the lens ρ.[10]

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. For example, if one visualizes a probability distribution as a mount of dirt, *Wasserstein distance* measures the effort required to move enough dirt to make one mount look like the other (thus, earth-mover's distance). Unlike Total Variation distance, which only measures the distance between the distribution curves, Wasserstein distance is greater with a higher shift in confidence; 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 4.1 contains a glossary of the notation we use, Section 4.2 formally defines the notion of *incoherence* that we measure, Section 4.3 defines what it means for an algorithm to be *demographically incoherent*, and Section 4.4 defines what it means for a data curation algorithm to be *coherence enforcing*.

4.1 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 $X \in \mathcal{X}^n$. [11]

- 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.

---

[10]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 $\mathcal{L}$.

[11]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.

- We define a lens ρ as a set of features from $\mathcal{X}$.

- 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 $\bot$.

- 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

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

4.2 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.

4.3 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 metric distance 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 under

our approach in Section 3.2, expect we are testing the coherence of an algorithm that gets the dataset in the clear. In Section 4.4, 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 $\mathcal{C}$.

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

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 $\text{dist}(\cdot,\cdot)$.

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

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

Incoherence Condition:
 Set $b = 0$ if there exists $C \in \mathcal{C}$ such that:

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

 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:

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

However, this definition cannot be realized with respect to arbitrary categories C and all datasets, because the sampling experiment in and of 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 meaningful and achievable, we need that the datasets considered are sufficiently representative that the incoherence due to sampling does not dominate. Thus,

15

in order to make the definition achievable, 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 one can prove may only hold for large choices 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 $X \in \mathcal{X}^n$, $\mathcal{C}^* = \{C \in \mathcal{C} \mid |C \cap X| \geq \gamma\}$

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

4.4 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 many any changes to the algorithm itself.

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 that $\mathcal{A}$ 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.

4.5 Instantiating Demographic Coherence with a Metric

In the definitions above, we do not use a specific metric. The metric we will use to instantiate these definitions in the following sections is 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 Algorithms that Enforce Wasserstein Demographic Coherence

Now, we show that Wasserstein coherence enforcement is instantiable. Firstly, in Section 5.1, we go over some technical preliminaries necessary to state and prove our results, Then, in Section 5.2, we prove Theorem 6 showing that algorithms with bounded max-information are coherence enforcing. Building on this, in Section 5.3, we leverage the connection between differential privacy and max-information to prove Theorem 7 which says that pure-DP algorithms enforce demographic coherence, and Theorem 8 which is says that approx-DP algorithms enforce demographic coherence as well.

5.1 Technical Preliminaries

For a recap of the notation used, see Section 4.1.

Definition 6 (Differential Privacy [24])**.** *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 differential privacy (approx-DP).

Definition 7 (Max-information of random variables [23])**.** *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 8 (Max-information of algorithms [23]). [12] *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 following definition of order-invariant algorithms appears as a technical assumption in some of our theorem statements.[13] 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 9 (Order-invariant algorithm). *An algorithm $\mathcal{A} : \mathcal{X}^m \to \mathcal{Y}$, is order invariant if for all $X \in \mathcal{X}^m$, the distribution of the random variable $\mathcal{A}(X)$ does not depend on the order of the elements of X.*

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

Theorem 3. (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 [47]. 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 4. (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*

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

---

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

[13]Since it is an assumption in the version of McDiarmid's Inequality for sampling without replacement (Theorem 9) that we use.

Definition 10 (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 $H(b, a, s)$).*

Theorem 5 ([33]). *Let K have a hypergeometric distribution $H(b, a, s)$. Then for every $\gamma \geq 2$,*

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

where

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

5.2 Bounded Max-Information Algorithms are Coherence Enforcing

Theorem 6. *Let $n \in \mathbb{N}, \zeta > 0$, $\beta \in (0,1)$, $\alpha \in (0,1]$. Let $\mathsf{dist}_{\mathcal{W}_1}(\cdot, \cdot)$ represent 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

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

The intuition behind the result in Theorem 6 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 6 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*

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

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

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

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

Proof sketch of Lemma 1: Considering a particular subpopulation $C \subseteq \mathcal{X}$, We need to show for all algorithms $\mathcal{L} : \mathcal{Y} \to \{\mathcal{X} \to [-1, 1]\}$, all datasets $X \in \mathcal{X}^n$, $\mu > 0$ that with high probability over a choice of split $X_a, X_b \xleftarrow{\$} X$, and predictor $h \leftarrow \mathcal{L} \circ \mathcal{A}(X_a)$ as in the $\mathsf{DemCoh}_{\mathcal{L} \circ \mathcal{A}, X, C^*, \rho}(\alpha)$ experiment in Figure 1, the following holds:

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

Note that instead of the split X_a, X_b which predictor h depends on, if we consider an independent split $S, \overline{S} \xleftarrow{\$} X$, then we could hope to use a concentration inequality to get the bound we desire. (The proof of Claim 2 below redefines the sampling process in a way that allows us to use a concentration bound of Hush and Scovel (Theorem 5) for the hypergeometric distribution to prove such a result.)

With the goal of analysing an independent split, we combine the fact that bounded max-information is preserved under post-processing with the intuition that the output of an algorithm with bounded max-information does not contain too much specific information about the input dataset to decouple the predictive hypothesis h from the dataset X_a in the following way (in Claim 1):

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

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, $I^\eta_{\infty, P}(\mathcal{A}, n/2) \le \zeta$, 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*

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

Then, in Claim 2, we bound $\Pr[\mathsf{dist}_{\mathcal{W}_1}(g(\boldsymbol{\pi}_\rho(S|_C)), h(\boldsymbol{\pi}_\rho(\overline{S}|_C))) > \alpha]$ 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 $|X \cap C| \ge \max\{\frac{4.15}{\alpha^2} \ln(4/\mu), \frac{5.3}{\alpha}, 8.3 \ln(4/\mu), 40\}$, we have that*

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

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 $(S, \mathcal{A}^*(X_a))$,

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

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 $(h(\boldsymbol{\pi}_\rho(\overline{S}|_C)))$.

Let $T = \{(S, h) \mid \mathsf{dist}_{\mathcal{W}_1}(h(\boldsymbol{\pi}_\rho(S|_C)), h(\boldsymbol{\pi}_\rho(\overline{S}|_C))) > \alpha\}$. 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:

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

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

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

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.

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

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

To show Equation (4), we redefine the sampling process in a way that will allow us to use Theorem 5, a concentration result for the hypergeometric distribution (See Definition 10 for a definition of this distribution): 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 $g(\pmb{\pi}_\rho(x)) \leq y$ 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 $g(\pmb{\pi}_\rho(x)) \leq y$. The random variables W and V follow hypergeometric distributions as follows:

$$V \sim H(n, m, n/2)$$
$$(W|V = v) \sim H(m, t, v).$$

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 5 and using $m > 40$ and $m > 8.3\ln(4/\mu)$ we have that

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

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., $\left|W - \frac{tv}{m}\right| \leq \zeta$). Then applying Theorem 5 for $\zeta \geq 2$ and c as in the theorem:

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

22

Observe that

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

Therefore,

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

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 $\Pr[\overline{E_2} \mid V = v] < \mu/2$.

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,

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

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:

23

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

$$= 2\exp\left(\frac{-4}{\frac{m}{2}+1} \cdot \left(\frac{9m^2\alpha^2}{16^2} - 1\right)\right)$$

$$\leq 2\exp\left(\frac{-8}{1.05m} \cdot \left(\frac{0.9 \cdot 9m^2\alpha^2}{16^2}\right)\right)$$

$$= 2\exp\left(\frac{8.1\alpha^2 m}{33.6}\right) < \mu/2.$$

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 6. ∎

Proof of Theorem 6. 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 6. 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 $X \in \mathcal{X}^n$, 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)

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

where the choice of split $X_a, X_b \xleftarrow{\$} X$ 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 need to show, instead, that for

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

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:

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

where the split $X_a, X_b \xleftarrow{\$} X$ 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 $2\,\eta$).

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.

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

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 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

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

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

5.3 Differentially Private Algorithms Enforce Demographic Coherence

In this section, we argue that our definition of demographic coherence can be achieved via differentially private algorithms. We do this by adapting known connections between differential privacy and max-information, and Theorem 6 connecting max-information and demographic coherence.

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 6 (connecting bounded max-information to demographic coherence) to argue that differential privacy implies demographic coherence.

Theorem 7. *[Pure-DP Enforces Wasserstein Coherence] Fix any $\varepsilon, \beta \in (0,1]$, $\alpha \in (0,1]$, $n \in \mathbb{N}$. 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*

$$\gamma = \max \left\{ \frac{8.3 \cdot (\varepsilon^2 n/4 + \varepsilon\sqrt{n\ln(4|\mathcal{C}|/\beta)}/2 + \ln(16|\mathcal{C}|/\beta))}{\alpha^2}, \frac{36\ln((3/\alpha)}{\alpha^2} \right.$$
$$\left. , 16.6 \cdot (\varepsilon^2 n/4 + \varepsilon\sqrt{n\ln(4|\mathcal{C}|/\beta)}/2 + \ln(16|\mathcal{C}|/\beta)), \frac{5.3}{\alpha}, 80 \right\}. \tag{12}$$

Proof. Fix $\beta > 0$. By Theorem 11 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 6 and substituting the bound on max-information then completes the proof. ∎

Theorem 8. *[Approx-DP Enforces Wasserstein Coherence] Fix any $\beta \in (0,1]$, $\alpha \in (0,1]$, $n \in N$. 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 [14] (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*

$$\gamma = \max \left\{ \frac{8.3 \cdot (265\varepsilon^2 n + 12\varepsilon\sqrt{n\ln(4|\mathcal{C}|/\beta)} + \ln(32|\mathcal{C}|/\beta))}{\alpha^2}, \frac{36\ln((3/\alpha)}{\alpha^2} \right.$$
$$\left. , 16.6 \cdot (265\varepsilon^2 n + 12\varepsilon\sqrt{n\ln(4|\mathcal{C}|/\beta)} + \ln(32|\mathcal{C}|/\beta)), \frac{16/3}{\alpha}, 80 \right\}. \tag{13}$$

---

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

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 6 and substituting the bound on max-information then completes the proof. ∎

6 Acknowledgments

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-

[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/simula May 2021.

[11] Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Moham-mad 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 Asso-ciation, 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. ATTAXON-OMY: 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] 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.

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

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[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] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

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

[40] 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-ris` Sep 2022.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

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

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

[55] 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.

[56] 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 11) 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 11 (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 11. 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 2.

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

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 $R_a \leftarrow \mathcal{A}(X_a)$.

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 6 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 [23, 47]. In this appendix we prove Theorem 11 and Theorem 12, which are analogs of those theorems for sampling without replacement.

B.1 Preliminaries

First we state Theorem 9, a version of McDiarmid's inequality that applies to the case of sampling without replacement. This result follows from Lemma 2 in [54]. This result in used in the proof of Theorem 11, which says that pure-DP algorithms have bounded max-information (even in the case of sampling without replacement.)

Definition 12. *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 9 (McDiarmid's for sampling without replacement [54])**.** *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 12 and Corollary 2 which say that approximate-DP algorithms have bounded max-information (even in the case of sampling without replacement.)

Definition 13 (Point-wise indistinguishibility [38])**.** *Two random variables A, B are pointwise (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 [38])**.** *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 [38])**.** *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 10 (Azuma's Inequality)**.** *Let* $C_1, \cdots, C_n$ *be a sequence of random variables such that for every* $i \in [n]$*, we have*

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

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

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

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 11, which is an analog of Theorem from [23]. The proof of this theorem works exactly as in [23], except replacing the application of McDiarmid's Lemma with a version of McDiarmid's for sampling without replacement (Theorem 9) which we state in Appendix B.1.

Theorem 11. (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 12, which is an analog of Theorem 1 from Rogers et al.[47]. In fact, the following proof is almost exactly the proof of them from [47] 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 12 to get Corollary 2, which is used in the proof of Theorem 8.

Theorem 12. (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

$$Z_i(a, \mathbf{x}_{[i]}) \stackrel{\text{def}}{=} \log \frac{\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]}. \tag{14}$$

$$\begin{aligned} Z(a, \mathbf{x}) &\stackrel{\text{def}}{=} \log\left(\frac{\Pr_{\mathbf{x}}\left[\mathcal{A}(x) = a, \mathbf{X} = \mathbf{x}\right]}{\Pr\left[\mathcal{A} = a\right] \cdot \Pr\left[\mathbf{X} = \mathbf{x}\right]}\right) \\ &= \sum_{i=1}^n Z_i(a, \mathbf{x}_{[i]}) \end{aligned} \tag{15}$$

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 ([23, 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 10) 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* $\delta'' \stackrel{\text{def}}{=} \frac{2\hat{\delta}}{1 - e^{-3\varepsilon}}$. *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:*

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

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

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

$\blacksquare$

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:*

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

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.

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 (where the first inequality follows from DP.):

$\Pr\left[(\mathcal{A}(\mathbf{X}), X_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right]$

$= \displaystyle\sum_{x_i \sim X_i|_{\mathbf{x}_{[i-1]}}} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}, x_i\right]$

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

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

$= \displaystyle\sum_{x_i \sim X_i|_{\mathbf{x}_{[i-1]}}} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \left(e^{\varepsilon}\Pr\left[(\mathcal{A}(\mathbf{X}), x_i) \in \mathcal{O} \mid \mathbf{x}_{[i-1]}\right] + \delta\right)$

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

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

[15]

Applying a very similar argument, will give us that

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

∎

---

[15]In the step where we apply DP, the parameters will double if the algorithm is not order-invariant.

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.

We will use the following shorthand notation for conditional expectation:

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

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

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 (16).

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:

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

$$= \sum_{x_i \in S} \Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] Z_i(a, x_{[i]})$$

$$= \sum_{x_i \in S} \Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] \log\left(\frac{\Pr\left[X_i = x_i | a, \mathbf{x}_{[i-1]}\right]}{\Pr\left[X_i = x_i | \mathbf{x}_{[i-1]}\right]}\right)$$

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.

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

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

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

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

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

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

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:

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

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

$$\leq \log(1 - q) - \log(1 - (e^{3\varepsilon} q + \hat{\delta}))$$

$$\leq \log(e) \cdot (-q + e^{3\varepsilon} q + \hat{\delta} + 2(e^{3\varepsilon} q + \hat{\delta})^2)$$

$$= \log(e) \cdot ((e^{3\varepsilon} - 1)q + \hat{\delta} + 2(e^{3\varepsilon} q + \hat{\delta})^2)$$

$$\stackrel{\text{def}}{=} \tau(\hat{\delta})$$

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:

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

$$\leq \sum_{x_i \in S} \Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] \log\left(\frac{\Pr[X_i = x_i | a, \mathbf{x}_{[i-1]}]}{\Pr[X_i = x_i | \mathbf{x}_{[i-1]}]}\right)$$

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

$$= \sum_{x_i \in S} \left(\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] - \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]\right) \log\left(\frac{\Pr[X_i = x_i | a, \mathbf{x}_{[i-1]}]}{\Pr[X_i = x_i | \mathbf{x}_{[i-1]}]}\right) + \tau(\hat{\delta})$$

$$\leq_{|Z_i| \leq 6\varepsilon} 6\varepsilon \sum_{x_i \in S} \left|\Pr\left[X_i = x_i \mid a, \mathbf{x}_{[i-1]}, |Z_i| \leq 6\varepsilon\right] - \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right]\right| + \tau(\hat{\delta})$$

$$\leq_{\text{Def of } S,\text{ Claim 3}} 6\varepsilon \sum_{x_i \in S} \Pr\left[X_i = x_i \mid \mathbf{x}_{[i-1]}\right] \max\left\{\frac{e^{6\varepsilon}}{\Pr\left[|Z_i| \leq 6\varepsilon \mid a, \mathbf{x}_{[i-1]}\right]} - 1, 1 - \frac{e^{-6\varepsilon}}{\Pr\left[|Z_i| \leq 6\varepsilon \mid a, \mathbf{x}_{[i-1]}\right]}\right\} + \tau(\hat{\delta})$$

$$\leq_{\text{Claim 3}} 6\varepsilon \left(\frac{e^{6\varepsilon}}{1 - \frac{2\hat{\delta}}{1 - e^{-3\varepsilon}}} - 1\right) + \tau(\hat{\delta})$$

$$\leq_{\text{Substituting for } \tau(\hat{\delta})} 6\varepsilon \left(e^{6\epsilon}\left(1 + \frac{4\hat{\delta}}{1 - e^{-3\epsilon}}\right) - 1\right) + \log(e) \cdot ((e^{3\varepsilon} - 1)q + \hat{\delta} + 2(e^{3\varepsilon} q + \hat{\delta})^2)$$

$$\leq_{\text{Upper bound for } q} 6\varepsilon \left(e^{6\epsilon}\left(1 + \frac{4\hat{\delta}}{1 - e^{-3\epsilon}}\right) - 1\right)$$

$$+ \log(e) \cdot \left((e^{3\varepsilon} - 1)\frac{2\hat{\delta}}{1 - e^{-3\epsilon}} + \hat{\delta} + 2\hat{\delta}^2 + 8\frac{\hat{\delta}^2 e^{6\varepsilon}}{(1 - e^{-3\varepsilon})^2} + 8\frac{\hat{\delta}^2 e^{3\varepsilon}}{1 - e^{-3\varepsilon}} \right)$$

$$=_{b = \frac{\hat{\delta}}{1 - e^{-3\epsilon}}} 6\varepsilon \left(e^{6\epsilon}(1 + 4b) - 1 \right) + \log(e) \cdot \left(b \left(2e^{3\varepsilon} - 2 + 8\frac{\hat{\delta} e^{6\varepsilon}}{(1 - e^{-3\varepsilon})} + 8\hat{\delta}e^{3\varepsilon} \right) + \hat{\delta} + 2\hat{\delta}^2 \right)$$

$$= b \left(24\varepsilon e^{6\varepsilon} + 2e^{3\varepsilon} - 2 + 8\frac{\hat{\delta} e^{6\varepsilon}}{(1 - e^{-3\varepsilon})} + 8\hat{\delta}e^{3\varepsilon} \right) + \hat{\delta} + 2\hat{\delta}^2 + 6\varepsilon(e^{6\varepsilon} - 1)$$

$$= \frac{\hat{\delta}}{1 - e^{-3\varepsilon}} \left(2e^{3\varepsilon}(4e^{3\varepsilon}(3\varepsilon + \frac{\hat{\delta}}{1 - e^{-3\varepsilon}})) + 4\hat{\delta} + 1) - 2 \right) + \hat{\delta} + 2\hat{\delta}^2 + 6\varepsilon(e^{6\varepsilon} - 1)$$

$$\leq_{e^{-3\varepsilon} \leq 1 - 1.5\varepsilon \text{ for } \varepsilon \in [0, 0.5]} 2\frac{\hat{\delta}}{1.5\varepsilon}e^{3\varepsilon} \left(4e^{3\varepsilon}(3\varepsilon + \frac{\hat{\delta}}{1.5\varepsilon})) + 4\hat{\delta} + 1 \right) + \hat{\delta} \left(\frac{-2}{1.5\varepsilon} + 2\hat{\delta} + 1 \right) + 6\varepsilon(e^{6\varepsilon} - 1)$$

$$\leq 8\frac{e^{6\varepsilon}\hat{\delta}}{1.5\varepsilon} \left(3\varepsilon + \frac{\hat{\delta}}{1.5\varepsilon} \right) + 2\frac{\hat{\delta}}{1.5\varepsilon}e^{3\varepsilon} \left(4\hat{\delta} + 1 \right) + \hat{\delta} \left(\frac{-2}{1.5\varepsilon} + 2\hat{\delta} + 1 \right) + 6\varepsilon(e^{6\varepsilon} - 1)$$

$$\leq_{e^{3\varepsilon} \leq 1 + 7\varepsilon, e^{6\varepsilon} \leq 1 + 40\varepsilon \text{ for } \varepsilon \in [0, 0.5]} 8\frac{(1 + 40\varepsilon)\hat{\delta}}{1.5\varepsilon} \left(3\varepsilon + \frac{\hat{\delta}}{1.5\varepsilon} \right) + 2\frac{(1 + 7\varepsilon)\hat{\delta}}{1.5\varepsilon} \left(4\hat{\delta} + 1 \right)$$

$$+ \hat{\delta} \left(\frac{-2}{1.5\varepsilon} + 2\hat{\delta} + 1 \right) + 6\varepsilon(40\varepsilon)$$

$$\leq \frac{(8 + 320\varepsilon)\hat{\delta}}{1.5\varepsilon} \left(3\varepsilon + \frac{\hat{\delta}}{1.5\varepsilon} \right) + \frac{(2 + 14\varepsilon)}{1.5\varepsilon} \left(4\hat{\delta}^2 + \hat{\delta} \right) - \frac{2\hat{\delta}}{1.5\varepsilon} + 2\hat{\delta}^2 + \hat{\delta} + 240\varepsilon^2$$

$$\leq_{\varepsilon < 0.5} \frac{168\hat{\delta}}{1.5\varepsilon}(3\varepsilon + \frac{\hat{\delta}}{1.5\varepsilon}) + \frac{8}{1.5}\frac{\hat{\delta}^2}{\varepsilon} + \frac{56}{1.5}\hat{\delta}^2 + \frac{14}{1.5}\hat{\delta} + 2\hat{\delta}^2 + \hat{\delta} + 240\varepsilon^2$$

$$\leq_{\varepsilon \leq 0.5} 347\hat{\delta} + 75 \left(\frac{\hat{\delta}}{\varepsilon} \right)^2 + 24\frac{\hat{\delta}^2}{\varepsilon} + 240\varepsilon^2$$

$$\stackrel{\text{def}}{=} \nu(\hat{\delta}) \tag{16}$$

∎

Finally, we need to apply Azuma's inequality (stated in Theorem 10) 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:

$$\mathcal{G}_i(\hat{\delta}) = \left\{ (a, \mathbf{x}_{[i]}) \ \middle| \ |Z_i(a, \mathbf{x}_{[i]})| \leq 6\varepsilon \ \& \ X_i\mid_{\mathbf{x}_{[i-1]}} \approx_{3\varepsilon, \hat{\delta}} X_i\mid_{a, \mathbf{x}_{[i-1]}} \right\}, \tag{17}$$

$$\mathcal{G}_{\leq i}(\hat{\delta}) = \left\{ (a, \mathbf{x}_{[i]}) : (a, x_1) \in \mathcal{G}_1(\hat{\delta}), \cdots, (a, \mathbf{x}_{[i]}) \in \mathcal{G}_i(\hat{\delta}) \right\} \tag{18}$$

$$T_i(a, \mathbf{x}_{[i]}) = \begin{cases} Z_i(a, \mathbf{x}_{[i]}) & \text{if } (a, \mathbf{x}_{[i]}) \in \mathcal{G}_{\leq i}(\hat{\delta}) \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

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 12. For any constant ν, we have:

$$\Pr\left[\sum_{i=1}^{n} Z_i(\mathcal{A}, \mathbf{X}_{[i]}) > n\nu + 6t\varepsilon\sqrt{n}\right]$$

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

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

We then substitute ν by $\nu(\hat{\delta})$ as defined in Equation (16), 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 10, respectively. Therefore,

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

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 5.3 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 12, 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. ∎