# SoK: Differential Privacy as a Causal Property

Michael Carl Tschantz*, Shayak Sen†, and Anupam Datta†
*International Computer Science Institute †Carnegie Mellon University

***Abstract*—We present formal models of the associative and causal views of differential privacy. Under the associative view, the possibility of dependencies between data points precludes a simple statement of differential privacy's guarantee as conditioning upon a single changed data point. However, we show that a simple characterization of differential privacy as limiting the effect of a single data point does exist under the causal view, without independence assumptions about data points. We believe this characterization resolves disagreement and confusion in prior work about the consequences of differential privacy. The associative view needing assumptions boils down to the contrapositive of the maxim that correlation doesn't imply causation: differential privacy ensuring a lack of (strong) causation does not imply a lack of (strong) association. Our characterization also opens up the possibility of applying results from statistics, experimental design, and science about causation while studying differential privacy.**

## I. Introduction

Differential Privacy (DP) is a precise mathematical property of an algorithm requiring that it produce almost identical distributions of outputs for any pair of possible input databases that differ in a single data point. Despite the popularity of DP in the research community, unease with the concept remains. For example, Cuff and Yu's paper states "an intuitive understanding can be elusive" and recommends that DP be related to more familiar concepts based on statistical associations, such as mutual information [8, p. 2]. This and numerous other works exploring similar connections between DP and statistical association each makes assumptions about the data points (e.g., [1, p. 9] [6, p. 32], [2, p. 4], [34, p. 14], [20, p. 6]).

The use of such assumptions has led to some papers stating that DP implicitly requires some assumption: that it requires the data points to be independent (e.g., [27, p. 2], [28, p. 1], [32, p. 2], [23, p. 3], [5, p. 7], [50, p. 232], [33, p. 1]), that the adversary must know all but one data point, the so-called *strong adversary assumption* (e.g., [8, p. 2], [32, p. 10]), or that either assumption will do (e.g., [49, §1.2]). (Appendix A3 provides quotations.) Conversely, other works assert that no such assumption exists (e.g., [3], [25], [36], [35]). How can such disagreements arise about a precise mathematical property of an algorithm?

We put to rest both the nagging feeling that DP should be expressed in more basic terms and the disagreement about whether it makes various implicit assumptions. We do so by showing that DP is better understood as a causal property than as an associative one. We show that DP constrains effect sizes, a basic concept from empirical science about how much changing one variable changes another. This view does not require any independence or adversary assumptions.

Furthermore, we show that the difference between the two views over whether DP makes assumptions is precisely captured as the difference between association and causation. That some fail to get what they want out of DP (without making an assumption) comes from the contrapositive of the maxim *correlation doesn't imply causation*: DP ensuring a lack of (strong) causation does not imply a lack of (strong) association. Given the common confusion of association and causation, and that DP does not make its causal nature explicit in its mathematical statement, we believe our work explains how disagreement could have arose in the research literature about the what assumptions DP requires.

### A. Motivating Example and Intuition

To provide more details, let us consider an example of using DP inspired by Kifer and Machanavajjhala [27]. Suppose Ada and her son Byron are considering participating in a differentially private survey with $n-2$ other people. The survey collects a data point from each participant about their health status with respect to a genetic disease. Since Ada and Byron are closely related, their data points are closely related. This makes them wonder whether the promise of DP becomes watered down for them, a worrying prospect given the sensitivity of their health statuses.

Figure 1 summarizes what would happen if both Ada and Byron participate in the survey. In it, each solid arrow represents a causal relationship where the quantity at the start of the arrow causally affects the quantity at the end of the arrow. For example, Arrow (1) represents that Ada's genetics has a causal effect on her son Byron's genetics. We use an arrow since causation is directional: Byron's genetics does not have a causal effect on Ada's. Arrow (2) represents a mechanism by which Ada provides her status to the survey. This information becomes a data point in the survey's data set, that is, a row in a database. This database comprises Ada's data point, Byron's data point, and $n-2$ other people's data points. Arrows (5), (6), and (7) together represent the algorithm that computes the survey's result, that is, the output produced from the database using a differentially private algorithm.

As mentioned, Ada's status also affects the status of her son Byron, shown with Arrow (1). Therefore, their statuses are statistically associated (i.e., not probabilistically independent). While causation is directional, such associations are not: seeing Byron's status reveals information about Ada's status despite not causing Ada's status. Furthermore, Ada's and Byron's data points will be statistically associated because they have a common cause, Ada's status. Thus, seeing Byron's data point reveals information about Ada's status and data point. Since both Ada's and Byron's data points reveal information about Ada's status, the output can be informed by two data points about Ada's status. This double dose of information is

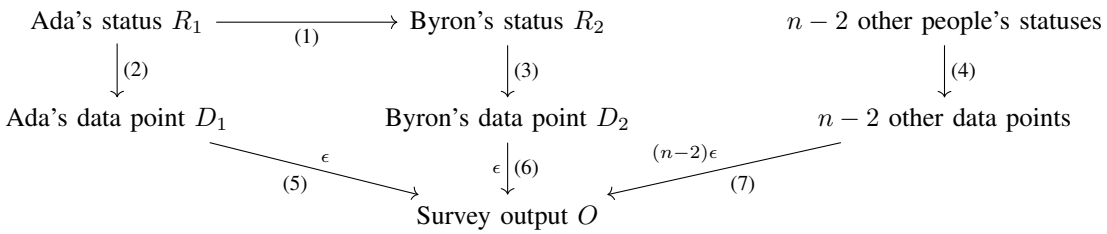



Fig. 1. A causal diagram approximating the process through which the output of a statistical query is generated and used. The arrows represent direct causal effects. Indirect cause effects can be inferred from taking the transitive closure of the arrows. $\epsilon$ labels causal effects bounded by $\epsilon$-differential privacy. (1)–(7) serve as labels naming arrows.

what gives Ada pause about participating. Furthermore, much the same applies to Byron.

In the words of Kasiviswanathan and Smith [25, p. 2], DP intuitively ensures that

> changing a single individual's data in the database leads to a small change in the distribution on outputs. $(*)$

This intuitive consequence of DP, denoted as "$(*)$", does not make explicit the notion of *change* intended. It implicitly compares the distribution over the output, a random variable $O$, in two hypothetical worlds, the pre- and post-change worlds. If we focus on the individual Ada and let $D_1$ be a random variable representing her data point as it changes values from $d_1$ to $d_1'$, then the comparison is between $\Pr[O{=}o \text{ when } D_1{=}d_1]$ and $\Pr[O{=}o \text{ when } D_1{=}d_1']$. The part of this characterization of DP that is informal is the notion of *when*, which leaves the notion of *change* imprecise. Our paper contrasts various interpretations of *change* and *when*.

The most obvious interpretation is that of *conditioning* upon two different values for the changed status. This interpretation implies an approximation of statistical independence between an individual's data point and the output: $\Pr[O{=}o \mid D_1{=}d_1] \approx \Pr[O{=}o \mid D_1{=}d_1']$. Presuming the data points are truthful, such an approximate independence implies (up to a factor) an approximate independence that compares probabilities over a status with or without knowing the output, that is, $\Pr[R_1{=}r_1 \mid O{=}o] \approx \Pr[R_1{=}r_1]$. In this case, observing the output reveals little about an individual's status, explaining this interpretation's appeal.

However, as discussed above, both Ada's and Byron's data points reveal information about each of their statuses since associations depend upon the full breadth of causal relations. This double dose of information about their statuses means that DP does not actually imply this appealing form of approximate independence. Thus, attempts to interpret DP in terms of conditioning fail to hold in the presence of the associations between the data points. Those desiring an associative guarantee from DP must rule out such double doses of information, for example, by assuming that the data points lack any associations or that the adversary already knows all but one data point, making such associations uninformative.

Now, let us instead consider interpreting DP in terms of *causal interventions*. This interpretation models artificially altering the value of random variables, as in a randomized experiment. The key difference between intervening upon a random variable and conditioning upon it is that while intervening tracks causal effects by accounting for how the intervention may cause other variables to change, it does not depend upon all the associations in the database since such interventions break them. Thus, while associative definitions using conditioning depends upon the distribution producing data points, causal ones can screen off this distribution to examine the behavior of just the DP algorithm itself by intervening upon all its outputs.

For example, suppose Byron is born without the genetic disease and a scientist flips a coin and ensures that Byron has the disease if it comes up heads and ensures that he does not if it comes up tails. (While the technology to execute this experiment is currently wanting, it is conceptually possible.) Since Bryon starts without the disease, the tails outcome does nothing and can be viewed a control treatment while the heads outcome causes a change. If it comes up heads, the scientist could measure various things about Byron to see what changed from giving him the disease. In particular, Byron's data point and the output computed from it would change. On the other hand, nothing would change about Ada since causation is directional. (Section IV makes this more precise.) In fact, after the randomization, Bryon's status and data point no longer reveals any information about Ada's status since the randomization broke the association between their statuses.

The scientist can measure the size of any changes to compute an *effect size*. The effect size for Byron's data point would be large since the data point is supposed to be equal to the status, but the effect size for the output will be small since it is computed by an algorithm with $\epsilon$-differential privacy. If we instead consider intervening on Ada's status, we find two paths to the output: one via Ada's data point (Arrows (2) and (5)) and another via Byron's (Arrows (1), (3), and (6)). These two paths mean that the effect size could be as much as double that of changing Byron's status. Thus, DP cannot be interpreted as limiting the effect of changing Ada's status to just $\epsilon$ in size.

Recall that the intuitive characterization $(*)$ of DP referred to data points, not statuses: "changing a single individual's data in the database…" [25, p. 2]. So, let us consider intervening upon the data points instead. Each data point is piped directly into the differentially private algorithm and has no other effects. Thus, DP does bound the effect size at $\epsilon$ for Ada's data point without making any assumptions about the statuses. For this reason, we believe DP is better understood as a bound on effect sizes than as a bound on associations.

We believe that ease of conflating associative and causal properties explains the disagreement in the research literature.

(See Appendix A for a history of this disagreement.) Our observation also reduces the benefits and drawbacks of these implicitly associative and causal views of privacy to those known from studying association and causation in general. For example, the causal view only requires looking at the system itself (causation is an inherent property of systems) while the associative view requires looking at the input distribution as well. This difference explains why papers implicitly with the associative view discuss the distribution over data points despite the definition of DP and the implicitly causal papers do not mention it.

The causal characterization also requires us to distinguish between an individual's attributes ($R_i$s) and the data that is input to an algorithm ($D_i$s), and intervenes on the latter. Under the assumption that individuals report their true statuses, the associative interpretation does not require this distinction since conditioning on one is identical to conditioning on the other. This distinction captures an aspect of the difference between protecting "secrets about you" ($R_i$) and protecting "secrets from you" ($D_i$) pointed out by McSherry [36], [35], where DP protects the latter in a causal sense. An individual's attribute $R_i$ is *about* him and its value is often outside of his control. On the other hand, an individual's data point $D_i$, at least in the setting typically envisioned for DP, is under his control and is volunteered by the individual, making it *from* him.

### B. Overview

Our main goal is to demonstrate that DP can be understood as a causal property without needing the sorts of assumptions made to view it as an associative property. We lay out the associative view by surveying definitions presented in prior work to show its awkward fit for DP and how it leads to suggestions that DP makes assumptions. We then turn to the causal view, replacing conditioning with interventions in the associative definitions. Doing so reveals three key insights; we find that the causal definitions (1) work without such assumptions, (2) provides a tight characterization of DP, and (3) explains how DP maps to a concept found throughout statistics and science, namely to a measure of effect sizes.

We start our analysis with the associative view, which uses conditioning (Section III). We first consider conditioning upon all the data points instead of just the changed one. After dealing with some annoyances involving the inability to condition on zero-probability data points, we get a precise characterization of DP (Definition 2). However, this associative definition does not correspond well to the intuitive characterization $(*)$ of differential privacy's key consequences: whereas the above-quoted characterization refers to just the changed data point, this associative definition refers to them all, thereby blurring the characterization's focus on change.

Next, we modify the associative definition to condition upon just the single changed data point (Definition 3). The resulting definition prohibits more than an $\epsilon$ degree of correlation between the data point and the output, hereby limiting what can be learned about the data point. While this definition is not implied by DP on its own, it is implied with an additional assumption of independence between data points (Definition 4). We believe that this explains the claim found in some papers that DP implicitly assumes independence.

However, we do not share this feeling since the independence assumption is not required to get DP to imply the intuitive consequence $(*)$ quoted above when interpreting *change* as a causal intervention instead of as associative conditioning. After reviewing the core concepts of causal modeling (Section IV), we consider intervening upon all the data points (Section V-A). As with conditioning upon all the data points, a definition intervening on all the data points (Definition 6) characterizes DP (Proposition 3) but without the intuitive focus on a single data point that we desire.

We then consider characterizing DP as intervening upon a single point (Definition 7 of Section V-B). A benefit of this causal characterization is that it is implied by DP without any assumptions about independence (Proposition 4). An additional benefit is that, unlike the associative characterizations, we do not need side conditions limiting the characterization to data points with non-zero probabilities. This benefit follows from causal interventions being defined for zero-probability events unlike conditioning upon them. These two benefits lead us to believe that DP is better viewed as a causal property than as an associative one.

In addition to considering the consequences of DP through the lenses of association and causation, we also consider how these two approaches can provide definitions equivalent to DP. Table I shows our key results about definitions that are either equivalent to DP or might be mistaken as such, which, in the sections below, we weave in with our aforementioned results about characterizations of the consequences of DP.

When intervening upon all data points, we get equivalence for free from Definition 6 that we already explored as a characterization of the consequences of DP. This free equivalence does not occur for conditioning upon all data points since the side condition ruling out zero-probability data points means those data points are not constrained. Since DP is a restriction on all data points, to get an equivalence, the definition must check all data points. To achieve this, we further require that the definition hold on all distributions over the data points, not just the naturally occurring distribution. (Alternatively, we could require the definition to hold for any one distribution with non-zero probabilities for all data points, such as the uniform distribution.) We also make similar alterations to the definitions looking at a single data point.

Having shown that DP can be viewed as a causal property, we then consider how this view can inform our understanding of it. We relate DP to a previously studied notion of effect size and discuss how this more general notion can make discussions about privacy more clear (Section VI). In particular, DP is a bound on the measure of effect size called *relative probabilities* (also known as *relative risk* and *risk ratio*). That is, DP bounds the relative probabilities for the effects of each data point upon the output. Since not all research papers are in agreement about what counts as an individual's data point, spelling out exactly which random variables have bounded relative probabilities may be more clear than simply asserting that DP holds for some implicit notion of data point.

We then consider in more detail the relationship between

TABLE I

DIFFERENTIAL PRIVACY AND VARIATIONS UPON IT. The left-most column gives the number of its definition later in the text. The point of comparison is the quantity computed for every pair of values $d_i$ and $d'_i$ for $\underline{d_i}$ to check whether the point of comparison's values for $d_i$ and for $d'_i$ are within a factor of $e^\epsilon$ of one another. The check is for all values of the index $i$. Some of the definitions only perform the comparison when the probability of the changed data point $D_i$ having the value $d_i$ (and $d'_i$, the changed value) is non-zero under $\mathcal{P}$. Others only perform the comparison when all the data points $D$ having the values $d$ (and $d'$ for changed value of $D_i$) has non-zero probability. do denotes a causal intervention instead of standard conditioning [42]. The definitions vary in whether they require performing these comparisons for just the actual probability distribution over data points $\mathcal{P}$ or over all such distributions. In one case (Definition 4), the comparison just applies to distributions where the data points are independent of one another.

| Num. | $\mathcal{P}$ | Conditions on population distribution $\mathcal{P}$ | Point of comparison (should be stable as $\underline{d_i}$ changes) | Relation |
|---|---|---|---|---|
| Original Differential Privacy | | | | |
| 1 | n/a | | $\Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, \underline{d_i}, \ldots, d_n\rangle){=}o]$ | is DP |
| Associative Variants | | | | |
| 2 | $\forall$ | $\Pr_{\mathcal{P}}[D_1{=}d_1, \ldots, D_i{=}\underline{d_i}, \ldots, D_n{=}d_n] > 0$ | $\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_1{=}d_1, \ldots, D_i{=}\underline{d_i}, \ldots, D_n{=}d_n]$ | $\leftrightarrow$ DP |
| 3 | $\forall$ | $\Pr_{\mathcal{P}}[D_i{=}d_i] > 0$ | $\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_i{=}\underline{d_i}]$ | $\rightarrow$ DP |
| 4 | $\forall$ indep. $D_i$ | $\Pr_{\mathcal{P}}[D_i{=}d_i] > 0$ | $\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_i{=}\underline{d_i}]$ | $\leftrightarrow$ DP |
| Causal Variants | | | | |
| 5 | $\forall$ | | $\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid \mathrm{do}(D_1{=}d_1, \ldots, D_i{=}\underline{d_i}, \ldots, D_n{=}d_n)]$ | $\leftrightarrow$ DP |
| 6 | given | | $\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid \mathrm{do}(D_1{=}d_1, \ldots, D_i{=}\underline{d_i}, \ldots, D_n{=}d_n)]$ | $\leftrightarrow$ DP |
| 7 | given | | $\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid \mathrm{do}(D_i{=}\underline{d_i})]$ | $\leftarrow$ DP |
| 8 | $\forall$ | | $\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid \mathrm{do}(D_i{=}\underline{d_i})]$ | $\leftrightarrow$ DP |

our work and that of Kasiviswanathan and Smith [25] (Section VII). In short, Kasiviswanathan and Smith provide a Bayesian interpretation of DP whereas we provide a complementary causal one.

As we elaborate in the conclusion (Section VIII), these results open up the possibility of using all the methods developed for working with causation to work with DP. Furthermore, it explains why researchers have found uses for DP out side of privacy (e.g., [15], [14], [12], [13], [30]): they are really trying to limit effect sizes.

## II. PRIOR WORK

The paper coining the term "differential privacy" recognized that causation is key to understanding DP: "it will not be the presence of her data that causes [the disclosure of sensitive information]" [11, p. 8]. Despite this causal view being present in the understanding of DP from the beginning, we believe we are first to make it mathematically explicit and precise, and to compare it explicitly with the associative view.

Tschantz et al. [46] reduces probabilistic noninterference (a notion of having no flow of information) to having no casual effect at all. We observe that DP with $\epsilon = 0$ is identical to noninterference, implying that the $\epsilon = 0$ case of DP could be reduced to causal effects. Our work generalizes from noninterference to DP and thereby differs in having additional bookkeeping to track the size of the effect for handling the $\epsilon > 0$ case, where an effect may be present but must be bounded. Importantly, this generalization allows us to compare the causal and associative views of DP, not a focus of [46].

Our work is largely motivated by wanting to explain the difference between two lines of research papers that have emerged from DP. The first line, associated with the inventors of DP, emphasizes differential privacy's ability to ensure that data providers are no worse off for providing data (e.g., [11], [25], [36], [35]). The second line, which formed in response to limitations in differential privacy's guarantee, emphasizes that an adversary should not be able to learn anything sensitive about the data providers from the system's outputs (e.g., [27], [28], [32], [29], [23], [5], [50], [33]). The second line notes that DP fails to provide this guarantee when the data points from different data providers are associated with one another unless one assumes that the adversary knows all but one data point. McSherry provides an informal description of the differences between the two lines [36]. While not necessary for understanding our technical development, Appendix A provides a history of the two views of DP.

Kasiviswanathan and Smith look at a different way of comparing the two views of DP, which they call *Semantic Privacy* [25]. They study the Bayesian probabilities that an adversary seeing the system's outputs would assign to a sensitive property. Whereas other works looking at an adversary's beliefs, such as Pufferfish [29], bounds the change in the adversary's probabilities before and after seeing the output, Kasiviswanathan and Smith bound the change between adversary's probabilities after seeing the output for two difference inputs, much as DP compares output distributions for two different inputs. They conclude that this posterior-to-posterior comparison captures the epistemic consequences of DP, unlike the anterior-to-posterior comparison made by Pufferfish-like definitions, since DP bounds it without additional assumptions, such as independent data points. Our causal definitions (Def. 5–8) instead expose differential privacy's causal nature with a modification of Pearl's causal framework as a frequentist effect size and we do not use any Bayesian probabilities in our causal definitions. We view their Bayesian non-causal characterization of DP as complimentary to our

frequentist causal characterization, with theirs focused on an adversary's knowledge and ours on physical constraints. (We conjecture that a Bayesian causal characterization should be possible, but leave that to future work.) Besides the conceptual difference, our characterization is tighter in that we show an exact equivalence between our central definition (Def. 8) and DP in that each implies the other with the same value of $\epsilon$, whereas their implications hold for an increased value of $\epsilon$. Section VII considers their work in more detail.

Others have explored how assumptions about the data or adversary enables alternative reductions of DP to information flow properties. Clarkson and Schneider prove an equivalence between DP and an information-theoretic notion of information suppression while making the strong adversary assumption [6, p. 32]. After making the strong adversary assumption, Cuff and Yu have argued that DP can be viewed a constraint on mutual information [8, p. 2], but McSherry points out that the connection is rather weak [37]. Alvim et al. bound the min-entropy and mutual information in terms of $\epsilon$ under assumptions about the data's distribution [1, p. 9]. Ghosh and Kleinberg provide inferential privacy bounds for DP mechanisms under assumptions about restricted background knowledge [20, p. 6]. We avoid such assumptions and our causal version of DP (Def. 8) is equivalent to the original, not merely a bound.

Instead of looking at how much an adversary learns about a single data point, Barthe and Köpf bound how much adversary learns, in terms of min entropy, about the whole database from a differentially private output, while sometimes making the strong adversary assumption [2, p. 4]. They prove that as the database increases size, the bound increases as well. McGregor et al. similarly bound the amount of information leaked, in terms of mutual information, about the whole database by a differentially private protocol (the information cost), while sometimes assuming independent data points [34, p. 14]. We focus on privacy consequences to individuals, that is, on one data point at a time.

Other papers have provided flexible or convenient associative definitions not limited to attempting to capture DP. For example, Pufferfish is a flexible framework for stating associative privacy properties [29]. Lee and Clifton explore bounding the probability that the adversary can assign to an individual being in a data set [31]. While such probabilities are more intuitive than the $\epsilon$ of DP, their central definition implicitly makes a strong adversary assumption [31, Def. 4].

## III. Differential Privacy as Association

Dwork provides a well known expression of DP [11, p. 8]. In our notation, it becomes

**Definition 1** (Differential Privacy)**.** *A randomized algorithm $\mathcal{A}$ is* $\epsilon$-differentially private *if for all $i$, for all data points $d_1, \ldots, d_n$ in $\mathcal{D}^n$ and $d'_i$ in $\mathcal{D}$, and for all output values $o$,*

$$\Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_n\rangle)=o] \leq e^{\epsilon} \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d'_i, \ldots, d_n\rangle)=o]$$

This formulation differs from Dwork's formulation in four minor ways. First, for simplicity, we restrict ourselves to only considering programs producing outputs over a finite domain, allowing us to use notationally simpler discrete probabilities. Second, we change some variable names. Third, we explicitly represent that the probabilities are over the randomization within the algorithm $\mathcal{A}$, which should be understood as physical probabilities, or *frequencies*, not as epistemic probabilities, or Bayesian *credences*. Fourth, we use the *bounded* formulation of DP, in which we presume a maximum number $n$ of individuals potentially providing data. In this formulation, it is important that one of the possible values for data points is the null data point containing no information to represent an individual deciding to not participate.

Both Dwork's expression of and our re-expression of DP make discussing the concerns about dependencies between data points raised by some papers difficult since it does not mention any distribution over data points. This omission is a reflection of the standard view that DP does not depend upon that distribution. However, to have a precise discussion of this issue, we should introduce notation for denoting the data points. We use Yang et al.'s expression of DP as a starting point [49, p. 749]:

> *Definition 4.* (Differential Privacy) A randomized mechanism $\mathcal{M}$ satisfies *$\epsilon$-differential privacy*, or $\epsilon$-DP, if
>
> $$DP(\mathcal{M}) := \sup_{i,\mathbf{x}_{-i},x_i,x'_i,S} \log \frac{\Pr(r \in S \mid x_i, \mathbf{x}_{-i})}{\Pr(r \in S \mid x'_i, \mathbf{x}_{-i})} \leq \epsilon.$$

We rewrite this definition in our notation as follows:

**Definition 2** (Strong Adversary Differential Privacy)**.** *A randomized algorithm $\mathcal{A}$ is* $\epsilon$-strong adversary differentially private *if for all population distributions $\mathcal{P}$, for all $i$, for all data points $d_1, \ldots, d_n$ in $\mathcal{D}^n$ and $d'_i$ in $\mathcal{D}$, and for all output values $o$, if*

$$\Pr_{\mathcal{P}}[D_1{=}d_1, \ldots, D_i{=}d_i, \ldots, D_n{=}d_n] > 0 \quad (1)$$

$$\textit{and} \quad \Pr_{\mathcal{P}}[D_1{=}d_1, \ldots, D_i{=}d'_i, \ldots, D_n{=}d_n] > 0 \quad (2)$$

*then*

$$\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_1{=}d_1, \ldots, D_i{=}d_i, \ldots, D_n{=}d_n]$$
$$\leq e^{\epsilon} * \Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_1{=}d_1, \ldots, D_i{=}d'_i, \ldots, D_n{=}d_n] \quad (3)$$

*where $O = \mathcal{A}(D)$ and $D = \langle D_1, \ldots, D_n\rangle$.*

This formulation differs from Yang et al.'s formulation in the following ways. As before, we change some variable names and only consider programs producing outputs over a finite domain. Also, rather than using shorthand, we write out variables explicitly and denote the distributions from which they are drawn. For example, for what they denoted as $\Pr(r \in S \mid x'_i, \mathbf{x}_{-i})$, we write $\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_1{=}d_1, \ldots, D_i{=}d'_i, \ldots, D_n{=}d_n]$, where the data points $D_1, \ldots, D_n$ are drawn from the population distribution $\mathcal{P}$ and the output $O$ uses the algorithm's internal randomization $\mathcal{A}$. This allows explicitly discussion of how the data points $D_1, \ldots, D_n$ may be correlated in the population $\mathcal{P}$ from which they come.

Finally, we explicitly deal with data points potentially having a probability of zero under $\mathcal{P}$. We ensure that we only attempt to calculate the conditional probability for databases with non-zero probability. This introduces a new problem:

if the probability distribution $\mathcal{P}$ over databases assigns zero probability to a data point value $d_i$, we will never examine the algorithm's behavior for it. While the algorithm's behavior on zero-probability events may be of little practical concern, it would allow the algorithm $\mathcal{A}$ to violate DP. (See Appendix B for an example.) To remove this possibility, we quantify over all probability distributions, which will include some with non-zero probability for every combination of data points.

Alternately, we could have used just one distribution that assigns non-zero probability to all possible input data points. We instead quantify over all distributions to make it clear that DP implies a property for all population distributions $\mathcal{P}$. While the population distribution $\mathcal{P}$ is needed to compute the probabilities used by Definition 2 and will change the probability of outcomes, whether or not $\mathcal{A}$ has DP does not actually depend upon the distribution beyond whether it assigns non-zero probability to data points. This lack of dependence explains why DP is typically defined without reference to a population distribution $\mathcal{P}$ and typically only mentions the algorithm's randomization $\mathcal{A}$.

For us, the population distribution $\mathcal{P}$ serves to link the algorithm to the data on which it is used, explaining the consequences of the algorithm for that population. Since the concerns of Yang et al. and others deal with differential privacy's behavior on populations with correlated data points, having this link proves useful. The following theorem shows that its introduction does not alter the concept.

**Proposition 1.** *Definitions 1 and 2 are equivalent.*

*Proof.* Assume Definition 1 holds. Consider any population $\mathcal{P}$, index $i$, data points $\langle d_1, \ldots, d_n\rangle$ in $\mathcal{D}^n$ and $d_i'$ in $\mathcal{D}$, and output $o$ such that the following holds: $\Pr_{\mathcal{P}}[D_1{=}d_1, \ldots, D_n{=}d_n] > 0$ and $\Pr_{\mathcal{P}}[D_1{=}d_1, \ldots, D_i{=}d_i', \ldots, D_n{=}d_n] > 0$. Since Definition 1 holds,

$$\Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_n\rangle){=}o] \leq e^{\epsilon} \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_i', \ldots, d_n\rangle){=}o]$$

Letting $O = \mathcal{A}(D)$ and $D = \langle D_1, \ldots, D_n\rangle$, the above implies

$$\begin{aligned}&\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_1{=}d_1, \ldots, D_n{=}d_n]\\ &\leq e^{\epsilon} * \Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_1{=}d_1, \ldots, D_i{=}d_i', \ldots, D_n{=}d_n]\end{aligned}$$

Thus, Definition 2 holds.

Assume Definition 2 holds. Let $\mathcal{P}$ be a population that is i.i.d. and assigns non-zero probabilities to all the sequences of $n$ data points. Consider any index $i$, data points $\langle d_1, \ldots, d_n\rangle$ in $\mathcal{D}^n$ and $d_i'$ in $\mathcal{D}$, and output $o$. $\mathcal{P}$ is such that $\Pr_{\mathcal{P}}[D_1{=}d_1, \ldots, D_n{=}d_n] > 0$ and $\Pr_{\mathcal{P}}[D_1{=}d_1, \ldots, D_i{=}d_i', \ldots, D_n{=}d_n] > 0$ both hold. Thus, since Definition 2 holds for $\mathcal{P}$,

$$\begin{aligned}&\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_1{=}d_1, \ldots, D_n{=}d_n]\\ &\leq e^{\epsilon} * \Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_1{=}d_1, \ldots, D_i{=}d_i', \ldots, D_n{=}d_n]\end{aligned}$$

where $O = \mathcal{A}(D)$ and $D = \langle D_1, \ldots, D_n\rangle$. Thus,

$$\Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_n\rangle){=}o] \leq e^{\epsilon} \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_i, \ldots, d_n\rangle){=}o]$$

Thus, Definition 1 holds. □

The standard intuition provided for the formulation of differential privacy found in Definition 2 is a Bayesian one in which we think of $\mathcal{P}$ as being prior information held by an adversary trying to learn about $D_i$. We condition upon and fix all the values of $D_1, \ldots, D_n$ except $D_i$ to model a "strong adversary" that knows every data point except $D_i$, whose value varies in (3). As the value of $D_i$ varies, we compare the probabilities of output values $o$. These probabilities can be thought of as measuring what the adversary knows about $D_i$ given all the other data points. The bigger the change in the probabilities as the value of $D_i$ varies, the bigger the flow of information from $D_i$ to $O$.

The origins of this characterization of DP go back to the original work of Dwork et al., who instead call strong adversaries "informed adversaries" [17, App. A]. However, their characterization is somewhat different than what is now viewed as the strong adversary characterization. This new characterization has since shown up in numerous places. For example, Alvim and Andrés rewrite DP this way [1, p. 5] while Yang et al. [49, Def. 4] and Cuff and Yu [8, Def. 1] even define it thus.

Despite this intuition, there's no mathematical requirement that we interpret the probabilities in terms of an adversary's Bayesian beliefs and we could instead treat them as frequencies over some population. In Section VII, we return to this issue where we explicitly mix the two interpretations. Either way, we term Definition 2 to be an *associative characterization* of DP since (3) compares probabilities that differ in the value of $D_i$ that is conditioned upon.

While it may seem intuitive that ensuring privacy against such a "strong" adversary would imply privacy against other "weaker" adversaries that know less, it turns out that the name is misleading. Suppose we measure privacy in terms of the association between $D_i$ and $O$, which captures what an adversary learns, as in (3). Depending upon the circumstances, either a more informed "stronger" adversary or a less informed "weaker" adversary will learn more from a data release [7], [27]. Intuitively, if the released data is esoteric information and only the informed adversary has enough context to make use of it, it will learn more. If, on the other hand, the released data is more basic information relating something that the informed adversary already knows but the uninformed one does not, then the "weaker" uninformed one will learn more.

One way to make this issue more precise is to model how informed an adversary is by the number of data points it knows, that is, the number conditioned upon. This leads to Yang et al.'s definition of *Bayesian Differential Privacy* [49, Def. 5]. Despite the name, its probabilities can be interpreted either as Bayesian credences or as frequencies. For simplicity, we state their definition for just the extreme case where the adversary knows zero data points:

**Definition 3** (Bayesian$_0$ Differential Privacy)**.** *A randomized algorithm $\mathcal{A}$ is* $\epsilon$-Bayesian$_0$ differentially private *if for all population distributions $\mathcal{P}$, for all $i$, for all data points $d_i$ and $d_i'$ in $\mathcal{D}$, and for all output values $o$, if* $\Pr_{\mathcal{P}}[D_i{=}d_i] > 0$ *and* $\Pr_{\mathcal{P}}[D_i{=}d_i'] > 0$ *then*

$$\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_i{=}d_i] \leq e^{\epsilon} * \Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_i{=}d_i'] \qquad (4)$$

*where $O = \mathcal{A}(D)$ and $D = \langle D_1, \ldots, D_n\rangle$.*

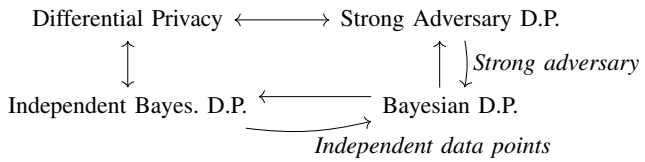



Fig. 2. Relationships between Differential Privacy and Associative Characterizations of It. Arrows show implications. Curved, labeled arrows show, in italics, assumptions required for the implication. For differential privacy to imply Bayesian Differential Privacy, one of two assumptions must be made.

One might expect that DP would provide Bayesian Differential Privacy from hearing informal descriptions of them. However, Yang et al. prove that Bayesian Differential Privacy implies DP but is strictly stronger [49, Thm. 2]. Indeed, it was already known that limiting the association between $D_i$ and the output $O$ requires limiting the associations between $D_i$ and the other data points [7], [27]. Doing so, Yang et al. proved that DP implies Bayesian Differential Privacy under the assumption that the data points are independent of one another [49, Thm. 1]. We state the resulting qualified form of DP as follows:

**Definition 4** (Independent Bayesian$_0$ Differential Privacy)**.** *A randomized algorithm $\mathcal{A}$ is* $\epsilon$-Bayesian$_0$ differentially private for independent data points *if for all population distributions $\mathcal{P}$ such that for all $i$ and $j$ where $i \neq j$, $D_i$ is independent of $D_j$ conditioned upon the other data points the following holds: for all data points $d_i$ and $d_i'$ in $\mathcal{D}$, and for all output values $o$, if* $\Pr_{\mathcal{P}}[D_i{=}d_i] > 0$ *and* $\Pr_{\mathcal{P}}[D_i{=}d_i'] > 0$ *then*

$$\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_i{=}d_i] \leq e^{\epsilon} * \Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D_i{=}d_i'] \qquad (5)$$

*where $O = \mathcal{A}(D)$ and $D = \langle D_1, \ldots, D_n \rangle$.*

On all the above math, everyone is in agreement, which we summarize in Figure 2 and below:

(a) Differential privacy and Strong Adversary Differential Privacy are equivalent,
(b) Differential privacy and Independent Bayesian Differential Privacy are equivalent,
(c) Bayesian Differential Privacy and related associative properties are strictly stronger than Differential Privacy,
(d) If we limit ourselves to strong adversaries, DP and Bayesian Differential Privacy become equivalent, and
(e) If we limit ourselves to independent data points, DP and Bayesian Differential Privacy become equivalent.

More controversially, some papers have pointed to these facts to say that DP makes implicit assumptions. Some have taken (d) to imply that DP has an implicit assumption of a strong adversary. For example, Cuff and Yu's paper states [8, p. 2]:

> The definition of $(\epsilon, \delta)$-DP involves a notion of neighboring database instances. Upon examination, one realizes that this has the effect of assuming that the adversary has already learned about all but one entry in the database and is only trying to gather additional information about the remaining entry. We refer to this as the strong adversary assumption, which is implicit in the definition of differential privacy.

Others have focused on (e) and independent data points. For example, Liu et al.'s paper asserts [33, p. 1]:

> To provide its guarantees, DP mechanisms assume that the data tuples (or records) in the database, each from a different user, are all independent.

Appendix A3 provides more examples.

Those promoting the original view of DP have re-asserted that DP was never intended to prevent all associative, or inferential, privacy threats and that doing so is impossible [3], [25], [36], [35]. However, this assertion raises the question: if DP is not providing some form of association-based inferential privacy, what is it providing?

## IV. A Primer on Causation

We believe that the right way of thinking about DP is that it is providing a causal guarantee. Before justifying this claim, we will review a framework for precisely reasoning about causation based upon Pearl's [42]. We choose Pearl's since it is the most well known in computer science, but our results can be translated into other frameworks.

To explain causation, let us return to the example of Section I-A. Suppose that the statistic being computed is the number of data points showing the genetic disease. A possible implementation of such a differentially private count algorithm $\mathcal{A}$ for a fixed number of three data points is

$$\begin{aligned}&\texttt{def } \texttt{prog}_{\mathcal{A}}(D_1, D_2, D_3):\\ &\quad D := \langle D_1, D_2, D_3 \rangle\\ &\quad O := \texttt{Lap}(1/\epsilon) + \sum_{i=1}^{3} (1 \texttt{ if } D[i] == \texttt{pos } \texttt{else } 0)\end{aligned}$$

It takes in 3 data points as inputs, representing the statuses reported by survey participants. It stores them in a database $D$ and then uses the Laplace Mechanism to provide a differentially private count of the number of data points recording the status as positive (pos) [17, Example 1].

One could use a tool like the GNU Project Debugger (GDB) to check the value of a variable as the program executes. We can think of this as making an observation. If you observed that $D[3]$ is negative (neg), you would know that $D_3$ and the third input were also neg. In a probabilistic setting, conditioning would carry out this update in knowledge.

One could also use GDB to intervene on the program's execution and alter $D[3]$ to be pos. This would probabilistically increase the output's value. But would one learn from this that $D_3$ is pos and no longer neg? No, since the program uses assignments and not equalities to shift the value of the right-hand-side variable into the left-hand-side variable. $D_3$ is a (partial) cause of $D$, but not the other way around. Altering the value of $D[3]$ only affects variables that it assigns a value to, those they assign values to, and so forth, that is, the ones it causes. In this example, that is only $O$. This reflects the difference between association and causation.

More formally, to develop a causal interpretation of DP, we start by replacing the equation $O = \mathcal{A}(D)$ with a stronger claim. Such equations say nothing about why this relation holds. We use a stronger causal relation asserting that the

value of the output $O$ is caused by the value of the input $D$, that is, we use a *structural equation*. We will denote this structural equation by $O := \mathcal{A}(D)$ since it is closer to an assignment than equality due to its directionality. To make this more precise, let $\mathrm{do}(D{=}d)$ denote an intervention setting the value of $D$ to $d$ (Pearl's *do* notation [42]). Using this notation, $\Pr[O{=}o \mid \mathrm{do}(D{=}d)]$ represents what the probability of $O = o$ would be if the value of $D$ were set to $d$ by intervention. Similar to normal conditioning on $D = d$, $\Pr[O{=}o \mid \mathrm{do}(D{=}d)]$ might not equal $\Pr[O{=}o]$. However, $\Pr[D{=}d \mid \mathrm{do}(O{=}o)]$ will surely equal $\Pr[D{=}d]$ since $O$ is downstream of $D$, and, thus, changing $O$ has no effects on $D$.

Similarly, we replace $D = \langle D_1, D_2, D_3\rangle$ with $D := \langle D_1, D_2, D_3\rangle$. That is, we consider the value of the whole database to be caused by the values of the data points and nothing more. Furthermore, we require that $D_1, D_2, D_3$ only cause $D$ and do not have any other effects. In particular, we do not allow $D_i$ to affect $D_j$ for $i \neq j$. Looking at our example program $\mathtt{prog}_{\mathcal{A}}$, this is the case.

This requirement might seem to prevent one person's attribute from affecting another's, for example, preventing a mother's genetic condition from affecting her child's genetic condition. This is not the case since $D_1, D_2, D_3$ represent the data points provided as inputs to the algorithm and not the actual attributes themselves. One could model the actual attributes, such as genetics itself, as random variables $R_1, R_2, R_3$ where $D_i := R_i$ for all $i$ and allow $R_i$ to affect $R_j$ without changing how intervening on the $D_i$s works. For example, $\mathtt{prog}_{\mathcal{A}}$ might be called in the following context:

$$
\begin{aligned}
&\mathtt{def}\ \mathtt{prog}_{\text{status}}(R_1, R_3):\\
&\quad R_2 := R_1\\
&\quad D_1 := R_1\\
&\quad D_2 := R_2\\
&\quad D_3 := R_3\\
&\quad \mathtt{prog}_{\mathcal{A}}(D_1, D_2, D_3)
\end{aligned}
$$

which does not say how the inputs $R_1$ or $R_3$ are set but does model that $R_2$ is assigned $R_1$. We can graphically represent these relationships as a *graphical model*, similar to the one in Figure 1 with $n - 2 = 1$ and an intermediate variable $D$ representing the database put between the data points and the output. Note that while $D_1$ and $D_2$ are associated, equal in fact, neither causes the other and they can be changed independently of one another, which can be seen from neither being downstream from the other.

To make the above intuitions about causation formal, we use *structural equation models* (SEMs). An SEM $\mathcal{M} = \langle \mathcal{V}_{\mathsf{en}}, \mathcal{V}_{\mathsf{ex}}, \mathcal{E}\rangle$ includes a set of *variables* partitioned into *endogenous* (or dependent) variables $\mathcal{V}_{\mathsf{en}}$ and *background* (or exogenous, or independent) variables $\mathcal{V}_{\mathsf{ex}}$. You can think of the endogenous variables as being those assigned values by the programs above and the background variables as being those provided as inputs to the programs. $\mathcal{M}$ also includes a set $\mathcal{E}$ of structural equations, corresponding to the assignments. Each endogenous variable $X$ has a structural equation $X := F_X(\vec{Y})$ where $F_X$ is a possibly randomized function and $\vec{Y}$ is a list of other variables, modeling the direct causes of $X$. To avoid circularity, $\vec{Y}$ may not include $X$. We call the variables $\vec{Y}$ the *parents* of $X$, denoted as $\mathtt{pa}(X)$.

We limit ourselves to *recursive* SEMs, those in which the variables may be ordered such that all background variables come before all endogenous variables and no variable has a parent that comes before it in the ordering. We may view such SEMs as similar to a program where the background variables are inputs to the program and the ordering determines the order of assignment statements in the program. We can make this precise by computing the values of endogenous variables from the values of the background variables using a method similar to assigning a semantics to a program.

The only difference is that, rather then a single value, the inputs are assigned probability distributions over values, which allows us to talk about the probabilities of the endogenous variables taking on a value. Let a *probabilistic SEM* $\langle \mathcal{M}, \mathcal{P}\rangle$ be an SEM $\mathcal{M}$ with a probability distribution $\mathcal{P}$ over its background variables. We can raise the structural equations (assignments) to work over $\mathcal{P}$ instead of a concrete assignment of values. (Appendix C provides details.)

Finally, to define causation, let $\mathcal{M}$ be an SEM, $Z$ be an endogenous variable of $\mathcal{M}$, and $z$ be a value that $Z$ can take on. Pearl defines the *sub-model* $\mathcal{M}[Z{:=}z]$ to be the SEM that results from replacing the equation $Z := F_Z(\vec{Z})$ in $\mathcal{E}$ of $\mathcal{M}$ with the equation $Z := z$. You can think of this as using GDB to assign a value to a variable or as aspect-oriented programming jumping into a function to alter a variable. The sub-model $\mathcal{M}[Z{:=}z]$ shows the *effect* of setting $Z$ to $z$. Let $\Pr_{\mathcal{M},\mathcal{P}}[Y{=}y \mid \mathrm{do}(Z{:=}z)]$ be $\Pr_{\mathcal{M}[Z:=z],\mathcal{P}}[Y{=}y]$. This is well defined even when $\Pr_{\mathcal{M},\mathcal{P}}[Z{=}z] = 0$ as long as $z$ is within in the range of values $\mathcal{Z}$ that $Z$ can take on.

Returning to our example, let $\mathcal{M}^{\mathcal{A}}_{\text{st}}$ be an SEM representing $\mathtt{prog}_{\text{status}}$ and $\mathcal{P}$ be the naturally occurring distribution of data points. $\Pr_{\mathcal{M}^{\mathcal{A}}_{\text{st}},\mathcal{P}}[O{=}o]$ is the probability of the algorithm's output being $o$ under $\mathcal{P}$ and coin flips internal to $\mathcal{A}$. $\Pr_{\mathcal{M}^{\mathcal{A}}_{\text{st}},\mathcal{P}}[O{=}o \mid D_i{=}\mathtt{pos}]$ is that probability conditioned upon seeing $D_1 = \mathtt{pos}$. $\Pr_{\mathcal{M}^{\mathcal{A}}_{\text{st}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{=}\mathtt{pos})]$ is that probability given an intervention setting the value of $D_1$ to $\mathtt{pos}$, which is $\Pr_{\mathcal{M}^{\mathcal{A}}_{\text{st}}[D_i:=\mathtt{pos}],\mathcal{P}}[O{=}o]$. $\mathcal{M}^{\mathcal{A}}_{\text{st}}[D_1{:=}\mathtt{pos}]$ is the program with the line assigning $R_1$ to $D_1$ replaced with $D_1 := \mathtt{pos}$. $\Pr_{\mathcal{M}^{\mathcal{A}}_{\text{st}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{=}\mathtt{pos})]$ depends upon how the intervention on $D_1$ will flow downstream to $O$.

This probability differs from the conditional probability in that setting $D_1$ to $\mathtt{pos}$ provides no information about $D_j$ for $j \neq 1$, whereas if $D_1$ and $D_j$ are associated, then seeing the value $D_1$ does provide information about $D_j$. Intuitively, this lack of information is because the artificial setting of $D_1$ to $\mathtt{pos}$ has no causal influence on $D_j$ due to the data points not affecting one another and the artificial setting, by being artificial, tells us nothing about the associations found in the naturally occurring world. On the other hand, artificially setting the attribute itself $R_1$ to $\mathtt{pos}$ will provide information about $D_2$ since $R_1$ has an effect on $D_2$ in addition to $D_1$. A second difference is that $\Pr_{\mathcal{M}^{\mathcal{A}}_{\text{st}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_i{=}d_i)]$ is defined even when $\Pr_{\mathcal{M}^{\mathcal{A}}_{\text{st}},\mathcal{P}}[D_i{=}d_i] = 0$.

Importantly, interventions on a data point $D_i$ do not model modifying the attributes they record nor affect other inputs.

Instead, interventions on $D_i$ model changing the values provided as inputs to the algorithm, which can be changed without affecting the attributes or other inputs. This corresponds to an *atomicity* property: the inputs $D_i$ are causally isolated from one another and they can be intervened upon separately.

Making the distinction between the inputs $D_i$ and the attributes $R_i$ might seem nitpicky, but it is key to understanding DP. Recall that its motivation is to make people comfortable with truthfully sharing data instead of withholding it or lying, which is an acknowledgment that the inputs people provide might not be the same as the attributes they describe. Furthermore, that changing inputs do not change attributes or other inputs is a reflection of how the program works. It is not an implicit or hidden assumption of independence; it is a fact about the program analyzed.

## V. Differential Privacy as Causation

Due to differential privacy's behavior on associated inputs and its requirement of considering zero-probability database values, DP is not a straightforward property about the independence or the degree of association of the database and the algorithm's output. The would-be conditioning upon zero-probability values corresponds to a form of counterfactual reasoning asking what the algorithm would have performed had the database taken on a particular value that it might never actually take on. Experiments with such counterfactuals, which may never naturally occur, form the core of causation. The behavior of DP on associated inputs corresponds to the atomicity property found in causal reasoning, that one can change the value of an input without changing the values of other inputs. With these motivations, we will show that DP is equivalent to a causal property that makes the change in a single data point explicit.

### A. With the Whole Database

We first show an equivalence between DP and a causal property on the whole database to echo Strong Adversary Differential Privacy (Def. 2). To draw out the parallels between the associative and causal properties, we quantify over all populations as we did in Definition 2, but as we will see, doing so is not necessary.

Let $\mathcal{M}^{\mathcal{A}}$ be an SEM modeling a slightly modified version of $\texttt{prog}_{\text{status}}$ that lacks the first assignment and treats all of any fixed number of attributes $R_i$ as inputs (i.e., as exogenous variables) with $D_i := R_i$. (Appendix C provides details.) We could instead use a version of $\mathcal{M}^{\mathcal{A}}$ that also accounts for $D_i$ possibly being assigned a value other than $R_i$ to model withholding an attribute's actual value. While the proofs would become more complex, the results would remain the same since we only intervene on the $D_i$ and not the $R_i$.

**Definition 5** (Universal Whole Database Intervention D.P.)**.** *A randomized algorithm $\mathcal{A}$ is* $\epsilon$-differentially private as universal intervention on the whole database *if for all population distributions $\mathcal{P}$, for all $i$, for all data points $d_1, \ldots, d_n$ in $\mathcal{D}^n$ and $d'_i$ in $\mathcal{D}$, and for all output values $o$,*

$$\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{=}d_1, \ldots, D_n{=}d_n)] \leq e^{\epsilon} * \Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{=}d_1, \ldots, D_i{=}d'_i, \ldots, D_n{=}d_n)]$$

*where $O := \mathcal{A}(D)$ and $D := \langle D_1, \ldots, D_n \rangle$.*

**Proposition 2.** *Definitions 1 and 5 are equivalent.*

*Proof.* Pearl's Property 1 says that conditioning upon all the parents of a variable and causally intervening upon them all yields the same probability [42, p. 24]. Intuitively, this is for the same reason that Strong Adversary Differential Privacy is equivalent to DP: it blocks other paths of influence from one data point to the output via another data point by fixing all the data points.

We can apply Property 1 since all the $D_i$s are being intervened upon and they make up all the parents of $D$. We can apply it again on $D$ and $O$. We then get that $\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{=}d_1, \ldots, D_n{=}d_n)]$ is equal to $\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid D_1{=}d_1, \ldots, D_n{=}d_n]$, that is to Strong Adversary Differential Privacy, which we already know to be equivalent to DP by Proposition 1. □

Notice that this causal property is simpler than the associative one in that it does not need qualifications around zero-probability data points because we can causally fix data points to values with zero probability. In fact, the population distribution $\mathcal{P}$ did not matter at all since intervening upon all the data points makes it irrelevant, intuitively by overwriting it. For this reason, we could instead look at any population, such as the naturally occurring one (or even elide it from the definition altogether, as in Definition 1, if we are not too picky about formalism). Next, we state such a simplified definition.

**Definition 6** (Whole Database Intervention D.P.)**.** *Given a population distribution $\mathcal{P}$, a randomized algorithm $\mathcal{A}$ is* $\epsilon$-differentially private as intervention on the whole database for $\mathcal{P}$ *if for all $i$, for all data points $d_1, \ldots, d_n$ in $\mathcal{D}^n$ and $d'_i$ in $\mathcal{D}$, and for all output values $o$,*

$$\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{=}d_1, \ldots, D_n{=}d_n)] \leq e^{\epsilon} * \Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{=}d_1, \ldots, D_i{=}d'_i, \ldots, D_n{=}d_n)]$$

*where $O := \mathcal{A}(D)$ and $D := \langle D_1, \ldots, D_n \rangle$.*

**Proposition 3.** Definitions 1 and 6 are equivalent.

*Proof.* The proof follows in the same manner as Proposition 2 since that proof applies to all population distributions $\mathcal{P}$. □

### B. With a Single Data Point

Definitions 5 and 6, by fixing every data point, do not capture the local nature of the decision facing a single potential survey participant. We can define a notion similar to DP that uses a causal intervention on a single data point as follows:

**Definition 7** (Data-point Intervention D.P.)**.** *Given a population $\mathcal{P}$, a randomized algorithm $\mathcal{A}$ is* $\epsilon$-differentially private as intervention on a data point for $\mathcal{P}$ *if for all $i$, for all data points $d_i$ and $d'_i$ in $\mathcal{D}$, and for all output values $o$,*

$$\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_i{=}d_i)] \leq e^{\epsilon} \Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_i{=}d'_i)]$$

*where $O := \mathcal{A}(D)$ and $D := \langle D_1, \ldots, D_n \rangle$.*

This definition is strictly weaker than DP. The reason is similar to why we had to quantify over all distributions $\mathcal{P}$

with Strong Adversary Differential Privacy. In both cases, we can give a counterexample with a population $\mathcal{P}$ that hides the effects of a possible value of the data point by assigning the value a probability of zero. For the associative definition, the counterexample involves only a single data point (Appendix B). However, for this causal definition, the counterexample has to have two data points. The reason is that, since the do operation acts on a single data point at a time, it can flush out the effects of a single zero-probability value but not the interactions between two zero-probability values.

**Proposition 4.** *Definition 1 implies Definition 7, but not the other way around.*

*Proof.* W.l.o.g., assume $i = n$. Assume Definition 1 holds:

$$\Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d_n\rangle){=}o] \leq e^{\epsilon} * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d'_n\rangle){=}o]$$

for all $o$ in $\mathcal{O}$, $\langle d_1, \ldots, d_n\rangle$ in $\mathcal{D}^n$, and $d'_n$ in $\mathcal{D}$. This implies that for any $\mathcal{P}$,

$$\Pr_{\mathcal{P}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d_n\rangle){=}o] \leq e^{\epsilon} * \Pr_{\mathcal{P}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d'_n\rangle){=}o]$$

for all $o$ in $\mathcal{O}$, $d_1, \ldots, d_n$ in $\mathcal{D}^n$, and $d'_n$ in $\mathcal{D}$. Thus,

$$\sum_{\langle d_1,\ldots,d_{n-1}\rangle \in \mathcal{D}^{n-1}} \Pr_{\mathcal{P}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d_n\rangle){=}o] \leq \sum_{\langle d_1,\ldots,d_{n-1}\rangle \in \mathcal{D}^{n-1}} e^{\epsilon} * \Pr_{\mathcal{P}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d'_n\rangle){=}o]$$

$$\sum_{\langle d_1,\ldots,d_{n-1}\rangle \in \mathcal{D}^{n-1}} \Pr_{\mathcal{P}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d_n\rangle){=}o] \leq e^{\epsilon} \sum_{\langle d_1,\ldots,d_{n-1}\rangle \in \mathcal{D}^{n-1}} \Pr_{\mathcal{P}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d'_n\rangle){=}o]$$

$$\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_n{=}d_n)] \leq e^{\epsilon} * \Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_n{=}d'_n)]$$

where the last line follows from Lemma 2 in Appendix C.

Definition 7 is, however, weaker than DP. Consider the case of a database holding two data points whose value could be 0, 1, or 2. Suppose the population $\mathcal{P}$ is such that $\Pr_{\mathcal{P}}[D_1{=}2] = 0$ and $\Pr_{\mathcal{P}}[D_2{=}2] = 0$. Consider an algorithm $\mathcal{A}$ such that

$$\Pr_{\mathcal{A}}[\mathcal{A}(\langle 2, 2\rangle){=}0] = 1 \qquad \Pr_{\mathcal{A}}[\mathcal{A}(\langle 2, 2\rangle){=}1] = 0$$
$$\Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, d_2\rangle){=}0] = 1/2 \quad \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, d_2\rangle){=}1] = 1/2$$

when $d_1 \neq 2$ or $d_2 \neq 2$. The algorithm does not satisfy Definition 1 due to its behavior when both of the inputs are 2. However, using Lemma 2 in Appendix C,

$$\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{=}d'_1)] = 1/2$$

for all $o$ and $d'_1$ since $\Pr_{\mathcal{P}}[D_2{=}2] = 0$. A similar result holds switching the roles of $D_1$ and $D_2$. Thus, the algorithm satisfies Definition 7 for $\mathcal{P}$ but not Definition 1. □

Despite being only implied by, not equivalent to, DP, Definition 7 captures the intuition behind the characterization $(*)$ of DP that "changing a single individual's data in the database leads to a small change in the distribution on outputs" [25, p. 2]. To get an equivalence, we can quantify over all populations as we did to get an equivalence for association, but this time we need not worry about zero-probability data points or independence. This simplifies the definition and makes it a more natural characterization of DP.

**Definition 8** (Universal Data-point Intervention D.P.)**.** *A randomized algorithm $\mathcal{A}$ is $\epsilon$-*differentially private as universal intervention on a data point *if for all population distributions $\mathcal{P}$, for all $i$, for all data points $d_i$ and $d'_i$ in $\mathcal{D}$, and for all output values $o$,*

$$\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_i{=}d_i)] \leq e^{\epsilon} \Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_i{=}d'_i)]$$

*where $O := \mathcal{A}(D)$ and $D := \langle D_1, \ldots, D_n\rangle$.*

**Proposition 5.** *Definitions 1 and 8 are equivalent.*

*Proof.* That Definition 1 implies 8 follows from Proposition 4.

Assume Definition 8 holds. W.l.o.g., assume $i = n$. Then, for all $\mathcal{P}$, $o$ in $\mathcal{O}$, and $d'_n$ in $\mathcal{D}$,

$$\begin{aligned}
&\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_i{=}d_i)] \leq e^{\epsilon} * \Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_i{=}d'_i)] \\
&\sum_{\langle d_1,\ldots,d_{n-1}\rangle \in \mathcal{D}^{n-1}} \Pr_{\mathcal{P}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d_n\rangle){=}o] \\
&\leq e^{\epsilon} \sum_{\langle d_1,\ldots,d_{n-1}\rangle \in \mathcal{D}^{n-1}} \Pr_{\mathcal{P}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d'_n\rangle){=}o]
\end{aligned} \tag{6}$$

follows from Lemma 2 in Appendix C.

For any $d^{\dagger}_1, \ldots, d^{\dagger}_{n-1}$ in $\mathcal{D}^{n-1}$, let $\mathcal{P}^{d^{\dagger}_1,\ldots,d^{\dagger}_{n-1}}$ be such that

$$\Pr_{\mathcal{P}^{d^{\dagger}_1,\ldots,d^{\dagger}_{n-1}}}\left[\wedge_{i=1}^{n-1} D_i{=}d^{\dagger}_i\right] = 1 \tag{7}$$

For any $d^{\dagger}_1, \ldots, d^{\dagger}_n$ in $\mathcal{D}^n$ and $d'_n$ in $\mathcal{D}$, (6) implies

$$\begin{aligned}
&\sum_{\langle d_1,\ldots,d_{n-1}\rangle \in \mathcal{D}^{n-1}} \Pr_{\mathcal{P}^{d^{\dagger}_1,\ldots,d^{\dagger}_{n-1}}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d^{\dagger}_n\rangle){=}o] \\
&\leq e^{\epsilon} \sum_{\langle d_1,\ldots,d_{n-1}\rangle \in \mathcal{D}^{n-1}} \Pr_{\mathcal{P}^{d^{\dagger}_1,\ldots,d^{\dagger}_{n-1}}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1, \ldots, d_{n-1}, d'_n\rangle){=}o]
\end{aligned}$$

Thus,

$$\Pr_{\mathcal{A}}[\mathcal{A}(\langle d^{\dagger}_1, \ldots, d^{\dagger}_{n-1}, d^{\dagger}_n\rangle){=}o] \leq e^{\epsilon} \Pr_{\mathcal{A}}[\mathcal{A}(\langle d^{\dagger}_1, \ldots, d^{\dagger}_{n-1}, d'_n\rangle){=}o]$$

since both sides has a non-zero probability for

$$\Pr_{\mathcal{P}^{d^{\dagger}_1,\ldots,d^{\dagger}_{n-1}}}\left[\wedge_{i=1}^{n-1} D_i{=}d_i\right]$$

at only the sequence of data point values $d^{\dagger}_1, \ldots, d^{\dagger}_{n-1}$. □

## VI. Bounding Effects: Generalizing D.P., Understanding Alternatives

To recap, we have shown that reasoning about DP as a causal property is more straightforward than reasoning about it as an associative property. Still, one might wonder, Why express DP in either form? Why not just stick with its even simpler expression in terms of functions in Definition 1?

In this section, we show what is gained by the causal view. We show that DP bounds a general notion of *effect size*. Essentially, DP limits the causal consequences of a decision to contribute data to a data set. If the consequences are small, then an individual will need less encouragement (e.g., financial incentives) to set aside privacy concerns.

We show that this general notion can also capture alternative privacy definitions, including some arising from concerns over dependent data points. A common causal framework allows us to precisely compare these definitions.

### A. Bounded Relative Probability (BRP)

Generalizing from the decision to participate in a data set, we define a more general notation for any two random variables $X$ and $Y$. To do so, we need a description of how $X$ and $Y$ relate to one another. Recall that a probabilistic SEM $\langle \mathcal{M}, \mathcal{P} \rangle$ shows the causal and statistical relations between random variables by providing a list of structural equations $\mathcal{M}$ and a distribution $\mathcal{P}$ over variables not defined in terms of others (exogenous variables). (See Appendix C for details.)

We will measure the size of the effects of $X$ on $Y$ using *relative probabilities*, better known as *relative risk* and as *risk ratio* with clinical studies of risks in mind. For three (binary) propositions $\rho$, $\phi$, and $\psi$, let

$$\mathrm{RP}_{\mathcal{M},\mathcal{P}}(\rho, \phi, \psi) = \frac{\Pr_{\mathcal{M},\mathcal{P}}[\rho \mid \mathrm{do}(\phi)]}{\Pr_{\mathcal{M},\mathcal{P}}[\rho \mid \mathrm{do}(\psi)]}$$

denote the relative probability. (Some authors also allow using conditioning instead of interventions.) For two random variables $X$ and $Y$, we can characterize the maximum effect of $X$ on $Y$ as

$$\bar{RP}_{\mathcal{M},\mathcal{P}}(Y, X) = \max_{y, x_1, x_2} \mathrm{RP}_{\mathcal{M},\mathcal{P}}(Y{=}y, X{=}x_1, X{=}x_2)$$

Expanding these definitions out shows that $\epsilon$-differential privacy places a bound on the maximum of the maximum relative probabilities:

$$\max_{\mathcal{P}, i} \bar{RP}_{\mathcal{M},\mathcal{P}}(O, D_i) \leq e^{\epsilon}$$

where $\mathcal{M}$ describes the differentially private algorithm $\mathcal{A}$. Note that our use of maximization is similar Yang et al. [49, p. 749, Def. 4], which we quote in Section III.

With this in mind, we propose to use $\bar{RP}$ for a general purpose effect-size restriction:

**Definition 9** (BRP)**.** A causal system described by $\mathcal{M}$ has *$\epsilon$-bounded relative probability (*BRP*) for $X$ to $Y$* iff

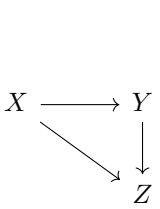
$$\max_{\mathcal{P}} \bar{RP}_{\mathcal{M},\mathcal{P}}(Y, X) \leq e^{\epsilon}$$

Differential privacy is equivalent to requiring $\epsilon$-BRP for all data points $D_i$.

### B. Composition

BRP enjoys many of the same properties as DP. Recall that DP has additive sequential composition for using two differentially private algorithms one after the next, even if the second is selected using the output of the first [38]. Similarly, BRP has additive sequential composition for two random variables.

To model the second output $Z$ depending upon the first $Y$, but not the other way around, we say random variables $X$, $Y$, and $Z$ are *in sequence* if $X$ may affect $Y$ and $Z$, and $Y$ may affect $Z$, but $Z$ may not affect $X$ nor $Y$, and $Y$ may not affect $X$. That is,




To model that the second output $Z$ could be computed with one of any of a set of algorithms but that each of algorithm has a bounded effect from $X$ to $Z$, we look at $Z$'s behavior in sub-models $\mathcal{M}[Y := y]$ where each setting of $Y$ corresponds to a selecting one available algorithm.

**Theorem 1.** For any SEM $\mathcal{M}$ such that $X$, $Y$, and $Z$ are in sequence and the parents of $Z$ are $\{X, Y\}$, if $X$ has $\epsilon_1$-BRP to $Y$ in $\mathcal{M}$ and $\epsilon_2$-BRP to $Z$ in $\mathcal{M}[Y := y]$ for all $y$ in $\mathcal{Y}$, then $X$ has $(\epsilon_1 + \epsilon_2)$-BRP to $\langle Y, Z \rangle$ in $\mathcal{M}$.

*Proof.* Consider any probability distribution $\mathcal{P}$, $x$ and $x'$ in $\mathcal{X}$, $y$ in $\mathcal{Y}$, and $z$ in $\mathcal{Z}$. Since the effect of $X$ on $Y$ is bounded by $\epsilon_1$-BRP,

$$\Pr_{\mathcal{M},\mathcal{P}}[Y{=}y \mid \mathrm{do}(X{=}x)] \leq e^{\epsilon_1} \Pr_{\mathcal{M},\mathcal{P}}[Y{=}y \mid \mathrm{do}(X{=}x')]$$

Since the parents of $Z$ are $\{X, Y\}$, Pearl's Property 1 [42, p. 24] shows that for any $y$ such that $\Pr_{\mathcal{M},\mathcal{P}}[Y{=}y] > 0$,

$$\begin{aligned}&\Pr_{\mathcal{M},\mathcal{P}}[Z{=}z \mid Y{=}y, \mathrm{do}(X{=}x)] \\ &\qquad = \Pr_{\mathcal{M},\mathcal{P}}[Z{=}z \mid \mathrm{do}(Y{=}y), \mathrm{do}(X{=}x)]\end{aligned}$$

Since there's $\epsilon_2$-BRP from $X$ to $Z$ in $M[Y := y]$ for all $y$, this implies that

$$\begin{aligned}&\Pr_{\mathcal{M},\mathcal{P}}[Z{=}z \mid Y{=}y, \mathrm{do}(X{=}x)] \\ &\qquad \leq e^{\epsilon_2} \Pr_{\mathcal{M},\mathcal{P}}[Z{=}z \mid Y{=}y, \mathrm{do}(X{=}x')]\end{aligned}$$

Thus,

$$\begin{aligned}&\Pr_{\mathcal{M},\mathcal{P}}[\langle Y, Z \rangle = \langle y, z \rangle \mid \mathrm{do}(X{=}x)] \\ &= \Pr_{\mathcal{M},\mathcal{P}}[Z{=}z \mid Y{=}y, \mathrm{do}(X{=}x)] \Pr_{\mathcal{M},\mathcal{P}}[Y{=}y \mid \mathrm{do}(X{=}x)] \\ &\leq e^{\epsilon_2} \Pr_{\mathcal{M},\mathcal{P}}[Z{=}z \mid Y{=}y, \mathrm{do}(X{=}x')] \\ &\qquad * e^{\epsilon_1} \Pr_{\mathcal{M},\mathcal{P}}[Y{=}y \mid \mathrm{do}(X{=}x')] \\ &= e^{\epsilon_1 + \epsilon_2} \Pr_{\mathcal{M},\mathcal{P}}[\langle Y, Z \rangle = \langle y, z \rangle \mid \mathrm{do}(X{=}x')]\end{aligned}$$

□

We can generalize this theorem for $Z$ having additional parents by requiring BRP for all of their values as well.

The special case of this theorem where $\epsilon_2 = 0$ is known as the *postprocessing* condition:

$$X \xrightarrow{\epsilon} Y \longrightarrow Z$$

For this causal diagram, Theorem 1 ensures that if the arrow from $X$ to $Y$ is $\epsilon$-BRP, then any subsequent consequence $Z$ of $Y$ is also going to be $\epsilon$-BRP. This captures the central intuition behind DP and BRP that they limit any downstream causal consequences of a variable $X$.

### C. Application

While the explicit causal reasoning in BRP can sharpen our intuitions about privacy, BRP is not itself a privacy definition. Only some choices of variables to bound yield reasonable privacy guarantees. Below, we use BRP to express some of well known variations of DP. Doing so both shows some reasonable ways of using BRP to provide privacy guarantees

and demonstrates that BRP provides a common framework for precisely stating and comparing these variations.

First, consider the randomized response method of providing privacy in which each survey participant adds noise to his own response before responding [48]. Let each person's actual attribute be $R_i$, let the noisy response he provides be $D_i$, and let $O$ be the output computed from all the $D_i$. Unlike with (standard) DP, the causal path from $D_i$ to $O$ has unbounded BRP, may not contain any random algorithms, and misses the privacy protection altogether. Similarly, the path from $R_i$ and $O$ has unbounded BRP due to the possibility of the $R_i$ having effects upon one another. However, the randomized response method does ensure $\epsilon$-BRP from $R_i$ to $D_i$ for all $i$ where $\epsilon$ depends upon the amount of noise added to $D_i$.

Second, we consider *group privacy*, the idea that a group of individuals may be so closely related that their privacy is intertwined. Differential privacy approaches group privacy by summing the privacy losses, measured in terms of $\epsilon$, of each individual in the group [11, p. 9]. Similarly, we can add the relative probabilities of multiple random variables to get a total effect size. Alternately, BRP can easily be extended to measure simultaneous joint interventions by using multiple instances of the $\mathrm{do}$ operator. The total effect size may be larger than the joint effect size since, in cases where the intervened upon variables affect one another, interventions on a downstream variable can mask interventions on its parents. Returning to the example of Section I-A, the total effect for both Ada's attribute $R_1$ and Byron's $R_2$ is $3\epsilon$ with $2\epsilon$ of that coming from $R_1$. However, the joint effect is $2\epsilon$ since $R_1$ achieved half of its effect via $R_2$. In examples like this where the variables correspond to different moral entities, the total effect size strikes us as more reasonable since it accounts for both Ada and Byron experiencing a privacy loss. If on the other hand, the variables correspond to a single topic about a single person, such as weight and waist size, then the joint effect size seems more reasonable. However, we see this choice as under explored since it does not emerge for DP given that data points cannot not affect one another.

Third, we consider a line of papers providing definitions of privacy that account for dependencies between data points, but which are ambiguous about association versus causation [5], [50], [33]. For example, Liu et al. use the word "cause" in a central definition of their work [33, Def. 3], but do no causal modeling, instead using a joint probability distribution to model just associations in their adversary model [33, §3]. Using causal modeling and BRP would allow them to actually model causation instead of approximating it with associations, or, if associations really is what they wish to model, would provide a foil making their goals more clear.

Fourth, as a more complex example, Kifer and Machanavajjhala consider applying DP to social networks [27, §3]. They note that DP applied to a network is typically taken to mean either considering nodes or edges labeled with an individual's id $i$ in the network as that individual's data point $D_i$, but that participation in a social network is likely to leave far more evidence than just those nodes and edges. They consider an example in which Bob joins a network and introduces Alice and Charlie to one another, leading them to create an edge between them that does not involve Bob. Arguably, protecting Bob's privacy requires counting this edge as Bob's as well despite neither edge nor node DP doing so. To capture this requirement, they distinguish between differential privacy's deleting of data points from a data set and their desire to "hide the *evidence of participation*" [27, §2.2.1].

Because "It is difficult to formulate a general and formal definition for what evidence of participation means" [27, §3, p. 5], they use correlations in its place for modeling public health and census records [27, §§2.1.3, 2.2, 4.1 & 4.3.1]. However, for modeling social networks, they use statistical models that they interpret as providing "a measure of the causal influence of Bob's edge", that is, informal causal models [27, §3, p. 6].

We believe that the causal framework presented herein provides the necessary mathematical tools to precisely reason about evidence of participation. Causal models would allow them to precisely state which aspects of the system they wish to protect, for example, by requiring that Bob's joining the network should have a bounded effect upon a data release. While accurately modeling a social process is a difficult task, at least the requirement is clearly stated, allowing us to return to empirical work. Furthermore, such formalism can allow for multiple models to be considered and we can demand privacy under each of them, and erring on the side of safety by overestimating effect sizes remains an option.

Finally, causal modeling can make the choices between privacy notions more clear. The distinction between *direct* and *indirect* effects [41] can model the difference between node privacy, which only captures the direct effects of joining a social network, and all of the evidence of participation, which includes hard-to-model indirect effects. Edge privacy captures the direct effect of posting additional content. Given that Facebook has reached near universal membership but worries about disengagement, this effect might be the more concerning one from paractical perspective.

## VII. Restrictions on Knowledge

Privacy is often thought of as preventing an adversary from learning sensitive information. To make this intuition precise, we can model an adversary's beliefs using Bayesian probabilities, or *credences*. We denote them with $\mathrm{Cr}$, instead of $\Pr$, which we have been using to denote natural frequencies over outcomes without regard to any agent's beliefs. We denote the adversary's background knowledge as $B$. The knowledge of an adversary about the database $D$ after observing the output can be expressed as $\mathrm{Cr}[D{=}d \mid O{=}o, B]$. A natural privacy property, termed *statistical nondisclosure* by Dalenius [9], requires that $\mathrm{Cr}[D{=}d \mid O{=}o, B] = \mathrm{Cr}[D{=}d \mid B]$, that is, that the beliefs about the database before and after observing the output are the same.

This requirement limiting the difference between prior and posterior beliefs has been shown to be impossible to achieve under arbitrary background knowledge by Dwork and Naor, even for approximate relaxations of statistical nondisclosure, as long as the output provides some information [18]. As DP also falls under the purview of this impossibility result,

it only provides this associative guarantee under restrictive background knowledge assumptions, such as independent data points or strong adversaries. To see the need for assumptions, consider that statistical nondisclosure implies 0-Bayesian$_0$ Differential Privacy (Def. 3) since both are equivalent to requiring independence between $D$ and $O$ in the case where the adversary's background information is the true distribution over data points. We believe such a need underlies the view that DP only works with assumptions (Appendix A3).

Kasiviswanathan and Smith's Semantic Privacy is a property about the adversary's ability to do inferences that does not require such assumptions [25]. It requires that the probability that the adversary assigns to the input data points does not change much whether an individual $i$ provides data or not. The probability assigned by the adversary when each person provides his data point is

$$\Pr[D{=}d \mid O{=}o, B] = \frac{\Pr_{\mathcal{A}}[\mathcal{A}(d){=}o] * \Pr[D{=}d \mid B]}{\sum_{d'} \Pr_{\mathcal{A}}[\mathcal{A}(d'){=}o] * \Pr[D{=}d' \mid B]}$$

where $D{=}d$ is shorthand for $\bigwedge_{j=1}^{n} D_j{=}d_j$ with $d = \langle d_1, \ldots, d_n \rangle$. The probability where person $i$ does not provide data or provides fake data is

$$\frac{\Pr_{\mathcal{A}}[\mathcal{A}(d_{-i}d_i'){=}o] * \Pr[D{=}d \mid B]}{\sum_{d'} \Pr_{\mathcal{A}}[\mathcal{A}(d_{-i}d_i'){=}o] * \Pr[D{=}d' \mid B]}$$

where $d_i'$ is the value (possibly the null value) provided instead of the real value and $d_{-1}d_i'$ is shorthand for $d$ with its $i$th component replaced with $d_i'$. While we leave fully formalizing the combining of Bayesian credences and frequentist probabilities to future work, intuitively, this probability is $\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[D{=}d \mid O{=}o, \mathrm{do}(D_i{=}d_i'), B]$ in our causal notation.

Kasiviswanathan and Smith prove that DP and Semantic Privacy are closely related [25, Thm. 2.2]. In essence, they show that DP ensures that $\Pr[D{=}d \mid O{=}o, B]$ and $\Pr[D{=}d \mid O{=}o, \mathrm{do}(D_i{=}d_i'), B]$ are close in nearly the same sense as it ensures that $\Pr[O{=}o]$ and $\Pr[O{=}o \mid \mathrm{do}(D_i{=}d_i')]$ are close. That is, it guarantees that an adversary's beliefs will not change much relative to whether you decide to provide data or not, providing an inference-based view of DP.

To gain intuition about these results, let us consider the findings of Wang and Kosinski [47], which show the possibility of training a neural network to predict a person's sexual orientation from a photo of their face. If this model had been produced with DP, then each study participant would know that their participation had little to do with the model's final form or success. However, inferential threats would remain. An adversary can use the model and a photo of an individual to infer the individual's sexual orientation, whether that individual participated in the study or not. Less obviously, an adversary might have some background knowledge allowing it to repurpose the model to predict people's risks of certain health conditions. Such difficult to predict associations may already be used for marketing [24] (cf. [43]).

An individual facing the option of participating in such a study may attempt to reason about how likely such repurposing is. Doing so requires the difficult task of characterizing the adversary's background knowledge since Dwork and Naor's proof shows that the possibility cannot be categorically eliminated [18]. Furthermore, if the individual decides that the study is too risky, merely declining to participate will do little to mitigate the risk since DP ensures that the individual's data would have had little effect on the model. Rather, the truly concerned individual would have to lobby others to not participate. For this reason, both the causal and associative views of privacy have their uses, with the causal view being relevant to a single potential participant's choice and the associative, to the participants collectively. One can debate whether such collective properties are privacy per se or some other value since it goes beyond protecting personal data [36].

## VIII. Conclusion and Further Implications

Although it is possible to view DP as an associative property with an independence assumption, we have shown that it is cleaner to view DP as a causal property without such an assumption. We believe that this difference in goals helps to explain why one line of research claims that DP requires an assumption of independence while another line denies it: the assumption is not required but does yield stronger conclusions.

We believe these results have implications beyond explaining the differences between these two lines. Having shown a precise sense in which DP is a causal property, we can use the results of statistics, experimental design, and science about causation while studying DP. For example, various papers have attempted to reverse engineer or test whether a system has DP [45], [10], [4]. Authors of follow up works may leverage by pre-existing experimental methods and statistical analyses for measuring effect sizes that apply with or without access to causal models.

In the opposite direction, the natural sciences can use DP as an effect-size metric, inheriting all the pleasing properties known of DP. For example, DP composes cleanly with itself, both in sequence and in parallel [39]. The same results would also apply to the effect-size metric that DP suggests.

Finally, showing that DP is in essence a measure of effect sizes explains why it, or properties based upon it, has shown up in areas other than privacy, including fairness [15], ensuring statistical validity [14], [12], [13], and adversarial machine learning [30]. While it may be surprising that privacy is related to such a diverse set of areas, it is not surprising that causation is, given the central role the concept plays in science. What is actually happening is that causal reasoning is making its importance felt in each of these areas, including in privacy. That it has implicitly shown up in at least four areas of research suggests that causal reasoning should play a more explicit role in computer science.

*Acknowledgements:* We thank Arthur Azevedo de Amorim, Deepak Garg, Ashwin Machanavajjhala, and Frank McSherry for comments on this work. We received funding from the NSF (Grants 1514509, 1704845, and 1704985) and DARPA (FA8750-16-2-0287). The opinions in this paper are those of the authors and do not necessarily reflect the opinions of any funding sponsor or the United States Government.

## REFERENCES


[1] M. Alvim, M. Andrés, K. Chatzikokolakis, and C. Palamidessi, "On the relation between differential privacy and quantitative information flow," in *38th Intl. Colloquium on Automata, Languages and Programming – ICALP 2011*, ser. LICS, vol. 6756, 2011, pp. 60–76.

[2] G. Barthe and B. Kopf, "Information-theoretic bounds for differentially private mechanisms," in *2011 IEEE 24th Computer Security Foundations Symp.*, ser. CSF '11, 2011, pp. 191–204.

[3] R. Bassily, A. Groce, J. Katz, and A. Smith, "Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy," in *2013 IEEE 54th Annual Symp. on Foundations of Computer Science*, 2013, pp. 439–448.

[4] B. Bichsel, T. Gehr, D. Drachsler-Cohen, P. Tsankov, and M. Vechev, "DP-Finder: Finding differential privacy violations by sampling and optimization," in *2018 ACM SIGSAC Conf. on Computer and Communications Security (CCS '18)*, 2018, pp. 508–524.

[5] R. Chen, B. C. Fung, P. S. Yu, and B. C. Desai, "Correlated network data publication via differential privacy," *The VLDB Journal*, vol. 23, no. 4, pp. 653–676, 2014.

[6] M. R. Clarkson and F. B. Schneider, "Quantification of integrity," *Mathematical Structures in Computer Science*, vol. 25, no. 2, pp. 207–258, 2015.

[7] G. Cormode, "Personal privacy vs population privacy: Learning to attack anonymization," in *17th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2011, pp. 1253–1261.

[8] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *2016 ACM SIGSAC Conf. on Computer and Communications Security*, ser. CCS '16, 2016, pp. 43–54.

[9] T. Dalenius, "Towards a methodology for statistical disclosure control," *Statistik Tidskrift*, vol. 15, pp. 429–444, 1977.

[10] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *2018 ACM SIGSAC Conf. on Computer and Communications Security (CCS '18)*, 2018, pp. 475–489.

[11] C. Dwork, "Differential privacy," in *Automata, Languages and Programming, 33rd Intl. Colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II*, ser. LICS, vol. 4052, 2006, pp. 1–12.

[12] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *28th Intl. Conf. on Neural Information Processing Systems - Volume 2 (NIPS'15)*, 2015, pp. 2350–2358.

[13] ——, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015. [Online]. Available: http://science.sciencemag.org/content/349/6248/636

[14] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, "Preserving statistical validity in adaptive data analysis," in *Forty-seventh Annual ACM Symp. on Theory of Computing (STOC '15)*, 2015, pp. 117–126.

[15] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *3rd Innovations in Theoretical Computer Science Conf.*, 2012, pp. 214–226.

[16] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *24th Annual Intl. Conf. on The Theory and Applications of Cryptographic Techniques*, 2006, pp. 486–503.

[17] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conf.*, 2006, pp. 265–284.

[18] C. Dwork and M. Naor, "On the difficulties of disclosure prevention in statistical databases or the case for differential privacy," *J. Privacy and Confidentiality*, vol. 2, no. 1, pp. 93–107, 2008.

[19] J. Gehrke, E. Lui, and R. Pass, "Towards privacy for social networks: A zero-knowledge based definition of privacy," in *8th Conf. on Theory of Cryptography*, 2011, pp. 432–449.

[20] A. Ghosh and R. Kleinberg, "Inferential privacy guarantees for differentially private mechanisms," *CoRR*, vol. abs/1603.01508v2, 2017, presented at the 8th Innovations in Theoretical Computer Science conference in 2017.

[21] S. Goldwasser and S. Micali, "Probabilistic encryption & how to play mental poker keeping secret all partial information," in *Fourteenth Annual ACM Symp. on Theory of Computing*, ser. STOC '82, 1982, pp. 365–377.

[22] ——, "Probabilistic encryption," *J. Computer and System Sciences*, vol. 28, no. 2, pp. 270–299, 1984.

[23] X. He, A. Machanavajjhala, and B. Ding, "Blowfish privacy: Tuning privacy-utility trade-offs using policies," in *ACM SIGMOD Intl. Conf. on Management of Data (SIGMOD 2014)*, 2014.

[24] K. Hill, "How Target figured out a teen girl was pregnant before her father did," *Forbes*, 2012.

[25] S. P. Kasiviswanathan and A. Smith, "On the 'semantics' of differential privacy: A Bayesian formulation," *J. Privacy and Confidentiality*, vol. 6, no. 1, pp. 1–16, 2014.

[26] S. P. Kasiviswanathan and A. D. Smith, "A note on differential privacy: Defining resistance to arbitrary side information," *ArXiv*, vol. 0803.3946, 2008.

[27] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *2011 ACM SIGMOD Intl. Conf. on Management of data*. ACM, 2011, pp. 193–204.

[28] ——, "A rigorous and customizable framework for privacy," in *31st ACM SIGMOD-SIGACT-SIGAI Symp. on Principles of Database Systems*, 2012, pp. 77–88.

[29] ——, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, pp. 3:1–3:36, 2014.

[30] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," *ArXiv*, vol. 1802.03471v3, 2018, to appear at IEEE S&P 2019.

[31] J. Lee and C. Clifton, "Differential identifiability," in *18th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2012, pp. 1041–1049.

[32] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang, "Membership privacy: A unifying framework for privacy definitions," in *2013 ACM SIGSAC Conf. on Computer and Communications Security*, ser. CCS '13, 2013, pp. 889–900.

[33] C. Liu, S. Chakraborty, and P. Mittal, "Dependence makes you vulnerable: Differential privacy under dependent tuples," in *Network and Distributed System Security Symposium (NDSS)*, 2016.

[34] A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. P. Vadhan, "The limits of two-party differential privacy," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 18, 2011, corrected version of paper presented at FOCS '10.

[35] F. McSherry, "Differential privacy and correlated data," Blog: https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-29.md, 2016.

[36] ——, "Lunchtime for data privacy," Blog: https://github.com/frankmcsherry/blog/blob/master/posts/2016-08-16.md, 2016.

[37] ——, "On "differential privacy as a mutual information constraint"," Blog: https://github.com/frankmcsherry/blog/blob/master/posts/2017-01-26.md, 2017.

[38] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symp. on*. IEEE, 2007, pp. 94–103.

[39] F. D. McSherry, "Privacy integrated queries: An extensible platform for privacy-preserving data analysis," in *2009 ACM SIGMOD Intl. Conf. on Management of data*. ACM, 2009, pp. 19–30.

[40] P. Mittal, "Differential privacy is vulnerable to correlated data – introducing dependent differential privacy," Freedom to Tinker blog: https://freedom-to-tinker.com/2016/08/26/differential-privacy-is-vulnerable-to-correlated-data-introducing-dependent-differential 2016.

[41] J. Pearl, "Direct and indirect effects," in *17th Conf. on Uncertainy in Artificial Intelligence*, 2001, pp. 411–420.

[42] ——, *Causality*, 2nd ed., 2009.

[43] G. Piatetsky, "Did Target really predict a teen's pregnancy? The inside story," *KDnuggets*, 2014.

[44] C. E. Shannon, "Communication theory of secrecy systems," *Bell Labs Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[45] J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang, "Privacy loss in Apple's implementation of differential privacy on MacOS 10.12," *ArXiv*, vol. 1709.02753, 2017.

[46] M. C. Tschantz, A. Datta, A. Datta, and J. M. Wing, "A methodology for information flow experiments," in *Computer Security Foundations Symp.*, 2015.

[47] Y. Wang and M. Kosinski, "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images," 2017.

[48] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. the American Statistical Association*, vol. 60, no. 309, pp. 63–69, 1965.

[49] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *2015 ACM SIGMOD Intl. Conf. on Management of Data*, ser. SIGMOD '15, 2015, pp. 747–762.

[50] T. Zhu, P. Xiong, G. Li, and W. Zhou, "Correlated differential privacy: Hiding information in non-IID data set," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 229–242, 2015.

## APPENDIX

### A. Two Views of Differential Privacy: A Brief History

Throughout this paper, we have mentioned two lines of work about DP. The historically first line, associated with its creators, views DP as not requiring additional assumptions, such as independent data points or an adversary that already knows all but one data point. The historically second line views such assumptions as needed by or implicit in DP. Here, we briefly recount the history of the two lines.

*1) Before Differential Privacy:* The idea of a precise framework for mathematically modeling the conditions under which an adversary does not learn something perhaps starts with Shannon's work on *perfect security* in 1949 [44]. In 1984, this idea led to Goldwasser and Silvio's cryptographic notion of *semantic security*, which relaxes Shannon's requirement by applying to only polynomially computationally bounded adversaries [22] (with antecedents in their earlier 1982 work [21]).

Apparently independently, the statistics community also considered limiting what an adversary would learn. One early work cited by DP papers (e.g., [11]) is Dalenius's 1977 paper on *statistical disclosure* [9]. Dalenius defines statistical disclosures in terms of a *frame* of *objects*, for example, a sampled population of people [9, §4.1]. The objects have data related to them [9, §4.2]. A survey releases some *statistics* over such data for the purpose of fulfilling some *objective* [9, §4.3]. Finally, the adversary may have access to *extra-objective data*, which is auxiliary information other than the statistics released as part of the survey. Dalenius defines a *statistical disclosure* as follows [9, §5]:

> If the release of the statistics $S$ makes it possible to determine the value $D_K$ more accurately than is possible without access to $S$, a disclosure has taken place [...]

where $D_K$ is the value of the attribute $D$ held by the object (e.g., person) $K$. The attribute $D$ and object $K$ may be used in the computation of $S$ or not. The extra-objective data may be used in computing the estimate of $D_K$.

As pointed out by Dwork [11], Dalenius's work is both similar to and different from the aforementioned work on cryptosystems. The most obvious difference is looking at databases and statistics instead of cryptosystems and messages. However, the more significant difference is the presence of the *objective* with a *benefit*, or the need for *utility* in Dwork's nomenclature. That is, the released statistics is to convey some information to the public; whereas, the encrypted message, the cryptosystem's analog to the statistic, only needs to convey information to the intended recipient. Dalenius recognized that this additional need makes the elimination of statistical disclosures "not operationally feasible" and "would place unreasonable restrictions on the kind of statistics that can be released" [9, §18].

Even before the statistics work on statistical nondisclosure, statistical research by S. L. Warner in 1965 introduced the *randomized response* method of providing DP [48]. (His work is more similar to the local formulation of DP [16].) The randomized response model and statistical disclosure can be viewed as the prototypes of the first and second lines of reseach respectively, although these early works appear to have had little impact on the actual formation of the lines of reseach over a quarter century later.

*2) Differential Privacy:* In March 2006, Dwork, McSherry, Nissim, and Smith presented a paper containing the first modern instance of DP under the name of "$\epsilon$-indistinguishable" [17]. The earliest use of the term "differential privacy" comes from a paper by Dwork presented in July 2006 [11]. This paper of Dwork explicitly rejects the view that DP provides associative or inferential privacy [11, p. 8]:

> Note that a bad disclosure can still occur [despite DP], but [DP] assures the individual that it will not be the presence of her data that causes it, nor could the disclosure be avoided through any action or inaction on the part of the user.

and further contains a proof that preventing Dalenius's statistical disclosures while releasing useful statistics is impossible. (The proof was joint work with Naor, with whom Dwork later further developed the impossibility result [18].) Later works further expound upon their position [16], [26].

*3) Questions Raised about Differential Privacy:* In 2011, papers started to question whether DP actually provides a meaningful notion of privacy [7], [27], [19]. These papers point to the fact that a released statistic can enable inferring sensitive information about a person, similar to the attacks Dalenius wanted to prevent [9], even when that statistic was computed using a differentially private algorithm. While the earlier work on DP acknowledged this limitation, these papers provide examples where correlations, or more generally associations, between data points can enable inferences that some people might not expect to be possible under DP. These works kicked off a second line of research (including, e.g., [28], [29], [23], [5], [50], [33]) attempting to find stronger definitions that account for such correlations. In some cases, these papers assert that such inferential threats are violations of privacy and not what people expect of DP. For example, Liu et al.'s abstract states that associations between data points can lead to "degradation in expected privacy levels" [33]. The rest of this subsection provides details about these papers.

In 2011, Kifer and Machanavajjhala published a paper stating that the first popularized claim about DP is that "It makes no assumptions about how data are generated" [27, p. 1]. The paper then explains that "a major criterion for a privacy definition is the following: can it hide the evidence of an individual's participation in the data generating process?" [27, p. 2]. It states [27, p. 2]:

> We believe that under any reasonable formalization of evidence of participation, such evidence can be encapsulated by exactly one tuple [as done by DP] only when all tuples are independent (but not necessarily generated from the same distribution). We believe this independence assumption is a good rule of thumb when considering the applicability of differential privacy.

For this reason, the paper goes on to say "Since evidence of participation requires additional assumptions about the data (as we demonstrate in detail in Sections 3 and 4), this addresses the first popularized claim – that differential privacy requires no assumptions about the data" [27, p. 2]. From context, we take "addresses" to mean *invalidates* since the paper states "The goal of this paper is to clear up misconceptions about differential privacy" [27, p. 2].

In 2012, Kifer and Machanavajjhala published follow up work stating that "we use [the Pufferfish framework] to formalize and prove the statement that differential privacy assumes independence between records" [28, p. 1]. It goes on to say "Assumptionless privacy definitions are a myth: if one wants to publish useful, privacy-preserving sanitized data then one *must* make assumptions about the original data and data-generating process" [28, p. 1, emphasis in original]. In 2014, Kifer and Machanavajjhala published a journal version of their 2012 paper, which makes a similar statement: "Note that assumptions are absolutely necessary – privacy definitions that can provide privacy guarantees without making any assumptions provide little utility beyond the default approach of releasing nothing at all" [29, p. 3:5]. However, this version is, overall, more qualified. For example, it states "The following theorem says that if we have any correlations between records, then some differentially private algorithms leak more information than is allowable (under the odds ratio semantics in Section 3.1)" [29, 3:12–13], which makes it clear that the supposed shortcoming of DP in the face of correlated data points is relative to a particular notion of privacy presented in that paper, roughly, reducing uncertainty about some sensitive fact about a person.

Also in 2014, He et al. published a paper building upon the Pufferfish framework [23]. Referring to the conference version [28], He et al. states [23, p. 1]:

> [Kifer and Machanavajjhala] showed that differential privacy is equivalent to a specific instantiation of the Pufferfish framework, where (a) every property about an individual's record in the data is kept secret, and (b) the adversary assumes that every individual is independent of the rest of the individuals in the data (no correlations). We believe that these shortcomings severely limit the applicability of differential privacy to real world scenarios that either require high utility, or deal with correlated data.

and "Recent work [by Kifer and Machanavajjhala] showed that differentially private mechanisms could still lead to an inordinate disclosure of sensitive information when adversaries have access to publicly known constraints about the data that induce correlations across tuples" [23, p. 3].

In 2013, Li et al. published a paper that states "differential privacy's main assumption is independence" [32, p. 2]. Similar, to the papers by Kifer and Machanavajjhala, this paper assumes a technical definition of privacy, *positive membership privacy*, and makes this assertion since independence is required for DP to imply it. The paper also claims that "the original definition of differential privacy assumes that the adversary has precise knowledge of all the tuples in the dataset" [32, p. 10], which we take as a reference to the strong adversary assumption.

Chen et al.'s 2014 paper is the first of three attempting to provide an associative version of privacy, motivated by Pufferfish, in the face of correlated data [5]. It states "$\epsilon$-differential privacy fails to provide the claimed privacy guarantee in the correlated setting" [5, p. 2] and "$\epsilon$-differential privacy is built on the assumption that all underlying records are independent of each other" [5, p. 7].

The second paper, Zhu et al.'s paper, published in 2015, provides a more accurate accounting of correlations [50]. It states [50, p. 229]:

> An adversary with knowledge on correlated information will have higher chance of obtaining the privacy information, and violating the definition of differential privacy. Hence, how to preserve rigorous differential privacy in a correlated dataset is an emerging issue that needs to be addressed.

It further asserts [50, p. 231]:

> In the past decade, a growing body of literature has been published on differential privacy. Most existing work assumes that the dataset consists of independent records.

and "a major disadvantage of traditional differential privacy is the overlook of the relationship among records, which means that the query result leaks more information than is allowed" [50, p. 232].

The third paper, by Liu et al. in 2016, provides an even more accurate accounting of correlations [33]. A blog post by one of the authors, Mittal, announcing the paper states "To provide its guarantees, DP implicitly assumes that the data tuples in the database, each from a different user, are all independent." [40]. In five comments on this blog post, McSherry posted a summary of his concerns about their paper and blog post. McSherry also treats the paper at length in a blog post [35]. McSherry highlights three statements made by the paper that he finds false [35]: (1) "For providing this guarantee, differential privacy mechanisms assume independence of tuples in the database" [33, p. 1], (2) "To provide its guarantees, DP mechanisms assume that the data tuples (or records) in the database, each from a different user, are all independent." [33, p. 1], and (3) "However, the privacy guarantees provided by the existing DP mechanisms are valid only under the assumption that the data tuples forming the database are pairwise independent" [33, p. 2].

A somewhat different tack is taken in a 2016 paper by Cuff and Yu, which instead focuses on the strong adversary assumption [8, p. 2]:

> The definition of $(\epsilon, \delta)$-DP involves a notion of neighboring database instances. Upon examination one realizes that this has the affect of assuming that the adversary has already learned about all but one entry in the database and is only trying to gather additional information about the remaining entry. We refer to this as the strong adversary assumption, which is implicit in the definition of differential privacy.

Yang et al.'s 2015 paper allows either assumption [49, §1.2]:

> Differential privacy is designed to preserve the privacy in the face of intrusions by the strongest adversary who exactly knows everything about all individual entities except the object of its attack. [...] In fact, as we will show in Section 3, differential privacy does guarantee privacy against intrusion by any adversary when all the entities in the database are independent.

*4) Responses:* In addition to the aforementioned blog post by McSherry [35], other works by those promoting the original view of DP have also re-asserted that DP was never intended to prevent all inferential privacy threats and that doing so is impossible [3], [25], [36]. In a different blog post, McSherry goes the furthest, questioning whether wholesale inferential privacy is the normal meaning of "privacy" or even an appealing concept [36]. He calls it "forgettability", invoking the European Union's right to be forgotten, and points out that preventing inferences prevents people from using data and scientific progress. He suggests that perhaps people should only have an expectation to the privacy of data they own, as provided by DP, and not to the privacy of data about them. He challenges the line of research questioning DP (Appendix A3) to justify the view that forgettability is a form of privacy.

We know no works explicitly responding to this challenge.

### B. Counterexample Involving Zero Probability for Strong Adversary D.P.

Consider Definition 2 modified to look at one distribution $\mathcal{P}$, which represents the actual distribution of the world.

**Definition 10.** *A randomized algorithm $\mathcal{A}$ is said to be $\epsilon$-*Strong Adversary Differentially Private for One Distribution *$\mathcal{P}$ if for all databases $d, d' \in \mathcal{D}^n$ at Hamming distance at most 1, and for all output values $o$, if $\Pr[D{=}d] > 0$ and $\Pr[D{=}d'] > 0$ then*

$$\Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D{=}d] \leq e^{\epsilon} * \Pr_{\mathcal{P},\mathcal{A}}[O{=}o \mid D{=}d'] \tag{8}$$

*where $O = \mathcal{A}(D)$ and $D = \langle D_1, \ldots, D_n \rangle$.*

To prove that this does not imply Definition 1, consider the case of a database holding a single data point whose value could be 0, 1, or 2. Suppose the population $\mathcal{P}$ is such that $\Pr_{\mathcal{P}}[D_1{=}2] = 0$. Consider an algorithm $\mathcal{A}$ such that for the given population $\mathcal{P}$,

$$\Pr_{\mathcal{A}}[\mathcal{A}(0){=}0] = 1/2 \qquad \Pr_{\mathcal{A}}[\mathcal{A}(0){=}1] = 1/2 \tag{9}$$

$$\Pr_{\mathcal{A}}[\mathcal{A}(1){=}0] = 1/2 \qquad \Pr_{\mathcal{A}}[\mathcal{A}(1){=}1] = 1/2 \tag{10}$$

$$\Pr_{\mathcal{A}}[\mathcal{A}(2){=}0] = 1 \qquad \Pr_{\mathcal{A}}[\mathcal{A}(2){=}1] = 0 \tag{11}$$

The algorithm does not satisfy Definition 1 due to its behavior on the input 2. However, using (3),

$$\Pr_{\mathcal{P},\mathcal{A}}[O{=}0 \mid D_1{=}0] = 1/2 \quad \Pr_{\mathcal{P},\mathcal{A}}[O{=}1 \mid D_1{=}0] = 1/2$$
$$\Pr_{\mathcal{P},\mathcal{A}}[O{=}0 \mid D_1{=}1] = 1/2 \quad \Pr_{\mathcal{P},\mathcal{A}}[O{=}1 \mid D_1{=}1] = 1/2$$

While (3) says nothing about $D_1{=}2$ since that has zero probability, this is sufficient to show that the algorithm satisfies Definition 10 since it only applies to data points of non-zero probability. Thus, the algorithm satisfies Definition 10 but not Definition 1.

### C. Details of Causation

We use a slight modification of Pearl's models. The models we use are suggested by Pearl for handling "inherent" randomness [42, p. 220] and differs from the model he typically uses (his Definition 7.1.6) by allowing randomization in the structural equations $F_V$. We find this randomization helpful for modeling the randomization within the algorithm $\mathcal{A}$.

Formally, let $[\![\mathcal{M}]\!](\vec{x}).\vec{Y}$ be the joint distribution over values for the variables $\vec{Y}$ that results from the background variables $\vec{X}$ taking on the values $\vec{x}$ (where these vectors use the same ordering). That is, $[\![\mathcal{M}]\!](\vec{x}).\vec{Y}(\vec{y})$ represents the probability of $\vec{Y} = \vec{y}$ given that the background variables had values $\vec{X} = \vec{x}$. Since the SEM is non-recursive this can be calculated in a bottom up fashion. We show this for the model $\mathcal{M}^{\mathcal{A}}$ with $D_i := R_i$ for all $i$, $D := \langle D_1, \ldots, D_n \rangle$, and $O := \mathcal{A}(D)$:

$$[\![\mathcal{M}^{\mathcal{A}}]\!](r_1, \ldots, r_n).R_i(r_i) = 1$$

$$\begin{aligned}&[\![\mathcal{M}^{\mathcal{A}}]\!](r_1, \ldots, r_n).D_i(r_i)\\ &\quad = \Pr_{F_{D_i}}[F_{D_i}(R_i){=}r_i] = \Pr_{F_{D_i}}[R_i{=}r_i] = 1\end{aligned}$$

$$\begin{aligned}&[\![\mathcal{M}^{\mathcal{A}}]\!](r_1, \ldots, r_n).D(\langle r_1, \ldots, r_n \rangle)\\ &\quad = \Pr_{F_D}[F_D(D_1, \ldots, D_n){=}\langle r_1, \ldots, r_n \rangle]\\ &\quad = \Pr_{F_D}[F_D(F_{D_1}(R_1), \ldots, F_{D_n}(R_n)){=}\langle r_1, \ldots, r_n \rangle]\\ &\quad = \Pr_{F_D}[F_D(R_1, \ldots, R_n){=}\langle r_1, \ldots, r_n \rangle]\\ &\quad = \Pr_{F_D}[\langle R_1, \ldots, R_n \rangle{=}\langle r_1, \ldots, r_n \rangle] = 1\end{aligned}$$

and

$$\begin{aligned}[\![\mathcal{M}^{\mathcal{A}}]\!](r_1, \ldots, r_n).O(o) &= \Pr_{F_O}[F_O(D){=}o]\\ &= \Pr_{\mathcal{A}}[\mathcal{A}(\langle r_1, \ldots, r_n \rangle){=}o]\end{aligned}$$

We can raise the calculations above to work over $\mathcal{P}$ instead of a concrete assignment of values $\vec{x}$. Intuitively, the only needed change is that, for background variables $\vec{X}$,

$$\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[\vec{Y}{=}\vec{y}] = \sum_{\vec{x} \in \vec{\mathcal{X}}} \Pr_{\mathcal{P}\mathcal{A}}[\vec{X}{=}\vec{x}] * [\![\mathcal{M}^{\mathcal{A}}]\!](\vec{x}).\vec{Y}(\vec{y})$$

where $\vec{X}$ are all the background variables.[1]

The following lemma will not only be useful, but will illustrate the above general points on the model $\mathcal{M}^{\mathcal{A}}$ that concerns us.

**Lemma 1.** For all algorithms $\mathcal{A}$, $\mathcal{P}$, all $o$, and all $d_1, \ldots, d_n$,

$$\begin{aligned}&\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O{=}o \mid \mathrm{do}(D_1{:=}d_1, \ldots, D_n{:=}d_n)]\\ &\qquad = \Pr_{\mathcal{A}}[\mathcal{A}(d_1, \ldots, d_n){=}o]\end{aligned}$$

*Proof.* Let $F_{d_i}()$ represent the constant function with no arguments that always returns $d_i$. The structural equation for $D_i$ is $F_{d_i}$ in $\mathcal{M}^{\mathcal{A}}[D_1{:=}d_1] \cdots [D_n{:=}d_n]$. As before, we compute bottom up, but this time on the modified SEM:

$$[\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d_1] \cdots [D_n{:=}d_n]]\!](r_1, \ldots, r_n).R_i(r_i) = 1$$

[1] This is Pearl's equation (7.2) raised to work on probabilistic structural equations $F_V$ [42, p. 205].

$$[\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d_1]\cdots[D_n{:=}d_n]]\!](r_1,\ldots,r_n).D_i(d_i) = \Pr_{F_{d_i}}[F_{d_i}()=d_i] = 1$$

$$\begin{aligned}[\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d_1]\cdots[D_n{:=}d_n]]\!](r_1,\ldots,r_n).D(\langle d_1,\ldots,d_n\rangle) &= \Pr_{F_D}[F_D(D_1,\ldots,D_n)=\langle d_1,\ldots,d_n\rangle]\\ &= \Pr_{F_D}[F_D(F_{D_1}(),\ldots,F_{D_n}())=\langle d_1,\ldots,d_n\rangle]\\ &= \Pr_{F_D}[F_D(d_1,\ldots,d_n)=\langle d_1,\ldots,d_n\rangle]\\ &= \Pr_{F_D}[\langle d_1,\ldots,d_n\rangle=\langle d_1,\ldots,d_n\rangle] = 1\end{aligned}$$

$$\begin{aligned}[\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d_1]\cdots[D_n{:=}d_n]]\!](r_1,\ldots,r_n).O(o) &= \Pr_{F_O}[F_O(D)=o] = \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1,\ldots,d_n\rangle)=o]\end{aligned}$$

Thus,

$$\begin{aligned}&\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O=o \mid \mathrm{do}(D_1{:=}d_1,\ldots,D_n{:=}d_n)]\\ &= \Pr_{\mathcal{M}^{\mathcal{A}}[D_1:=d_1]\cdots[D_n:=d_n],\mathcal{P}}[O=o]\\ &= \sum_{\vec{r}\in\mathcal{R}^n}\Pr_{\mathcal{P}}[\vec{R}=\vec{r}] * [\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d_1]\cdots[D_n{:=}d_n]]\!](\vec{r}).O(o)\\ &= \sum_{\vec{r}\in\mathcal{R}^n}\Pr_{\mathcal{P}}[\vec{R}=\vec{r}] * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1,\ldots,d_n\rangle)=o]\\ &= \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1,\ldots,d_n\rangle)=o] * \textstyle\sum_{\vec{r}\in\mathcal{R}^n}\Pr_{\mathcal{P}}[\vec{R}=\vec{r}]\\ &= \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1,\ldots,d_n\rangle)=o] * 1\\ &= \Pr_{\mathcal{A}}[\mathcal{A}(\langle d_1,\ldots,d_n\rangle)=o]\end{aligned}$$

□

**Lemma 2.** For all algorithms $\mathcal{A}$, $\mathcal{P}$, $o$, $j$, and $d'_j$,

$$\begin{aligned}&\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O=o \mid \mathrm{do}(D_j{=}d'_j)]\\ &= \sum_{\langle r_1,\ldots,r_{j-1},r_{j+1},\ldots,r_n\rangle\in\mathcal{R}^{n-1}}\Pr_{\mathcal{P}}\left[\wedge_{i\in\{1,\ldots,j-1,j+1,\ldots,n\}}R_i{=}r_i\right]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(r_1,\ldots,r_{j-1},d'_j,r_{j+1},\ldots,r_n)=o]\\ &= \sum_{\langle d_1,\ldots,d_{j-1},d_{j+1},\ldots,d_n\rangle\in\mathcal{D}^{n-1}}\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}\left[\wedge_{i\in\{1,\ldots,j-1,j+1,\ldots,n\}}D_i{=}d_i\right]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(d_1,\ldots,d_{j-1},d'_j,d_{j+1},\ldots,d_n)=o]\end{aligned}$$

*Proof.* With out loss of generality, assume $j$ is 1. Let $F_{d'_1}()$ represent the constant function with no arguments that always returns $d'_1$. The structural equation for $D_1$ is $F_{d'_1}$ in $\mathcal{M}^{\mathcal{A}}[D_1{:=}d'_1]$. As before, we compute bottom up, but this time on the modified SEM:

$$[\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d'_1]]\!](r_1,\ldots,r_n).R_i(r_i) = 1$$

holds as before. The behavior of $D_i$ varies based on whether $i=1$:

$$\begin{aligned}[\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d'_1]]\!](r_1,\ldots,r_n).D_1(d'_1) &= \Pr_{F_{d'_1}}[F_{d'_1}()=d'_1] = 1\\ [\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d'_1]]\!](r_1,\ldots,r_n).D_i(r_i) &= \Pr_{F_{D_i}}[F_{D_i}(R_i)=r_i]\\ &= \Pr_{F_{D_i}}[R_i=r_i] = 1\end{aligned}$$

for all $i\neq 1$. Thus,

$$\begin{aligned}&[\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d'_1]]\!](r_1,\ldots,r_n).D(\langle d'_1,r_2,\ldots,r_n\rangle)\\ &= \Pr_{F_D}[F_D(D_1,D_2\ldots,D_n)=\langle d'_1,r_2,\ldots,r_n\rangle]\\ &= \Pr_{F_D}[F_D(F_{d'_1}(),F_{D_2}(R_2),\ldots,F_{D_n}(R_n))=\langle d'_1,r_2,\ldots,r_n\rangle]\\ &= \Pr_{F_D}[F_D(d'_1,r_2,\ldots,r_n)=\langle d'_1,r_2,\ldots,r_n\rangle]\\ &= \Pr_{F_D}[\langle d'_1,r_2,\ldots,d_n\rangle=\langle d'_1,r_2,\ldots,r_n\rangle] = 1\end{aligned}$$

and

$$\begin{aligned}[\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d'_1]]\!](r_1,\ldots,r_n).O(o) &= \Pr_{F_O}[F_O(D)=o] = \Pr_{\mathcal{A}}[\mathcal{A}(\langle d'_1,r_2,\ldots,r_n\rangle)=o]\end{aligned}$$

Thus,

$$\begin{aligned}&\Pr_{\mathcal{M}^{\mathcal{A}},\mathcal{P}}[O=o \mid \mathrm{do}(D_1{:=}d'_1)]\\ &= \Pr_{\mathcal{M}^{\mathcal{A}}[D_1:=d'_1],\mathcal{P}}[O=o]\\ &= \sum_{r_1,\ldots,r_n\in\mathcal{R}^n}\Pr_{\mathcal{P}}[R_1=r_1,\ldots,R_n=r_n]\\ &\qquad * [\![\mathcal{M}^{\mathcal{A}}[D_1{:=}d'_1]]\!](r_1,\ldots,r_n).O(o)\\ &= \sum_{r_1,\ldots,r_n\in\mathcal{R}^n}\Pr_{\mathcal{P}}[R_1=r_1,\ldots,R_n=r_n]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d'_1,r_2,\ldots,r_n\rangle)=o]\\ &= \sum_{r_1,\ldots,r_n\in\mathcal{R}^n}\Pr_{\mathcal{P}}[R_1=r_1 \mid R_2=r_2,\ldots,R_n=r_n]\\ &\qquad * \Pr_{\mathcal{P}}[R_2=r_2,\ldots,R_n=r_n]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d'_1,r_2,\ldots,r_n\rangle)=o]\\ &= \sum_{r_2,\ldots,r_n\in\mathcal{R}^n}\sum_{r_1\in\mathcal{R}}\Pr_{\mathcal{P}}[R_1=r_1 \mid R_2=r_2,\ldots,R_n=r_n]\\ &\qquad * \Pr_{\mathcal{P}}[R_2=r_2,\ldots,R_n=r_n]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d'_1,r_2,\ldots,r_n\rangle)=o]\\ &= \sum_{r_2,\ldots,r_n\in\mathcal{R}^n}\Pr_{\mathcal{P}}[R_2=r_2,\ldots,R_n=r_n]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d'_1,r_2,\ldots,r_n\rangle)=o]\\ &\qquad * \sum_{r_1\in\mathcal{R}}\Pr_{\mathcal{P}}[R_1=r_1 \mid R_2=r_2,\ldots,R_n=r_n]\\ &= \sum_{r_2,\ldots,r_n\in\mathcal{R}^n}\Pr_{\mathcal{P}}[R_2=r_2,\ldots,R_n=r_n]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d'_1,r_2,\ldots,r_n\rangle)=o] * 1\\ &= \sum_{r_2,\ldots,r_n\in\mathcal{R}^n}\Pr_{\mathcal{P}}[R_2=r_2,\ldots,R_n=r_n]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d'_1,r_2,\ldots,r_n\rangle)=o]\\ &= \sum_{d_2,\ldots,d_n\in\mathcal{D}^n}\Pr_{\mathcal{P}}[D_2=d_2,\ldots,D_n=d_n]\\ &\qquad * \Pr_{\mathcal{A}}[\mathcal{A}(\langle d'_1,d_2,\ldots,d_n\rangle)=o]\end{aligned}$$

where the last line follows since $D_i = R_i$ for $i\neq 1$. □