Inferential Privacy Guarantees for Differentially Private Mechanisms

Arpita Ghosh[*]   Robert Kleinberg[†]

Abstract

The correlations and network structure amongst individuals in datasets today—whether explicitly articulated, or deduced from biological or behavioral connections—pose new issues around privacy guarantees, because of inferences that can be made about one individual from another's data. This motivates quantifying privacy in networked contexts in terms of 'inferential privacy'—which measures the change in beliefs about an individual's data from the result of a computation—as originally proposed by Dalenius in the 1970's. Inferential privacy is implied by differential privacy when data are independent, but can be much worse when data are correlated; indeed, simple examples, as well as a general impossibility theorem of Dwork and Naor, preclude the possibility of achieving non-trivial inferential privacy when the adversary can have arbitrary auxiliary information. In this paper, we ask how differential privacy guarantees translate to guarantees on inferential privacy in networked contexts: specifically, under what limitations on the adversary's information about correlations, modeled as a prior distribution over datasets, can we deduce an inferential guarantee from a differential one?

We prove two main results. The first result pertains to distributions that satisfy a natural positive-affiliation condition, and gives an upper bound on the inferential privacy guarantee for any differentially private mechanism. This upper bound is matched by a simple mechanism that adds Laplace noise to the sum of the data. The second result pertains to distributions that have weak correlations, defined in terms of a suitable "influence matrix". The result provides an upper bound for inferential privacy in terms of the differential privacy parameter and the spectral norm of this matrix.

---

[*]Cornell University, Ithaca, NY, USA. `arpitaghosh@cornell.edu`

[†]Cornell University, Ithaca, NY, USA, and Microsoft Research New England, Cambridge, MA, USA. `robert.kleinberg@cornell.edu`

1   Introduction

Privacy has always been a central issue in the discourse surrounding the collection and use of personal data. As the nature of data collected online grows richer, however, fundamentally new privacy issues emerge. In a thought-provoking piece entitled "Networked Rights and Networked Harms" [22], the sociologists Karen Levy and danah boyd argue that the 'networks' surrounding data today—whether articulated (as in explicitly declared friendships on social networks), behavioral (as in connections inferred from observed behavior), or biological (as in genetic databases)—raise conceptually new questions that current privacy law and policy cannot address. Levy and boyd present case studies to demonstrate how the current individual-centric legal frameworks for privacy do not provide a means to account for the networked contexts now surrounding personal data.

An analogous question arises on the formal front. One of computer science's fundamental contributions to the public debate about private data—most prominently via the literature on *differential privacy*[1] [10]— has been to provide a means to *measure* privacy loss, which enables evaluating the privacy implications of proposed data analyses and disclosures in quantitative terms. However, differential privacy focuses on the privacy loss to an individual by *her* contribution to a dataset, and therefore—by design—does not capture all of the privacy losses from inferences that could be made about one person's data due to its correlations with *other* data in networked contexts. For instance, the privacy implications of a database such as 23andme for one individual depend not just on that person's own data and the computation performed, but also on her siblings' data.

In this paper, we look to understand the implications of such 'networked' data for formal privacy guarantees. How much can be learnt about a single individual from the result of a computation on correlated data, and how does this relate to the differential privacy guarantee of the computation?

Inferential privacy.   A natural way of assessing whether a mechanism $\mathcal{M}$ protects the privacy of an individual is to ask, "Is it possible that someone, after observing the mechanism's output, will learn a lot about the individual's private data?" In other words, what is the *inferential privacy*—the largest possible ratio between the posterior and prior beliefs about an individual's data after observing the result of a computation on the database? (This quantity is identical to the differential privacy parameter of the mechanism when individuals' data are independent; see §2 and [18].)

The inferential privacy guarantee will depend, of course, on both the nature of the correlations in the database and on the precise mechanism used to perform the computation. Instead of seeking to *design* algorithms that achieve a particular inferential privacy guarantee—which would necessitate choosing a particular computational objective and correlation structure—we instead seek to *analyze* the inferential privacy guarantees provided by differentially private algorithms. Specifically, we ask the following question: consider the class of all mechanisms providing a certain differential privacy guarantee, say ε. What is the worst possible inferential privacy guarantee for a mechanism in this class?

This question is pertinent to a policy-maker who can prescribe that analysts provide some degree of differential privacy to individuals while releasing their results, but cannot control how—*i.e.*, using what specific algorithm—the analyst will provide this guarantee. In other words, rather than an algorithm designer who wants to design an inferential privacy-preserving algorithm (for a particular scenario), this question adopts the perspective of a policy-maker who can set privacy standards that analysts must obey, but is agnostic to the analysts' computational objectives. We choose the differential privacy guarantee as our measure of privacy for many reasons: it is, at present, the only widely-agreed-upon privacy guarantee known to provide strong protections even against arbitrary side information; there is a vast toolbox of differentially

---

[1]Differential privacy, which measures privacy via the relative amount of new information disclosed about an individual's data by her participation in a dataset, has emerged as the primary theoretical framework for quantifying privacy loss.

private algorithms and a well-understood set of composition rules for combining them to yield new ones; finally, differential privacy is now beginning to make its way into policy and legal frameworks as a potential means for quantifying privacy loss.

Measuring privacy loss via inferential privacy formalizes Dalenius's [4] desideratum that "access to a statistical database should not enable one to learn anything about an individual that could not be learned without access". While it is well known[2] that non-trivial inferential privacy guarantees are incompatible with non-trivial utility guarantees in the presence of arbitrary auxiliary information, our primary contribution is modeling and quantifying *what* degree of inferential privacy is in fact achievable under a *particular* form of auxiliary information, such as that resulting from a known correlation structure or a limited set of such structures. For example, as noted earlier, if the individuals' rows in the database are conditionally independent given the adversary's auxiliary information, then the inferential privacy guarantee for any individual collapses to her differential privacy guarantee. At the other extreme, when all individuals' data are perfectly correlated, the inferential privacy parameter can exceed the differential privacy parameter by a factor of n (the number of individuals in the database) as we will see below. What happens for correlations that lie somewhere in between these two extremes? Do product distributions belong to a broader class of distributions with *benign correlations* which ensure that an individual's inferential privacy is not much worse than her differential privacy? A key contribution of our paper (Theorem 4.2) answers this question affirmatively while linking it to a well-known sufficient condition for 'correlation decay' in mathematical physics.

Correlations in networked datasets and their privacy consequences. We start with a caricature example to begin exploring how one might address these questions in a formal framework. Consider a database which contains an individual Athena and her (hypothetical) identical twin Adina, who is so identical to Athena that the rows in the database corresponding to Athena and Adina are identical in (the databases corresponding to) every possible state of the world. A differential privacy guarantee of ϵ to all database participants translates to an inferential privacy guarantee of only 2ϵ to Athena (and her twin), since the "neighboring" database where Athena and Adina are different simply cannot exist.[3]

The erosion of Athena's privacy becomes even more extreme if the database contains $n > 2$ individuals and they are all clones of Athena; a generalization of the preceding calculation now shows that the inferential privacy parameter is $n\varepsilon$. However, in reality one is unlikely to participate in a database with many identical clones of oneself. Instead, it is interesting to consider cases with non-extreme correlations. For example, suppose now that the database contains data from Zeus and all of his descendants, and that every child's bit matches the parent's bit with probability $p > \frac{1}{2}$. The degree of privacy afforded to Zeus now depends on many aspects of the model: the strength of the correlation (p), the number of individuals in the database (n), and structural properties of the tree of family relationships—its branching factor and depth, for instance. Which of these parameters contribute most crucially to inferential privacy? Is Zeus more likely to be implicated by his strong correlation with a few close relatives, or by a diffuse "dragnet" of correlations with his distant offspring?

In general, of course, networked databases, and the corresponding inferential privacy guarantees, do not come with as neat or convenient a correlation structure as in this example. In full generality, we can represent the idea of networked similarity via a joint distribution on databases that gives the prior probability of each particular combination of bits. So, for example, a world where all individuals in the database are "twins" would correspond to a joint distribution which has non-zero probability only on the all-zeros and all-ones

---

[2]see, *e.g.*, [9, 10]

[3]Differential privacy guarantees that the probability of an outcome o changes by at most a factor e^ϵ amongst databases at Hamming distance one, so that if $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ denote the databases where the bits of Athena and Adina are $(0,0)$, $(1,0)$ and $(1,1)$ respectively, differential privacy guarantees that $\Pr(o|\mathbf{x}_1) \leq e^\epsilon \cdot \Pr(o|\mathbf{x}_2) \leq e^{2\epsilon} \cdot \Pr(o|\mathbf{x}_3)$. From here, a simple calculation using Bayes' Law—see equation (3) in Section 2—implies that: $\frac{\Pr(Athena=1|o)/\Pr(Athena=0|o)}{\Pr(Athena=1)/\Pr(Athena=0)} \leq e^{2\epsilon}$, so that the inferential privacy guarantee is 2ϵ.

databases, whereas a world where everyone's data is independent has multiplicative probabilities for each database.

Such a model of correlations allows capturing a rich variety of networked contexts: in addition to situations where a single database contains sensitive information about n individuals whose data have known correlations, it also captures the situation—perhaps closest to reality—where there are multiple databases to which multiple individuals contribute different (but correlated) pieces of information. In this latter interpretation, an inferential privacy guarantee limits the amount that an adversary may learn about *one* individual's contribution to *one* database, despite the correlations both across individuals and between a single individual's contributions to different databases.[4]

Our results.   Consider a policy-maker who specifies that an analyst must provide a certain differential privacy guarantee, and wants to comprehend the inferential privacy consequences of this policy for the population whose (correlated) data is being utilized. Our two main results can be interpreted as providing guidance to such a policy maker. The first result (Theorem 3.4) supplies a closed-form expression for the inferential privacy guarantee as a function of the differential privacy parameter when data are *positively affiliated*[5][24]. The second result (Theorem 4.2) allows understanding the behavior of the inferential privacy guarantee as a function of the degree of correlation in the population; it identifies a property of the joint distribution of data that ensures that the policy-maker can meet a given inferential privacy target via a differential privacy requirement that is a constant-factor scaling of that target.

Among all mechanisms with a given differential privacy guarantee, which ones yield the worst inferential privacy when data are correlated? Our first main result, Theorem 3.4, answers this question when data are positively affiliated, in addition to giving a closed-form expression for the inferential privacy guarantee. The answer takes the following form: we identify a simple property of mechanisms (Definition 3.3) such that any mechanism satisfying the property achieves the worst-case guarantee. Strikingly, the form of the worst-case mechanism does not depend on the joint distribution of the data, but *only* on the fact that the distribution satisfies positive affiliation. We also provide one example of such a mechanism: a "noisy-sum mechanism" that simply adds Laplace noise to the sum of the data. This illustrates that the worst inferential privacy violations occur even with one of the most standard mechanisms for implementing differential privacy, rather than some contrived mechanisms.

The aforementioned results provide a sharp bound on the inferential privacy guarantee for positively affiliated distributions, but they say little about whether this bound is large or small in comparison to the differential privacy guarantee. Our second main result fills this gap: it provides an upper bound on the inferential privacy guarantee when a *bounded affiliation* condition is satisfied on the correlations between individuals' rows in a database. Representing the strengths of these correlations by an *influence matrix* Γ, Theorem 4.2 asserts that if all row sums of this matrix are bounded by $1-\delta$ then every individual's inferential privacy is bounded by $2\epsilon/\delta$, regardless of whether or not the data are positively affiliated. Thus, Theorem 4.2 shows that in order to satisfy ν-inferential privacy against all distributions with $(1-\delta)$-bounded affiliation, it suffices for the policy-maker to set $\epsilon = \delta\nu/2$. We complement this result with an example showing that the ratio of inferential privacy to differential privacy can indeed be as large as $\Omega(\frac{1}{\delta})$, as the row sums of the influence matrix approach 1. Thus, the equivalence between inferential and differential privacy, $\nu = \epsilon$, which holds for independent distributions, degrades gracefully to $\nu = O(\epsilon)$ as one introduces correlation into the distribution, but only up to a point: as the row sums of the influence matrix approach 1, the ratio ν/ϵ can diverge to infinity, becoming unbounded when the row sums exceed 1.

---

[4]We are grateful to Kobbi Nissim for suggesting this interpretation of our model.

[5]Positive affiliation (Definition 3.1) is a widely used notion of positive correlation amongst random variables. It is satisfied, for example, by graphical models whose edges encode positively-correlated conditional distributions on pairs of variables.

Techniques. Our work exposes a formal connection between the analysis of inferential privacy in networked contexts and the analysis of spin systems in mathematical physics. In brief, application of a differentially private mechanism to correlated data is analogous to application of an external field to a spin system. Via this analogy, physical phenomena such as phase transitions can be seen to have consequences for data privacy: they imply that small variations in the amount of correlation between individuals' data, or in the differential privacy parameter of a mechanism, can sometimes have gigantic consequences for inferential privacy (§A.2 elaborates on this point). Statistical physics also supplies the blueprint for Theorem 4.2 and its proof: our bounded affiliation condition can be regarded as a multiplicative analogue of Dobrushin's Uniqueness Condition [5, 6], and our proof of Theorem 4.2 adapts the proof technique of the Dobrushin Comparison Theorem [6, 11, 21] from the case of additive approximation to multiplicative approximation. Since Dobrushin's Uniqueness Condition is known to be one of the most general conditions ensuring exponential decay of correlations in physics, our Theorem 4.2 can informally be interpreted as saying that differential privacy implies strong inferential privacy guarantees when the structure of networked correlations is such that, conditional on the adversary's side information, the correlations between individuals' data decay rapidly as their distance in the network increases.

Related work. Our paper adopts the term *inferential privacy* as a convenient shorthand for a notion that occurs in many prior works, dating back to Dalenius [4], which is elsewhere sometimes called "before/after privacy" [10], "semantic privacy" [18], or "noiseless privacy" [3]. Dwork and McSherry observed that differentially private mechanisms supply inferential privacy against adversaries whose prior is a product distribution; this was stated implicitly in [7] and formalized in [18]. However, when adversaries can have arbitrary auxiliary information, inferential privacy becomes unattainable except by mechanisms that provide little or no utility; see [9, 19] for precise impossibility results along these lines. Responses to this predicament have varied: some works propose stricter notions of privacy based on simulation-based semantics, *e.g. zero-knowledge privacy* [13], others propose weaker notions based on restricting the set of prior distributions that the adversary may have, *e.g. noiseless privacy* [3], and others incorporate aspects of both responses, *e.g. coupled-world privacy* [2] and the *Pufferfish* framework [20]. Our work is similar to some of the aforementioned ones in that we incorporate restrictions on the adversary's prior distribution, however our goal is quite different: rather than proposing a new privacy definition or a new class of mechanisms, we quantify how effectively an existing class of mechanisms (ε-differentially private mechanisms) achieves an existing privacy goal (inferential privacy).

Relations between differential privacy and network analysis have been studied by many authors—*e.g.* [17] and the references therein—but this addresses a very different way in which networks relate to privacy: the network in those works is part of the data, whereas in ours it is a description of the auxiliary information.

The exponential mechanism of McSherry and Talwar [23] can be interpreted in terms of Gibbs measures, and Huang and Kannan [16] leveraged this interpretation and applied a non-trivial fact about free-energy minimization to deduce consequences about incentive compatibility of exponential mechanisms. Aside from their work, we are not aware of other applications of statistical mechanics in differential privacy.

2 Defining Inferential Privacy

In this section we specify our notation and basic assumptions and definitions. A population of n individuals is indexed by the set $[n] = \{1, \ldots, n\}$. Individual i's private data is represented by the element $x_i \in X$, where X is a finite set. Except in §4 we will assume throughout, for simplicity, that $X = \{0, 1\}$, *i.e.* each individual's private data is a single bit. When focusing on the networked privacy guarantee for a particular individual, we denote her index by $a \in [n]$ and sometimes refer to her as "Athena".

A database is an n-tuple $\mathbf{x} \in X^n$ representing the private data of each individual. As explained in

Section 1, our model encodes the 'network' structure of the data using a probability distribution on X^n; we denote this distribution by μ. A computation performed on the database $\mathbf{x}$, whose outcome will be disclosed to one or more parties, is called a *mechanism* and denoted by $\mathcal{M}$. The set of possible outcomes of the computation is $\mathcal{O}$, and a generic outcome will be denoted by $o \in \mathcal{O}$. %

Differential privacy [7, 8, 10].   For a database $\mathbf{x} = (x_1, \ldots, x_n)$ and an individual $i \in [n]$, we use $\mathbf{x}_{-i}$ to denote the $(n-1)$-tuple formed by omitting x_i from $\mathbf{x}$, *i.e.* $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. We define an equivalence relation $\sim_i$ by specifying that $\mathbf{x} \sim_i \mathbf{x}' \Leftrightarrow \mathbf{x}_{-i} = \mathbf{x}'_{-i}$. For a mechanism $\mathcal{M}$ and individual i, the differential privacy parameter ϵ_i is defined by

$$e^{\epsilon_i} = \max \left\{ \frac{\Pr(\mathcal{M}(\mathbf{x}) = o)}{\Pr(\mathcal{M}(\mathbf{x}') = o)} \;\middle|\; \mathbf{x} \sim_i \mathbf{x}', o \in \mathcal{O} \right\}.$$

For any vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ we say that $\mathcal{M}$ is $\boldsymbol{\epsilon}$-differentially private if the differential privacy parameter of $\mathcal{M}$ with respect to i is at most ϵ_i, for every individual i.

Inferential privacy.   We define *inferential privacy* as an upper bound on the (multiplicative) change in $\frac{\Pr(x_a = z_1)}{\Pr(x_a = z_0)}$ when performing a Bayesian update from the prior distribution μ to the posterior distribution after observing $\mathcal{M}(\mathbf{x}) = o$. (If $\mathcal{M}$ has uncountably many potential outcomes, we must instead consider doing a Bayesian update after observing a positive-probability event $\mathcal{M}(\mathbf{x}) \in S$ for some set of outcomes S.)

Definition 2.1. We say that mechanism $\mathcal{M}$ satisfies ν-inferential privacy (with respect to individual a) if the inequality $\frac{\Pr(x_a = z_1 | \mathcal{M}(\mathbf{x}) \in S)}{\Pr(x_a = z_0 | \mathcal{M}(\mathbf{x}) \in S)} \leq e^{\nu} \cdot \frac{\Pr(x_a = z_1)}{\Pr(x_a = z_0)}$ holds for all $z_0, z_1 \in X$ and all $S \subset \mathcal{O}$ such that $\Pr(\mathcal{M}(\mathbf{x}) \in S) > 0$. The *inferential privacy parameter* of $\mathcal{M}$ is the smallest ν with this property.

Inferential versus differential privacy.   A short calculation using Bayes' Law illuminates the relation between these two privacy notions.

$$\frac{\Pr(x_a = z_1 \mid \mathcal{M}(\mathbf{x}) \in S)}{\Pr(x_a = z_0 \mid \mathcal{M}(\mathbf{x}) \in S)} = \frac{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid x_a = z_1)}{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid x_a = z_0)} \cdot \frac{\Pr(x_a = z_1)}{\Pr(x_a = z_0)}.$$

Thus, the inferential privacy parameter of mechanism $\mathcal{M}$ with respect to individual a is determined by:

$$e^{\nu_a} = \sup \left\{ \frac{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid x_a = z_1)}{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid x_a = z_0)} \;\middle|\; z_0, z_1 \in X, \Pr(\mathcal{M}(\mathbf{x}) \in S) > 0 \right\}. \tag{1}$$

Equivalently, if μ^0, μ^1 denote the conditional distributions of $\mathbf{x}_{-a}$ given that $x_a = z_0$ and $x_a = z_1$, respectively, then $\mathcal{M}$ is ν_a-inferentially private if

$$\Pr(\mathcal{M}(z_1, \mathbf{y}_1) \in S) \leq e^{\nu_a} \Pr(\mathcal{M}(z_0, \mathbf{y}_0) \in S) \quad \text{when } \mathbf{y}_0 \sim \mu^0, \, \mathbf{y}_1 \sim \mu^1. \tag{2}$$

For comparison, differential privacy asserts

$$\Pr(\mathcal{M}(z_1, \mathbf{y}) \in S) \leq e^{\epsilon_a} \Pr(\mathcal{M}(z_0, \mathbf{y}) \in S) \quad \forall \mathbf{y}. \tag{3}$$

When individuals' rows in the database are independent, $\mu^0 = \mu^1$ and (3) implies (2) with $\nu_a = \epsilon_a$ by averaging over $\mathbf{y}$. In other words, when bits are independent, ϵ_a-differential privacy implies ϵ_a-inferential privacy. When bits are correlated, however, this implication breaks down because the databases $\mathbf{y}_0, \mathbf{y}_1$ in (2) are sampled from different distributions. The 'twins example' from §1 illustrates concretely why this makes a difference: if μ^0 and μ^1 are point-masses on $(0, \ldots, 0)$ and $(1, \ldots, 1)$, respectively, then the inferential privacy parameter of $\mathcal{M}$ is determined by the equation $e^{\nu} = \sup_S \left\{ \frac{\Pr(\mathcal{M}(1,\ldots,1) \in S)}{\Pr(\mathcal{M}(0,\ldots,0) \in S)} \right\}$. For an ϵ-differentially-private mechanism this ratio may be as large as $e^{n\epsilon}$ since the Hamming distance between $(0, \ldots, 0)$ and $(1, \ldots, 1)$ is n.

3 Positively Affiliated Distributions

Suppose a designer wants to ensure that Athena receives an inferential privacy guarantee of ν, given a joint distribution μ on the data of individuals in the database. What is the largest differential privacy parameter ϵ that ensures this guarantee? The question is very challenging even in the special case of binary data (*i.e.*, when $X = \{0, 1\}$) because the ratio defining inferential privacy (Equation 2) involves summing exponentially many terms in the numerator and denominator. Determining the worst-case value of this ratio over all differentially private mechanisms $\mathcal{M}$ can be shown to be equivalent to solving a linear program with exponentially many variables (the probability of the event $\mathcal{M}(\mathbf{x}) \in S$ for every potential database $\mathbf{x}$) and exponentially many constraints (a differential privacy constraint for every pair of adjacent databases).

Our main result in this section answers this question when individuals' data are binary-valued and *positively affiliated* [12, 24], a widely used notion of positive correlation: Theorem 3.4 gives a closed-form formula (Equation 6) that one can invert to solve for the maximum differential privacy parameter ϵ that guarantees inferential privacy ν when data are positively affiliated. The theorem also characterizes the 'extremal' mechanisms achieving the worst-case inferential privacy guarantee in (6) as those satisfying a 'maximally biased' property (Definition 3.3). Intuitively, if one wanted to signal as strongly as possible that Athena's bit is 1 (resp., 0), a natural strategy—given that Athena's bit correlates positively with everyone else's—is to have a distinguished outcome (or set of outcomes) whose probability of being output by the mechanism increases with the number of 1's (resp., the number of 0's) in the database 'as rapidly as possible', subject to differential privacy constraints. Theorem 3.4 establishes that this intuition is valid under the positive affiliation assumption. (Interestingly, the intuition is not valid if one merely assumes that Athena's own bit is positively correlated with every other individual's bit; see Remark 3.6.) Lemma 3.5 provides one simple example of a maximally-biased mechanism, namely a "noisy-sum mechanism" that simply adds Laplace noise to the sum of the bits in the database. Thus, the worst-case guarantee in Theorem 3.4 is achieved not by contrived worst-case mechanisms, but by one of the most standard mechanisms in the differential privacy literature.

We begin by defining positive affiliation, a concept that has proven extremely valuable in auction theory (the analysis of interdependent value auctions), statistical mechanics, and probabilistic combinatorics. Affiliation is a strong form of positive correlation between random variables: informally, positive affiliation means that if some individuals' bits are equal to 1 (or more generally, if their data is 'large'), other individuals' bits are more likely to equal 1 as well (and similarly for 0). We formally define positive affiliation for our setting below and then state a key lemma concerning positively affiliated distributions, the FKG inequality.

Definition 3.1 (Positive affiliation). Given any two strings $\mathbf{x}_1, \mathbf{x}_2 \in \{0, 1\}^n$, let $\mathbf{x}_1 \vee \mathbf{x}_2$ and $\mathbf{x}_1 \wedge \mathbf{x}_2$ denote their pointwise maximum and minimum, respectively. A joint distribution μ on $\{0, 1\}^n$ satisfies positive affiliation if

$$\mu(\mathbf{x}_1 \vee \mathbf{x}_2) \cdot \mu(\mathbf{x}_1 \wedge \mathbf{x}_2) \ \geq \ \mu(\mathbf{x}_1) \cdot \mu(\mathbf{x}_2)$$

for all possible pairs of strings $\mathbf{x}_1, \mathbf{x}_2$. Equivalently, μ satisfies positive affiliation if $\log \mu(\mathbf{x})$ is a supermodular function of $\mathbf{x} \in \{0, 1\}^n$.

Lemma 3.2 (FKG inequality; Fortuin et al. [12]). *If f, g, h are three real-valued functions on $\{0, 1\}^n$ such that f and g are monotone and $\log h$ is supermodular, then*

$$\left[\sum_{\mathbf{x}} f(\mathbf{x}) g(\mathbf{x}) h(\mathbf{x}) \right] \left[\sum_{\mathbf{x}} h(\mathbf{x}) \right] \geq \left[\sum_{\mathbf{x}} f(\mathbf{x}) h(\mathbf{x}) \right] \left[\sum_{\mathbf{x}} g(\mathbf{x}) h(\mathbf{x}) \right]. \tag{4}$$

In order to state the main result of this section, Theorem 3.4, we must define a property that characterizes the mechanisms whose inferential privacy parameter meets the worst-case bound stated in the theorem. We

defer the task of describing a mechanism that satisfies the definition (or even proving that such a mechanism exists) until Lemma 3.5 below.

Definition 3.3. For $z \in \{0, 1\}$, a mechanism $\mathcal{M}$ mapping $\{0, 1\}^n$ to outcome set $\mathcal{O}$ is called *maximally z-biased*, with respect to a vector of differential privacy parameters $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$, if there exists a set of outcomes $S \subset \mathcal{O}$ such that $\Pr(\mathcal{M}(\mathbf{x}) \in S) \propto \prod_{i=1}^n e^{-\epsilon_i |x_i - z|}$ for all $\mathbf{x} \in \{0, 1\}^n$. In this case, we call S a *distinguished outcome set* for $\mathcal{M}$.

Theorem 3.4. *Suppose the joint distribution μ satisfies positive affiliation. Then for any $z \in \{0, 1\}$ and any vector of differential privacy parameters, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$, the maximum of the ratio*

$$\frac{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid x_a = z)}{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid x_a \neq z)}, \tag{5}$$

over all $\boldsymbol{\epsilon}$-differentially private mechanisms $\mathcal{M}$ and outcome sets S, is attained when $\mathcal{M}$ is maximally z-biased, with distinguished outcome set S. Therefore, the inferential privacy guarantee to individual a in the presence of correlation structure μ and differential privacy parameters $\epsilon_1, \ldots, \epsilon_n$ is given by the formula

$$\nu_a = \max_{z \in \{0,1\}} \left\{ \ln \left| \frac{\sum_{\mathbf{x}=(z,\mathbf{y})} \mu^z(\mathbf{y}) \exp\left(-\sum_{i=1}^n \epsilon_i |x_i - z|\right)}{\sum_{\mathbf{x}=(1-z,\mathbf{y})} \mu^{1-z}(\mathbf{y}) \exp\left(-\sum_{i=1}^n \epsilon_i |x_i - z|\right)} \right| \right\}. \tag{6}$$

Proof. Suppose $z = 0$ and consider any $\boldsymbol{\epsilon}$-differentially private mechanism $\mathcal{M}$ and outcome set S. Letting $p(\mathbf{x}) = \Pr(\mathcal{M}(\mathbf{x}) \in S)$, we have the identity

$$\frac{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid x_a = 0)}{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid x_a = 1)} = \frac{\Pr(\mathcal{M}(\mathbf{x}) \in S \text{ and } x_a = 0)}{\Pr(\mathcal{M}(\mathbf{x}) \in S \text{ and } x_a = 1)} \bigg/ \frac{\Pr(x_a = 0)}{\Pr(x_a = 1)} = \frac{\sum_{\mathbf{x}=(0,\mathbf{y})} \mu^0(\mathbf{y}) p(\mathbf{x})}{\sum_{\mathbf{x}=(1,\mathbf{y})} \mu^1(\mathbf{y}) p(\mathbf{x})}. \tag{7}$$

When $\mathcal{M}$ is maximially 0-biased, with distinguished outcome set S, the right side of (7) is equal to $\frac{\sum_{\mathbf{x}=(0,\mathbf{y})} \mu^0(\mathbf{y}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}}}{\sum_{\mathbf{x}=(1,\mathbf{y})} \mu^1(\mathbf{y}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}}}$. Thus, the $z = 0$ case of the theorem is equivalent to the assertion that

$$\frac{\sum_{\mathbf{x}=(0,\mathbf{y})} \mu^0(\mathbf{y}) p(\mathbf{x})}{\sum_{\mathbf{x}=(1,\mathbf{y})} \mu^1(\mathbf{y}) p(\mathbf{x})} \leq \frac{\sum_{\mathbf{x}=(0,\mathbf{y})} \mu^0(\mathbf{y}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}}}{\sum_{\mathbf{x}=(1,\mathbf{y})} \mu^1(\mathbf{y}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}}}. \tag{8}$$

After cross-multiplying and simplifying, this becomes

$$\left[\sum_{\mathbf{x}=(0,\mathbf{y})} \mu(\mathbf{x}) p(\mathbf{x}) \right] \cdot \left[\sum_{\mathbf{x}=(1,\mathbf{y})} \mu(\mathbf{x}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}} \right] \leq \left[\sum_{\mathbf{x}=(1,\mathbf{y})} \mu(\mathbf{x}) p(\mathbf{x}) \right] \cdot \left[\sum_{\mathbf{x}=(0,\mathbf{y})} \mu(\mathbf{x}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}} \right]. \tag{9}$$

If we add $\left[\sum_{\mathbf{x}=(1,\mathbf{y})} \mu(\mathbf{x}) p(\mathbf{x}) \right] \cdot \left[\sum_{\mathbf{x}=(1,\mathbf{y})} \mu(\mathbf{x}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}} \right]$ to both sides, we find that (9) is equivalent to

$$\left[\sum_{\mathbf{x} \in \{0,1\}^n} \mu(\mathbf{x}) p(\mathbf{x}) \right] \cdot \left[\sum_{\mathbf{x}=(1,\mathbf{y})} \mu(\mathbf{x}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}} \right] \leq \left[\sum_{\mathbf{x}=(1,\mathbf{y})} \mu(\mathbf{x}) p(\mathbf{x}) \right] \cdot \left[\sum_{\mathbf{x} \in \{0,1\}^n} \mu(\mathbf{x}) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}} \right]. \tag{10}$$

To prove (10) we will apply the FKG inequality. Set $h(\mathbf{x}) = \mu(x) e^{-\boldsymbol{\epsilon} \cdot \mathbf{x}}$ and note that $\log h$ is the sum of $\log \mu$—a supermodular function—and $(-\boldsymbol{\epsilon}) \cdot \mathbf{x}$, a linear function. Hence $\log h$ is supermodular. Now define $f(\mathbf{x}) = p(\mathbf{x}) e^{\boldsymbol{\epsilon} \cdot \mathbf{x}}$ and $g(\mathbf{x}) = x_a$. The differential privacy constraint for p implies that f is monotonically non-decreasing; observe that g is monotonically non-decreasing as well. The FKG inequality implies

$$\left[\sum_{\mathbf{x}} f(\mathbf{x}) h(\mathbf{x}) \right] \left[\sum_{\mathbf{x}} g(\mathbf{x}) h(\mathbf{x}) \right] \leq \left[\sum_{\mathbf{x}} f(\mathbf{x}) g(\mathbf{x}) h(\mathbf{x}) \right] \left[\sum_{\mathbf{x}} h(\mathbf{x}) \right]. \tag{11}$$

Substituting the definitions of f, g, h into (11) we readily see that it is equivalent to (10), which completes the proof. $\qquad \square$

Finally, as promised at the start of this section, we show that a noisy-sum mechanism that adds Laplace noise to the sum of the bits in the database is maximally z-biased for every $z \in \{0, 1\}$. Together with Theorem 3.4, this shows that any inferential privacy guarantee that can be proven for the noisy-sum mechanism automatically extends to a guarantee for all differentially private mechanisms, when data are positively affiliated.

Lemma 3.5. *Suppose that all individuals have the same differential privacy parameter, i.e. that $\boldsymbol{\epsilon} = (\epsilon, \epsilon, \dots, \epsilon)$ for some $\epsilon > 0$. Consider the* noisy-sum *mechanism $\mathcal{NS}$ that samples a random Y from the Laplace distribution with scale parameter $1/\epsilon$ and outputs the sum $Y + \sum_{i=1}^{n} x_i$. For all $z \in \{0, 1\}$ the mechanism $\mathcal{NS}$ is maximally z-biased.*

Proof. For any $\mathbf{x} \in \{0, 1\}^n$, let $|\mathbf{x}| = \sum_{i=1}^{n} x_i$. When $z = 0$ and $\boldsymbol{\epsilon} = (\epsilon, \epsilon, \dots, \epsilon)$, the definition of a maximally z-biased mechanism requires the existence of an outcome set S such that $\Pr(\mathcal{NS}(\mathbf{x}) \in S) \propto e^{-\epsilon|\mathbf{x}|}$. For the set $S = (-\infty, 0]$, the event $\mathcal{NS}(\mathbf{x}) \in S$ coincides with the event $Y \leq -|x|$. Since Y is a Laplace random variable with scale parameter $1/\epsilon$, this event has probability proportional to $e^{-\epsilon|x|}$, as desired. When $z = 1$ the proof of the lemma proceeds identically, using the set $S = [n, \infty)$. $\qquad\square$

Remark 3.6. Intuitively, one might expect Theorem 3.4 to hold whenever the joint distribution μ is such that each pair of bits is positively correlated, a weaker property than positive affiliation which requires each pair of bits to be positively correlated *even after conditioning on any possible tuple of values for the remaining bits*. In Appendix A.1 we present an example illustrating that the theorem's conclusion can be violated (in fact, quite drastically violated) when one only assumes pairwise positive correlation. The basic reason is that when bits are pairwise positively correlated, it may still be the case that one individual's bit correlates much more strongly with a non-monotone function of the others' bits than with any monotone function.

Remark 3.7. The quantities appearing in Theorem 3.4 have precise analogues in the physics of spin systems, and this analogy sheds light on inferential privacy. Appendix A.2 delves into this connection in detail; in this remark we merely sketch a dictionary for translating between inferential privacy and statistical mechanics and discuss some consequences of this translation.

In brief, an adversary's prior distribution on $\{0, 1\}^n$ corresponds to the Gibbs measure of a two-spin system with Hamiltonian $H(\mathbf{x}) = -\ln \mu(\mathbf{x})$. Under this correspondence, positively affiliated distributions correspond to ferromagnetic spin systems. The adversary's posterior distribution after applying a maximally 0-biased (resp., maximally 1-biased) mechanism is equivalent to the Gibbs measure of the spin system after applying the external field $\frac{1}{2}\epsilon$ (resp., $-\frac{1}{2}\epsilon$). The worst-case inferential privacy guarantee for Athena in Theorem 3.4 is therefore equivalent (up to a bijective transformation) to the magnetization at Athena's site when the external field $\pm\epsilon$ is applied to the spin system.

One of the interesting implications of this correspondence concerns phase transitions. Statistical-mechanical systems such as magnets are known to undergo sharp transitions in their physical properties as one varies thermodynamic quantities such as temperature and external field strength. Translating these results from physics to the world of privacy using the dictionary outlined above, one discovers that inferential privacy guarantees can undergo surprisingly sharp variations as one varies a mechanism's differential privacy parameter or an adversary's belief about the strength of correlations between individuals' bits in a database. Theorem A.2 in the appendix formalizes these observations about phase transitions in inferential privacy.

4 Bounded Affiliation Distributions

In this section we present a general upper bound for inferential privacy that applies under a condition that we call *bounded affiliation*. Roughly speaking, bounded affiliation requires that correlations between individuals are sufficiently weak, in the sense that the *combined influence* of all other individuals on any particular

8

one is sufficiently small. A very similar criterion in the statistical mechanics literature, *Dobrushin's uniqueness condition* [5, 6], is identical to ours except that it defines "influence" in terms of additive approximation and we define it multiplicatively (Theorem 4.1). Dobrushin showed that this condition implies uniqueness of the Gibbs measure for a specified collection of conditional distributions. Its implications for correlation decay [14, 11, 21] and mixing times of Markov chains [1, 25, 15] were subsequently explored. Indeed, our proof of network differential privacy under the assumption of bounded affiliation draws heavily upon the methods of Dobrushin [6], Gross [14], and Künsch [21] on decay of correlations under Dobrushin's uniqueness condition.

Throughout this section (and its corresponding appendix) we assume that each individual's private data belongs to a finite set X rather than restricting to $X = \{0, 1\}$. This assumption does not add any complication to the theorem statements and proofs, while giving our results much greater generality. We now define the notion of influence that is relevant to our results on distributions with bounded affiliation.

Definition 4.1. If $x_0, \ldots, x_n$ are jointly distributed random variables, the *multiplicative influence* of x_j on x_i, denoted by γ_{ij}, is defined by the equation

$$e^{2\gamma_{ij}} = \max \left\{ \left. \frac{\Pr(x_i \in S \mid \mathbf{x}_{-i})}{\Pr(x_i \in S \mid \mathbf{x'}_{-i})} \right| S \subseteq \operatorname{supp}(x_i), \mathbf{x}_{-i} \sim_j \mathbf{x'}_{-i} \right\}.$$

In other words, the influence of x_j on x_i is one-half of the (individual) differential privacy parameter of x_i with respect to x_j, when one regards x_i as a randomized function of the database $\mathbf{x}_{-i}$. When $i = j$ one adopts the convention that $\gamma_{ij} = 0$. The *multiplicative influence matrix* is the matrix $\Gamma = (\gamma_{ij})$.

Theorem 4.2. *Suppose that the joint distribution μ has a multiplicative influence matrix Γ whose spectral norm is strictly less than 1. Let $\Phi = (\phi_{ij})$ denote the matrix inverse of $I - \Gamma$. Then for any mechanism with individual privacy parameters $\boldsymbol{\epsilon} = (\epsilon_i)$, the inferential privacy guarantee satisfies*

$$\forall i \ \ \nu_i \leq 2 \sum_{j=1}^{n} \phi_{ij} \epsilon_j. \tag{12}$$

If the matrix of multiplicative influences satisfies $\forall i \ \sum_{j=1}^{n} \gamma_{ij} \epsilon_j \leq (1-\delta)\epsilon_i$ for some $\delta > 0$, then $\nu_i \leq 2\epsilon_i/\delta$ for all i.

Proof sketch. Let S be any set of potential outcomes of the mechanism $\mathcal{M}$ such that $\Pr(\mathcal{M}(\mathbf{x}) \in S) > 0$. Let π^1 denote the conditional distribution on databases $\mathbf{x} \in X^n$, given that $\mathcal{M}(\mathbf{x}) \in S$, and let π^2 denote the unconditional distribution μ, respectively. For $i \in \{1, 2\}$ and for any function $f : X^n \to \mathbb{R}$, let $\pi^i(f)$ denote the expected value of f under distribution π^i. Also define the Lipschitz constants $\rho_i(f) = \max\{f(\mathbf{x}) - f(\mathbf{x'}) \mid \mathbf{x} \sim_i \mathbf{x'}\}$. The heart of the proof lies in showing that if f takes values in $\mathbb{R}_+$ then

$$|\ln \pi^1(f) - \ln \pi^2(f)| \leq \frac{1}{2} \sum_{i,j=1}^{n} \Phi_{ij} \epsilon_j \rho_i(\ln f). \tag{13}$$

This is done by studying the set of all vectors $\boldsymbol{\kappa}$ that satisfy $|\ln \pi^1(f) - \ln \pi^2(f)| \leq \sum_{i=1}^{n} \kappa_i \rho_i(f)$ for all f, and showing that this set is non-empty and is preserved by an affine transformation T that is a contracting mapping of $\mathbb{R}^n$ (when the spectral norm of Γ is less than 1) with fixed point $\frac{1}{2}\Phi\boldsymbol{\epsilon}$. To derive (12) from (13), use the definition of ν_i to choose two distinct values $z_0 \neq z_1$ in X such that $\nu_i = \left|\ln\left(\frac{\Pr(x_i=z_1|\mathcal{M}(\mathbf{x})\in S) \,/\, \Pr(x_i=z_0|\mathcal{M}(\mathbf{x})\in S)}{\Pr(x_i=z_1) \,/\, \Pr(x_i=z_0)}\right)\right| = \left|\ln\left(\frac{\pi^1(f)/\pi^1(g)}{\pi^2(f)/\pi^2(g)}\right)\right|$, where f, g are the indicator functions of $x_i = z_0$ and $x_i = z_1$, respectively. Unfortunately $\rho_i(\ln f) = \rho_i(\ln g) = \infty$ so direct application of (13) is not useful; instead, we define a suitable averaging operator τ to smooth out f and g, thereby improving their Lipschitz constants and enabling application of (13). A separate argument is then used to

bound the error introduced by smoothing f and g using τ, which completes the proof of (12). Under the hypothesis that $\Gamma \epsilon \preceq (1 - \delta)\epsilon$, the relation $\boldsymbol{\nu} \preceq \frac{2}{\delta}\boldsymbol{\epsilon}$ is easily derived from (12) by applying the formula $\Phi = \sum_{m=0}^{\infty} \Gamma^m$. The full proof is presented in Appendix B. $\qquad\square$

The bound $\nu_i \leq 2\epsilon_i/\delta$ in the theorem is tight up to a constant factor. This is shown in §A.2 by considering an adversary whose prior is the Ising model of a complete d-ary tree T at inverse temperature $\beta = \tanh^{-1}\left(\frac{1-\delta}{d}\right)$ The entries of the influence matrix satisfy $\gamma_{ij} = \beta$ if $(i, j) \in E(T)$, 0 otherwise. Thus, the row sum $\sum_{j=1}^{n} \gamma_{ij}$ is maximized when i is an internal node, with degree $d + 1$, in which case the row sum is $(d + 1)\tanh^{-1}\left(\frac{1-\delta}{d}\right) = 1 - \delta - o(1)$ as $d \to \infty$. In §A.2 we apply Theorem 3.4 to show that the inferential privacy guarantee for the Ising model on a tree satisfies $\nu = \Omega(\frac{\epsilon}{\delta})$, matching the upper bound in Theorem 4.2 up to a constant factor.

5 Conclusion

A number of immediate questions are prompted by our results, such as incorporating (ε, δ)-privacy into our analysis of inferential guarantees (for product distributions this was achieved in [18]) and extending the analysis in §3 to non-binary databases where an individual's data cannot be summarized by a single bit. A key challenge here is to find an analogue of positive affiliation for databases whose rows cannot naturally be interpreted as elements of a lattice. More excitingly, however, the scenario of datasets with networked correlations raises several broad directions for future work.

Designing for inferential privacy: Our work takes differentially private algorithms as a primitive and *analyzes* what inferential privacy is achievable with given differential privacy guarantees. This allows leveraging the vast body of work on, and adoption of, differentially private algorithms, while remaining agnostic to the data analyst's objective or utility function. However if one instead assumes a particular measure of utility, one can directly investigate the design of inferential-privacy preserving algorithms to obtain stronger guarantees: given some joint distribution(s) and utility objectives, what is the best inferential privacy achievable, and what algorithms achieve it?

Inferential privacy and network structure: An intriguing set of questions arises from returning to the original network structures that led to the model of correlated joint distributions. Note that our results in Theorem 3.4 give the inferential privacy guarantee for a particular individual: how do inferential privacy guarantees depend on the *position* of an individual in the network (for instance, imagine the central individual in a large star graph versus the leaf nodes), and how does the relation between the correlations and the network structure play in?

References

[1] Aizenman, M. and Holley, R. (1987). Rapid convergence to equilibrium of stochastic Ising models in the Dobrushin-Shlosman regime. In Kesten, H., editor, *Percolation Theory and Ergodic Theory of Infinite Particle Systems (Minneapolis, Minn., 1984)*, volume 8 of *IMS Volumes in Math. and Appl.*, pages 1–11. Springer.

[2] Bassily, R., Groce, A., Katz, J., and Smith, A. (2013). Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pages 439–448. IEEE.

[3] Bhaskar, R., Bhowmick, A., Goyal, V., Laxman, S., and Thakurta, A. (2011). Noiseless database privacy. In *Advances in Cryptology–ASIACRYPT 2011*, pages 215–232. Springer.

[4] Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, 15(429-444):2–1.

[5] Dobrushin, R. L. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory of Probability and Its Applications*, 13(2):197–224.

[6] Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486.

[7] Dwork, C. (2006). Differential privacy. In *Automata, Languages and Programming (ICALP)*.

[8] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06.

[9] Dwork, C. and Naor, M. (2008). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1):8.

[10] Dwork, C. and Roth, A. (2013). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*.

[11] Föllmer, H. (1982). A covariance estimate for Gibbs measures. *Journal of Functional Analysis*, 46:387–395.

[12] Fortuin, C., Kasteleyn, P., and Ginibre, J. (1971). Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103.

[13] Gehrke, J., Lui, E., and Pass, R. (2011). Towards privacy for social networks: A zero-knowledge based definition of privacy. In *Theory of Cryptography*, pages 432–449. Springer.

[14] Gross, L. (1979). Decay of correlations in classical lattice models at high temperature. *Communications in Mathematical Physics*, 68(1):9–27.

[15] Hayes, T. P. (2006). A simple condition implying rapid mixing of single-site dynamics on spin systems. In *Proc. 47th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 39–46. IEEE.

[16] Huang, Z. and Kannan, S. (2012). The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 140–149. IEEE.

[17] Kasiviswanathan, S. P., Nissim, K., Raskhodnikova, S., and Smith, A. (2013). Analyzing graphs with node differential privacy. In *Proceedings of the 10th Theory of Cryptography Conference on Theory of Cryptography*, TCC'13.

[18] Kasiviswanathan, S. P. and Smith, A. (2014). On the 'semantics' of differential privacy: A Bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1):1.

[19] Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204. ACM.

[20] Kifer, D. and Machanavajjhala, A. (2014). Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):3.

[21] Künsch, H. (1982). Decay of correlations under Dobrushin's uniqueness condition and its applications. *Communications in Mathematical Physics*, 84(2):207–222.

[22] Levy, K. and boyd, d. (2014). Networked rights and networked harms. working paper, presented at Privacy Law School Conference (June 6, 2014) and Data & Discrimination (May 14, 2014).

[23] McSherry, F. and Talwar, K. (2007). Mechanism design via differential privacy. In *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*, pages 94–103. IEEE.

[24] Milgrom, P. R. and Weber, R. J. (1982). A Theory of Auctions and Competitive Bidding. *Econometrica*, 50(5).

[25] Weitz, D. (2005). Combinatorial criteria for uniqueness of Gibbs measures. *Random Structures & Algorithms*, 27(4):445–475.

A   Appendix to §3: Positively Affiliated Distributions

This appendix contains material accompanying §3 that was omitted from that section for space reasons.

A.1   Pairwise positive correlation

A weaker condition than positive affiliation is *pairwise positive correlation*. This property of a joint distribution μ on databases $\mathbf{x} \in \{0,1\}^n$ requires that for each pair of indices $i, j \in [n]$, the (unconditional) marginal distribution of the bits x_i, x_j satisfies

$$\mathbb{E}[x_i x_j] \geq \mathbb{E}[x_i] \cdot \mathbb{E}[x_j].$$

If the inequality is strict for every i, j then we say μ is *pairwise strictly positively correlated.*

Recall Theorem 3.4, which establishes that when a joint distribution μ satisfies positive affiliation then the worst-case inferential privacy guarantee is attained by any maximally z-biased distribution. The intuition supporting the theorem statement might seem to suggest that the same conclusion holds whenever μ satisfies pairwise positive correlation. In this section we show that this is not the case: if μ satisfies pairwise positive correlation (or even strict pairwise positive correlation) there may be a mechanism whose inferential privacy guarantee is much worse than that of any maximally z-biased mechanism.

Our construction applies when n is of the form $n = 1 + rs$ for two positive integers r, s. For a database $\mathbf{x} \in \{0,1\}^n$ we will denote one of its entries by x_a and the others by x_{ij} for $(i,j) \in [r] \times [s]$. The joint distribution μ is uniform over the solution set of the system of congruences

$$x_a + \sum_{j=1}^{s} x_{ij} \equiv 0 \pmod{2} \qquad \text{for } i = 1, \ldots, r \tag{14}$$

Thus, to sample from μ one draws the bits x_a and x_{ij} for $(i,j) \in [r] \times [s-1]$ independently from the uniform distribution on $\{0,1\}$, then one sets x_{is} for all i so as to satisfy (14).

The distribution μ is pairwise independent, hence it is pairwise positively correlated. (The calculation of privacy parameters is much easier in the pairwise-independent case. At the end of this section we apply a simple continuity argument to modify the example to one with pairwise strict positive correlation without significantly changing the privacy parameters.)

Let us first calculate the inferential privacy for a mechanism $\mathcal{M}_1$ that calculates the number of odd integers in the sequence

$$x_a, \sum_{j=1}^{s} x_{1j}, \sum_{j=1}^{s} x_{2j}, \ldots, \sum_{j=1}^{s} x_{rj} \tag{15}$$

and adds Laplace noise (with scale parameter $1/\epsilon$) to the result. This is ϵ-differentially private since changing a single bit of $\mathbf{x}$ changes the parity of only one element of the sequence. However, when $\mathbf{x}$ is sampled from μ the number of odd integers in the sequence (15) is either 0 if $x_a = 0$ or n if $x_a = 1$. Hence

$$\Pr(\mathcal{M}_1(\mathbf{x}) \leq 0 \mid x_a = 0) = \tfrac{1}{2}$$
$$\Pr(\mathcal{M}_1(\mathbf{x}) \leq 0 \mid x_a = 1) = \tfrac{1}{2} e^{-(r+1)\epsilon}$$

implying that the inferential privacy parameter of $\mathcal{M}_1$ is at least $(r+1)\epsilon$.

Now let us calculate the inferential privacy parameter of a maximally 0-biased mechanism $\mathcal{M}_2$, with outcome $o \in \mathcal{O}$ such that $\Pr(\mathcal{M}_2(\mathbf{x}) = o \mid \mathbf{x}) \propto e^{-\epsilon|\mathbf{x}|}$, where $|\mathbf{x}|$ denotes the sum of the bits in $\mathbf{x}$. Let T_0 (resp. T_1) denote the set of bit-strings in $\{0,1\}^s$ having even (resp. odd) sum, and let T_0^r, T_1^r denote the rth Cartesian powers of these sets. The conditional distribution of (x_{ij}) given $x_a = 0$ is the uniform distribution on T_0^r, and the conditional distribution of (x_{ij}) given $x_a = 1$ is the uniform distribution on T_1^r. For $\mathbf{y} = (y_{ij}) \in \{0,1\}^{rs}$ and $i \in [r]$, let $\mathbf{y}_{i*}$ denote the s-tuple $(y_{i1}, \ldots, y_{is})$. We have

$$\begin{aligned}
\Pr(\mathcal{M}_2(\mathbf{x}) = o \mid x_a = 0) &= \sum_{\mathbf{x}=(0,\mathbf{y})} \Pr(\mathcal{M}_2(\mathbf{x}) = o \mid \mathbf{x}) \cdot \Pr(\mathbf{x} \mid x_a = 0) \\
&= \sum_{\mathbf{y} \in T_0^r} e^{-\epsilon|\mathbf{y}|} \cdot 2^{-r(s-1)} \\
&= \sum_{\mathbf{y} \in T_0^r} \prod_{i=1}^{r} \left(e^{-\epsilon|\mathbf{y}_{i*}|} \cdot 2^{1-s} \right) \\
&= \left(2^{(1-s)} \sum_{\mathbf{z} \in T_0} e^{-\epsilon|\mathbf{z}|} \right)^r.
\end{aligned} \tag{16}$$

Similarly,

$$\Pr(\mathcal{M}_2(\mathbf{x}) = o \mid x_a = 1) = e^{-\epsilon} \cdot \left(2^{(1-s)} \sum_{\mathbf{z} \in T_1} e^{-\epsilon|\mathbf{z}|} \right)^r. \tag{17}$$

(The extra factor of $e^{-\epsilon}$ on the right side comes from the fact that $x_a = 1$, which inflates the exponent in the expression $e^{-\epsilon|\mathbf{x}|}$ by ϵ.) To evaluate the expressions on the right sides of (16)-(17), it is useful to let

$A_0 = \sum_{z \in T_0} e^{-\epsilon|z|}$ and $A_1 = \sum_{z \in T_1} e^{-\epsilon|z|}$. Then we find that

$$A_0 + A_1 = \sum_{z \in \{0,1\}^s} e^{-\epsilon|z|} = (1 + e^{-\epsilon})^s$$

$$A_0 - A_1 = \sum_{z \in \{0,1\}^s} (-1)^{|z|} \cdot e^{-\epsilon|z|} = (1 - e^{-\epsilon})^s$$

$$A_0 = \frac{1}{2}\left[(1 + e^{-\epsilon})^s + (1 - e^{-\epsilon})^s\right]$$

$$A_1 = \frac{1}{2}\left[(1 + e^{-\epsilon})^s - (1 - e^{-\epsilon})^s\right].$$

Substituting these expressions into (16)-(17) we may conclude that

$$\frac{\Pr(\mathcal{M}_2(\mathbf{x}) = o \mid x_a = 0)}{\Pr(\mathcal{M}_2(\mathbf{x}) = o \mid x_a = 1)} = \frac{e^\epsilon \cdot A_0^r}{A_1^r} = e^\epsilon \left[\frac{(1 + e^{-\epsilon})^s + (1 - e^{-\epsilon})^s}{(1 + e^{-\epsilon})^s - (1 - e^{-\epsilon})^s}\right]^r. \tag{18}$$

The inferential privacy parameter of $\mathcal{M}_2$ is therefore given by

$$\nu = \epsilon + r \ln\left[\frac{(1 + e^{-\epsilon})^s + (1 - e^{-\epsilon})^s}{(1 + e^{-\epsilon})^s - (1 - e^{-\epsilon})^s}\right]$$

$$< \epsilon + r \ln\left[1 + 2\left(\frac{1 - e^{-\epsilon}}{1 + e^{-\epsilon}}\right)^s\right]$$

$$= \epsilon + r \ln\left[1 + 2\tanh^s(\epsilon)\right] < \epsilon + 2r\epsilon^s.$$

Comparing the inferential privacy parameters of $\mathcal{M}_1$ and $\mathcal{M}_2$, they are $(r+1)\epsilon$ and $\epsilon + 2r\epsilon^s$, respectively, so the inferential privacy parameter of $\mathcal{M}_1$ exceeds that of the maximally 0-biased mechanism, $\mathcal{M}_2$, by an unbounded factor as $r, s \to \infty$.

Under the distribution μ we have analyzed thus far, the bits of $\mathbf{x}$ are pairwise independent. However, we may take a convex combination of μ with any distribution in which all pairs of bits are strictly positively correlated—for example, a distribution that assigns equal probability to the two databases $(1, \ldots, 1)$ and $(0, \ldots, 0)$ and zero probability to all others. In this way we obtain a distribution μ' which satisfies pairwise strict positive correlation and may can be made arbitrarily close to μ by varying the mixture parameter of the convex combination. Since the inferential privacy parameter of a mechanism with respect to a given prior distribution is a continuous function of that distribution, it follows that the inferential privacy parameters of $\mathcal{M}_1$ and $\mathcal{M}_2$ can remain arbitrarily close to the values calculated above while imposing a requirement that the prior on $\mathbf{x}$ satisfies pairwise strict positive correlation.

A.2 Connection to Ferromagnetic Spin Systems

The quantities appearing in Theorem 3.4 have precise analogues in the physics of spin systems, and this analogy sheds light on inferential privacy. In statistical mechanics, a two-spin system composed of n sites has a state space $\{\pm 1\}^n$ and an *energy function* or *Hamiltonian*, $H : \{\pm 1\}^n \to \mathbb{R}$. The Gibbs measure of the spin system is a probability distribution assigning to each state a probability proportional to $e^{-\beta H(\boldsymbol{\sigma})}$ where $\beta > 0$ is a parameter called the *inverse temperature*. Application of an *external field* $\boldsymbol{h} \in \mathbb{R}^n$ to the spin system is modeled by subtracting a linear function from the Hamiltonian, so that it becomes $H(\boldsymbol{\sigma}) - \boldsymbol{h} \cdot \boldsymbol{\sigma}$. The probability of state $\boldsymbol{\sigma}$ under the Gibbs measure then becomes

$$\Pr(\boldsymbol{\sigma}) = e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]}/Z(\beta, \boldsymbol{h}),$$

where $Z(\cdot)$ is the *partition function*

$$Z(\beta, \boldsymbol{h}) = \sum_{\boldsymbol{\sigma} \in \{\pm 1\}^n} e^{\beta[\sum_i h_i \sigma_i - H(\boldsymbol{\sigma})]}.$$

Databases $\mathbf{x} \in \{0, 1\}^n$ are in one-to-one correspondence with states $\boldsymbol{\sigma} \in \{\pm 1\}^n$ under the mapping $\sigma_i = (-1)^{x_i}$ and its inverse mapping $x_i = \frac{1}{2}(1 - \sigma)$. Any joint distribution $\mu = \{0, 1\}^n$ has a corresponding Hamiltonian $H(\boldsymbol{\sigma}) = -\ln \mu(\mathbf{x})$ whose Gibbs distribution (at $\beta = 1$) equals μ. The positive affiliation condition is equivalent to requiring that H is submodular, a property which is expressed by saying that the spin system is *ferromagnetic*.

For a maximally 0-biased mechanism $\mathcal{M}$ with distinguished outcome set S, the probabilities $p(\mathbf{x}) = \Pr(\mathcal{M}(\mathbf{x}) \in S)$ satisfy $p(\mathbf{x}) \propto \exp(-\sum_{i=1}^n \epsilon_i x_i) = \exp(-\frac{n}{2} + \frac{1}{2}\sum_i \epsilon_i \sigma_i)$, so

$$\mu(\mathbf{x})p(\mathbf{x}) \propto e^{-\frac{n}{2} + \frac{1}{2}\sum_i \epsilon_i \sigma_i - H(\boldsymbol{\sigma})}.$$

Application of the mechanism $\mathcal{M}$ is thus analogous to application of the external field $\frac{1}{2}\epsilon$ at inverse temperature 1. (The additive constant $-\frac{n}{2}$ in the Hamiltonian is irrelevant, since the Gibbs measure is unchanged by an additive shift in the Hamiltonian.) Similarly, applying a maximally 1-biased mechanism is analogous to applying the external field $-\frac{1}{2}\epsilon$ at inverse temperature 1.

Let $\rho = \frac{\mu(x_a = 1)}{\mu(x_a = 0)}$ denote the prior probability ratio for Athena's bit. For the networked privacy guarantee in Theorem 3.4, when the maximum on the right side of (6) is achieved by a maximally 0-biased mechanism, we have

$$\frac{e^{\nu_a} - \rho}{e^{\nu_a} + \rho} = \frac{\sum_{\mathbf{x}=(0,\mathbf{y})} \mu(\mathbf{x})p(\mathbf{x}) - \sum_{\mathbf{x}=(1,\mathbf{y})} \mu(\mathbf{x})p(\mathbf{x})}{\sum_{\mathbf{x}=(0,\mathbf{y})} \mu(\mathbf{x})p(\mathbf{x}) + \sum_{\mathbf{x}=(1,\mathbf{y})} \mu(\mathbf{x})p(\mathbf{x})} = \frac{\sum_{\boldsymbol{\sigma}} \sigma_a e^{\boldsymbol{\epsilon}\cdot\boldsymbol{\sigma}/2 - H(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma}} e^{\boldsymbol{\epsilon}\cdot\boldsymbol{\sigma}/2 - H(\boldsymbol{\sigma})}} = \mathbb{E}[\sigma_a \mid \boldsymbol{h} = \tfrac{\boldsymbol{\epsilon}}{2}],$$

where the operator $\mathbb{E}[\cdot \mid \boldsymbol{h} = \frac{\boldsymbol{\epsilon}}{2}]$ denotes the expectation under the Gibbs measure corresponding to external field $\boldsymbol{h} = \frac{\boldsymbol{\epsilon}}{2}$. A similar calculation in the case that a maximally 1-biased mechanism maximizes the right side of (6) yields the relation $\frac{e^{\nu_a} - \rho^{-1}}{e^{\nu_a} - \rho^{-1}} = \mathbb{E}[-\sigma_a \mid \boldsymbol{h} = -\frac{\boldsymbol{\epsilon}}{2}]$. Combining these two cases, we arrive at:

$$\nu_a = \max \left\{ \ln \rho + \ln \left(\frac{1 + \mathbb{E}[\sigma_a \mid \boldsymbol{h} = \boldsymbol{\epsilon}/2]}{1 - \mathbb{E}[\sigma_a \mid \boldsymbol{h} = \boldsymbol{\epsilon}/2]} \right), \ -\ln \rho - \ln \left(\frac{1 + \mathbb{E}[\sigma_a \mid \boldsymbol{h} = -\boldsymbol{\epsilon}/2]}{1 - \mathbb{E}[\sigma_a \mid \boldsymbol{h} = -\boldsymbol{\epsilon}/2]} \right) \right\}. \qquad (19)$$

We will refer to $\mathbb{E}[\sigma_a]$ as the *magnetization at site a*, by analogy with the usual definition of magnetization in statistical mechanics as the average $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\sigma_i]$. Equation (19) thus shows that the inferential privacy guarantee for a positively affiliated distribution is completely determined by the magnetization at site a when an external field of strength $\pm\epsilon/2$ is applied.

A.2.1 Ising models and phase transitions

Let us now apply this circle of ideas to analyze the "Zeus's family tree" example from §1. Represent Zeus and his progeny as the nodes of a rooted tree, and suppose that the joint distribution of the individuals' bits is defined by the following sampling rule: sample the bits in top-down order (from root to leaves), setting the root's bit to 0 or 1 with equal probability and each other node's bit equal to the parent's value with probability $p > \frac{1}{2}$ and the opposite value otherwise. This leads to a probability distribution μ in which the probability of any $\mathbf{x} \in \{0, 1\}^{V(T)}$ is proportional to $p^{a(\mathbf{x})}(1 - p)^{b(\mathbf{x})}$ where $a(\mathbf{x})$ denotes the number of tree edges whose endpoints receive the same label, and $b(\mathbf{x})$ is the number of edges whose endpoints receive opposite labels. Letting $J = \tanh^{-1}(2p - 1)$ so that $\ln(p) = \ln(1 - p) + 2J$, and associating $\boldsymbol{\sigma} \in \{\pm 1\}^n$ to $\mathbf{x} \in \{0, 1\}^n$ via $\sigma_i = (-1)^{x_i}$ as before, we find that up to an additive constant, $\ln \mu(\mathbf{x}) = J \sum_{(i,j) \in E} \sigma_i \sigma_j$,

where E denotes the edge set of the tree. Hence, the joint distribution of Zeus's family tree is equivalent to the Gibbs measure of the Hamiltonian $H(\boldsymbol{\sigma}) = -J \sum_{(i,j) \in E} \sigma_i \sigma_j$. Models whose Hamiltonian takes this form (for any graph, not just trees) are known as *Ising models* (with *interaction strength J*) and are among the most widely studied in mathematical physics.

Ising models are known to undergo *phase transitions* as one varies the inverse temperature or external field. For example, in an infinite two-dimensional lattice or Δ-regular tree, there is a phenomenon known as *spontaneous magnetization* where the magnetization does not converge to zero as the external field converges to zero from above, but this phenomenon only occurs if the inverse temperature is above a critical value, β_c, that is equal to $\ln(1+\sqrt{2})$ for the two-dimensional lattice and to $\frac{1}{2} \ln(1 + \frac{2}{\Delta - 2})$ for the Δ-regular tree. This phenomenon of phase transitions has consequences for inferential privacy, as articulated in Theorem A.2 below. To state the theorem it is useful to make the following definition.

Definition A.1. Let $\mathscr{D}$ be a family of joint distributions on $\{0,1\}^*$, with each distribution $\mu \in \mathscr{D}$ being supported on $\{0,1\}^n$ for a specific value $n = n(\mu)$. For a differential privacy parameter $\epsilon > 0$, let $\nu(\epsilon, \mathscr{D})$ denote the supremum, over all joint distributions $\mu \in \mathscr{D}$, of the inferential privacy guarantee corresponding to differential privacy parameter ϵ. We say that ν is *differentially enforceable* with respect to $\mathscr{D}$ if there exists $\epsilon > 0$ such that $\nu \leq \nu(\epsilon, \mathscr{D})$.

In other words, to say that ν is differentially enforceable means that a regulator can ensure ν-inferential privacy for the individuals participating in a datasest by mandating that an analyst must satisfy ϵ-differential privacy when releasing the results of an analysis performed on the dataset.

Theorem A.2. *For a family of graphs $\mathscr{G}$ and a given $J > 0$, let $\mathscr{D}$ be the family of Ising models with interaction strength J and zero external field on graphs in $\mathscr{G}$. Then*

a. **(Sensitivity to strength of correlations.)** *If $\mathscr{G}$ is the set of trees of maximum degree $\Delta = d + 1$ and $J = \tanh^{-1}\left(\frac{1-\delta}{d}\right)$ for some $\delta > 0$, then every $\nu > 0$ is differentially enforceable, and in fact $\nu(\epsilon, \mathscr{D}) = \Theta(\epsilon/\delta)$ for $0 < \delta < \epsilon \ll 1$. On the other hand, if $J > \tanh^{-1}\left(\frac{1}{d}\right)$ then the set of all differentially enforceable ν has a strictly positive infimum, $\nu_{\min}(J, \Delta)$.*

b. **(Sensitivity to differential privacy parameter.)** *For any $0 < \epsilon_0 < \epsilon_1$ and any $1 < r < R$, there exists a joint distribution μ whose inferential privacy guarantee satisfies $\nu/\epsilon < r$ when $\epsilon = \epsilon_0$ but $\nu/\epsilon > R$ when $\epsilon = \epsilon_1$.*

Part (b) is particularly striking because it implies, for instance, that when a policy-maker contemplates whether to mandate differential privacy parameter $\epsilon = 0.19$ or $\epsilon = 0.2$, this seemingly inconsequential decision could determine whether the inferential privacy guarantee will be $\nu = 0.2$ or $\nu = 20$.

§A.2.2 is devoted to proving the theorem. The proof combines known facts about phase transitions with some calculations regarding magnetization of Ising models on a tree subjected to an external field.

A.2.2   The Bethe lattice and the proof of Theorem A.2

The infinite Δ-regular tree is known in mathematical physics as the *Bethe lattice with coordination number Δ*. Most of the results stated in Theorem A.2 can be derived by analyzing the Ising model on the Bethe lattice and calculating the magnetization of the root when the lattice is subjected to an external field. Throughout this section, we will use the notation $\langle \sigma_a \rangle$ to denote the expectation of the random variable σ_a under the distribution defined by the Ising model with interaction strength J at inverse temperature 1 and external field h.

Lemma A.3. *For given $J > 0$, $d \geq 2$, and $h \in \mathbb{R}$, define a function $y(x)$ by*

$$y(x) = e^{2h} \left(\frac{e^J x + e^{-J}}{e^J + e^{-J} x} \right)^d.$$

The sequence $x_0, x_1, \ldots$ defined recursively by $x_0 = 1$ and $x_{n+1} = y(x_n)$ for $n \geq 0$ converges to a limit point $x = x(J, h)$. This limit point is determined as follows.

- *If $h = 0$ then $x(J, h) = 1$.*

- *If $h > 0$ then $x(J, h)$ is the unique solution of the equation $x = y(x)$ in the interval $(1, \infty)$.*

- *If $h < 0$ then $x(J, h)$ is the unique solution of the equation $x = y(x)$ in the interval $(0, 1)$.*

The behavior of $x(J, h)$ near $h = 0$ depends on the value of J. If $\tanh(J) < \frac{1}{d}$, then $x(J, h)$ varies continuously with h and $\lim_{h \to 0} x(J, h) = 1$. If $\tanh(J) > \frac{1}{d}$, then the function $x(J, h)$ is discontinuous at $h = 0$, and it satisfies $\lim_{h \searrow 0} x(J, h) > 1$ and $\lim_{h \nearrow 0} x(J, h) < 1$.

Proof. Rewriting the formula for $y(x)$ as

$$y(x) = e^{2(h+Jd)} \left[1 - \frac{e^J - e^{-3J}}{e^J + e^{-J} x} \right]^d$$

it is clear that for $0 \leq x < \infty$, $y(x)$ is continuous and monotonically increasing in x and takes values between $e^{2(h-Jd)}$ and $e^{2(h+Jd)}$. Since y is monotonic, the sequence $x_0, x_1, x_2, \ldots$ defined in the lemma must be monotonic: if $x_0 \leq x_1$ then an easy induction establishes that $x_n \leq x_{n+1}$ for all n, and likewise if $x_0 \geq x_1$ then $x_n \geq x_{n+1}$ for all n. Any monotonic sequence in a closed, bounded interval must converge to a limit, so the limit point $x(J, h)$ is well-defined.

If $h = 0$ then a trivial calculation shows that $x_n = 1$ for all n, and thus $x(J, h) = 1$. For $h > 0$ or $h < 0$ we must show that $x(J, h)$ is the unique solution of $y(x) = x$ in the interval $(1, \infty)$ or $(0, 1)$, respectively. First note that $y(1) = e^{2Jh}$, so the sequence $x_0, x_1, \ldots$ is monotonically increasing when $h > 0$ and decreasing when $h < 0$. Thus $x = x(J, h) = \lim_{n \to \infty} x_n$ belongs to $(1, \infty)$ when $h > 0$ and to $(0, 1)$ when $h < 0$. The continuity of y implies that

$$y(x) = \lim_{n \to \infty} y(x_n) = \lim_{n \to \infty} x_{n+1} = x.$$

Thus, x satisfies $x = y(x)$. It remains to show that this equation has a unique solution in $(1, \infty)$ when $h > 0$ and a unique solution in $(0, 1)$ when $h < 0$.

A solution to $x = y(x)$ is also a solution to $\ln x - \ln y(x) = 0$. The function $g(x) = \ln x - \ln y(x)$ has derivative

$$g'(x) = \frac{1}{x} - \frac{de^J}{e^J x + e^{-J}} + \frac{de^{-J}}{e^J + e^{-J} x} = \frac{1}{x} - \frac{d(e^{2J} - e^{-2J})}{x^2 + (e^{2J} + e^{-2J})x + 1}$$

The equation $g'(x) = 0$ is equivalent to the quadratic equation $x^2 - 2[d \sinh(2J) - \cosh(2J)]x + 1 = 0$. This has at most two real roots, and if it has any real roots at all then all roots are real and their product is equal to 1. Therefore, it has at most one root in the interval $(0, 1)$ and at most one root in the interval $(1, \infty)$. Furthermore, $g'(x)$ is strictly positive at $x = 0$ and as $x \to \infty$. Summarizing this discussion, there exist positive numbers $x_0 \leq x_1$ such that $x_0 \cdot x_1 = 1$ and the set $\{x \mid g'(x) > 0\}$ intersects the intervals $(0, 1)$ and $(1, \infty)$ in the subintervals $(0, x_0)$ and (x_1, ∞), respectively.

Now suppose $h > 0$. The set $\{x \mid x > 1 \text{ and } y(x) = x\}$ is non-empty; for example, it contains $x(J, h)$. Let $x_{\inf}$ denote the infimum of this set. By continuity, $y(x_{\inf}) = x_{\inf}$. Since $g(x_{\inf}) = 0$ whereas

$g(1) < 0$, we must have $g'(z) > 0$ for some z in the interval $(1, x_{\inf})$. Recalling the number x_1 defined in the previous paragraph, we must have $x_1 < z < x_{\inf}$. Consequently g' is strictly positive throughout the interval $(x_{\inf}, \infty)$, implying that there are no other solutions of $g(x) = 0$ in that interval. Thus, $x_{\inf}$ is the unique solution of $y(x_{\inf}) = x_{\inf}$ in $(1, \infty)$, and $x(J, h) = x_{\inf}$. When $h < 0$ an analogous argument using $x_{\sup} = \sup\{x \mid x < 1 \text{ and } y(x) = x\}$ proves that $x(J, h) = x_{\sup}$ is the unique solution of $y(x) = x$ in the interval $(0, 1)$.

To analyze the behavior of $x(J, h)$ near $h = 0$, it is useful to first analyze the zero set of $g'(x)$. Recall that $g'(x) = 0$ if and only if $x^2 - 2[d\sinh(2J) - \cosh(2J)]x + 1 = 0$. The discriminant test tells us that this quadratic equation has zero, one, or two real roots according to whether $d\sinh(2J) - \cosh(2J) - 1$ is less than, equal to, or greater than 0. Using the identities $\sinh(2J) = 2\sinh(J)\cosh(J) = 2\cosh^2(J)\tanh(J)$ and $\cosh(2J) = 2\cosh^2(J) - 1$ we find that $d\sinh(2J) - \cosh(2J) - 1 = 2\cosh^2(J)[d\tanh(J) - 1]$. So when $\tanh(J) > \frac{1}{d}$, $g'(x) > 0$ for all x and the equation $y(x) = x$ has the unique solution $x(J, h)$. Implicit differentiation, applied to the equation $x(J, h) = y(x(J, h))$, yields:

$$\frac{\partial x}{\partial h} = \frac{\partial y}{\partial h} + \frac{\partial y}{\partial x} \cdot \frac{\partial x}{\partial h} \tag{20}$$

which can be rearranged to yield

$$\frac{\partial x}{\partial h} = \frac{\partial y/\partial h}{1 - \partial y/\partial x}. \tag{21}$$

The function y is C^∞ in the region $x > 0$, and at $h = 0, x = 1$ we have

$$\frac{\partial y}{\partial x} = d\tanh(J)$$

$$\frac{\partial y}{\partial h} = 2$$

$$\frac{\partial x}{\partial h} = \frac{2}{1 - d\tanh(J)}, \tag{22}$$

so when $\tanh(J) < \frac{1}{d}$ the implicit function theorem implies that $x(J, h)$ is a differentiable, increasing function of h in a neighborhood of $h = 0$.

When $\tanh(J) > \frac{1}{d}$ and $h = 0$, we have $g(1) = 0$ and $g'(1) < 0$, so for some sufficiently small $\delta > 0$ we have $g(1 + \delta) < 0$, $g(1 - \delta) > 0$. On the other hand, the fact that $\ln y(x)$ is bounded between $-2Jd$ and $2Jd$ implies that $g(x) = \ln x - \ln y(x)$ tends to $-\infty$ as $x \to 0$ and to ∞ as $x \to \infty$. The intermediate value theorem implies that there exist $x_+ \in (1 + \delta, \infty)$ and $x_- \in (0, 1 - \delta)$ such that $g(x_+) = g(x_-) = 0$. In fact, the equation $g(x) = 0$ can have at most three solutions since $g'(x) = 0$ has only two solutions. So, the entire solution set of $g(x) = 0$ is $\{x_-, 1, x_+\}$. Denote the function $y(x)$ in the case $h = 0$ by $y_0(x)$, to distinguish it from the case of general h; similarly define $g_0(x) = \ln x - \ln y_0(x)$. Note that $g_0(x) \leq 0$ when $1 \leq x \leq x_+$, so $x \leq y_0(x)$ on that interval. When $h > 0$ we have $y(x) > y_0(x)$ for all x, hence $y(x) > x$ for $1 \leq x \leq x_+$. As $x(J, h)$ is the unique solution of $y(x) = x$ in the interval $(1, \infty)$ it follows that $x(J, h) > x_+$. On the other hand, for any $\delta > 0$, we have $g_0(x_+ + \delta) > 0$ and hence, for sufficiently small $h > 0$, we also have $g(x_+ + \delta) > 0$. Since $g(x_+) < 0$ and $g(x(J, h)) = 0$, the intermediate value theorem implies $x(J, h)$ belongs to the interval $(x_+, x_+ + \delta)$ for all sufficiently small $h > 0$. In other words, $\lim_{h \searrow 0} x(J, h) = x_+$. The analogous argument for $h < 0$ proves that $x(J, h) < x_-$ and that $\lim_{h \nearrow 0} x(J, h) = x_-$. $\qquad\square$

Lemma A.4. *If T is a subtree of T' and a is any node of T, let $\langle \sigma_a \rangle_T$ and $\langle \sigma_a \rangle_{T'}$ denote the expectation of σ_a in the Ising models on T and T', respectively, with interaction strength $J > 0$. For $h > 0$ we have $\langle \sigma_a \rangle_{T'} \geq \langle \sigma_a \rangle_T$ while for $h < 0$ we have $\langle \sigma_a \rangle_{T'} \leq \langle \sigma_a \rangle_T$.*

Proof. It suffices to prove the lemma in the case that $h > 0$ (since the $h < 0$ case is symmetric under exchanging the signs $+1$ and -1) and that T is obtained from T' by deleting a single leaf node, b. The lemma then follows by induction, since any subtree can be obtained from a tree by successively deleting leaves.

18

Let c denote the parent of b in T', *i.e.*, assume that (b, c) is the unique edge of T' containing b. For state $\boldsymbol{\sigma} \in \{\pm 1\}^{V(T)}$, $(\boldsymbol{\sigma}, +)$ and $(\boldsymbol{\sigma}, -)$ denote the states in $\{\pm 1\}^{V(T')}$ obtained by setting $\sigma_b = +1$ or $\sigma_b = -1$, respectively, while keeping the spin at every node of T the same. If $H(\boldsymbol{\sigma}) = -J \sum_{(i,j) \in E(T)} \sigma_i \sigma_j$ is the Hamiltonian of the Ising model on T, then the Hamiltonian of the Ising model on T' is given by

$$H'(\boldsymbol{\sigma}, +) = -J\sigma_c + H(\boldsymbol{\sigma})$$
$$H'(\boldsymbol{\sigma}, -) = J\sigma_c + H(\boldsymbol{\sigma}).$$

Thus, the partition functions $Z(\beta, \boldsymbol{h})$, $Z'(\beta, \boldsymbol{h})$ of T, T' respectively satisfy

$$Z(\beta, \boldsymbol{h}) = \sum_{\boldsymbol{\sigma}} e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]}$$

$$Z'(\beta, \boldsymbol{h}) = \sum_{\boldsymbol{\sigma}} e^{\beta[\boldsymbol{h} \cdot (\boldsymbol{\sigma}, +) - H'(\boldsymbol{\sigma}, +)]} + e^{\beta[\boldsymbol{h} \cdot (\boldsymbol{\sigma}, -) - H'(\boldsymbol{\sigma}, -)]}$$

$$= 2 \sum_{\boldsymbol{\sigma}} \cosh(\beta[h + J\sigma_c]) \, e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]}.$$

Furthermore, we have

$$\langle \sigma_a \rangle_T = \frac{1}{Z(\beta, \boldsymbol{h})} \sum_{\boldsymbol{\sigma}} \sigma_a \, e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]}$$

$$\langle \sigma_a \rangle_{T'} = \frac{1}{Z'(\beta, \boldsymbol{h})} \sum_{\boldsymbol{\sigma}} \sigma_a \, e^{\beta[\boldsymbol{h} \cdot (\boldsymbol{\sigma}, +) - H'(\boldsymbol{\sigma}, +)]} + e^{\beta[\boldsymbol{h} \cdot (\boldsymbol{\sigma}, -) - H'(\boldsymbol{\sigma}, -)]}$$

$$= \frac{2}{Z'(\beta, \boldsymbol{h})} \sum_{\boldsymbol{\sigma}} \sigma_a \, \cosh(\beta[h + J\sigma_c]) \, e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]}.$$

Associating to each $\mathbf{x} \in \{0, 1\}^n$ a state $\boldsymbol{\sigma}(\mathbf{x}) \in \{\pm 1\}^n$ via $\sigma_i = (-1)^{x_i}$ as before, we find that the logarithm of the function $\mathbf{x} \mapsto e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma}(\mathbf{x}) - H(\boldsymbol{\sigma}(\mathbf{x}))]}$ is supermodular. Furthermore, the functions $\mathbf{x} \mapsto (-1)^{x_a}$ and $\mathbf{x} \mapsto 2\cosh(\beta[h + (-1)^{x_c}J])$ are both monotonically decreasing. Thus, we may apply the FKG inequality (Lemma 3.2) to conclude that

$$\left[\sum_{\boldsymbol{\sigma}} e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]} \right] \left[2 \sum_{\boldsymbol{\sigma}} \sigma_a \cosh(\beta[h + J\sigma_c]) e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]} \right]$$

$$\geq \left[2 \sum_{\boldsymbol{\sigma}} \cosh(\beta[h + J\sigma_c]) \, e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]} \right] \left[\sum_{\boldsymbol{\sigma}} \sigma_a e^{\beta[\boldsymbol{h} \cdot \boldsymbol{\sigma} - H(\boldsymbol{\sigma})]} \right].$$

Dividing both sides by $Z(\beta, \boldsymbol{h}) \cdot Z'(\beta, \boldsymbol{h})$, we obtain the inequality asserted in the lemma. $\qquad \square$

Lemma A.5. *If T is a finite tree of maximum degree $\Delta = d + 1$, a is any node of T, and $\langle \sigma_a \rangle$ denotes the expectation of σ_a in the Ising model on T with interaction strength J, inverse temperature 1, and external field $h > 0$, then*

$$\ln \left(\frac{1 + \langle \sigma_a \rangle}{1 - \langle \sigma_a \rangle} \right) < \frac{d + 1}{d} \ln x(J, h) - \frac{2h}{d}. \tag{23}$$

The difference between the left and right sides converges to zero as the distance from a to the nearest node of degree less than Δ tends to infinity.

Proof. Define a sequence of rooted trees $T_0, T_1, \ldots$ recursively, by stating that T_0 is a single node and T_{n+1} consists of a root joined to $d = \Delta - 1$ children, each of whom is the root of a copy of T_n. Also define a

sequence of trees $T_0^*, T_1^*, \ldots$, by stating that $T_0^* = T_0$ while for $n > 0$, T_n^* consists of a root joined to Δ children, each of whom is the root of a copy of T_{n-1}. (In other words, T_n^* is like T_n, with the root modified to have Δ instead of $\Delta - 1$ children.)

If T is any tree of maximum degree Δ containing a node labeled a, then let r denote the distance from a to the nearest node of degree less than Δ, and let s denote the distance from a to the farthest leaf. We can embed T_r^* as a subtree of T rooted at a, and we can embed T as a subtree of T_s^* with a at the root. Applying Theorem A.4,

$$\langle \sigma_a \rangle_{T_r^*} \leq \langle \sigma_a \rangle_T \leq \langle \sigma_a \rangle_{T_s^*}.$$

To complete the proof we will show that $\ln \left(\frac{1 + \langle \sigma_a \rangle_{T_n^*}}{1 - \langle \sigma_a \rangle_{T_n^*}} \right)$ converges to $\frac{d+1}{d} \ln x(J, h) - \frac{2h}{d}$ (from below) as $n \to \infty$.

For any tree T with root node k, let

$$Z^+(T) = \sum_{\boldsymbol{\sigma}: \sigma_k = +1} e^{J \sum_{i,j} \sigma_i \sigma_j + h \sum_i \sigma_i}$$

$$Z^-(T) = \sum_{\boldsymbol{\sigma}: \sigma_k = -1} e^{J \sum_{i,j} \sigma_i \sigma_j + h \sum_i \sigma_i}$$

We have $\langle \sigma_k \rangle_T = \frac{Z^+(T) - Z^-(T)}{Z^+(T) + Z^-(T)}$, so $\frac{1 + \langle \sigma_k \rangle_T}{1 - \langle \sigma_k \rangle_T} = \frac{Z^+(T)}{Z^-(T)}$. For the tree T_n defined in the preceding paragraph, the quantity $x_n = Z^+(T_n)/Z^-(T_n)$ satisfies the recurrence

$$x_{n+1} = \frac{e^h (e^J Z^+(T_n) + e^{-J} Z^-(T_n))^d}{e^{-h}(e^{-J} Z^+(T_n) + e^J Z^-(T_n))^d} = e^{2h} \left(\frac{e^J x_n + e^{-J}}{e^J + e^{-J} x_n} \right)^d = y(x_n)$$

where the function $y(\cdot)$ is defined as in Lemma A.3. Applying the conclusion of that lemma, we find that x_n increases with n and $x_n \to x(J, h)$ from below as $n \to \infty$. Finally, for the quantity $x_n^* = Z^+(T_n^*)/Z^-(T_n^*)$ we have

$$x_n^* = \frac{e^h (e^J Z^+(T_{n-1}) + e^{-J} Z^-(T_{n-1}))^{d+1}}{e^{-h}(e^{-J} Z^+(T_{n-1}) + e^J Z^-(T_n))^{d+1}} = e^{2h} \left(\frac{e^J x_{n-1} + e^{-J}}{e^J + e^{-J} x_{n-1}} \right)^{d+1} = e^{-2h/d} x_n^{(d+1)/d}.$$

Finally,

$$\ln \left(\frac{1 + \langle \sigma_a \rangle}{1 - \langle \sigma_a \rangle} \right) = \ln x_n^* = \frac{d+1}{d} \ln(x_n) - \frac{2h}{d}.$$

The lemma follows because $\ln(x_n)$ increases with n and converges to $\ln x(J, h)$ from below as $n \to \infty$. $\quad\square$

Corollary A.6. *Let $\mathscr{D}$ denote the family of Ising models with interaction strength J and zero external field on trees of maximum degree Δ. For any $\epsilon > 0$,*

$$\nu(\epsilon, \mathscr{D}) = \frac{\Delta}{\Delta - 1} \ln x \left(J, \frac{\epsilon}{2} \right) - \frac{\epsilon}{\Delta - 1}. \tag{24}$$

Proof. For the Ising model with zero external field, the joint distribution μ is symmetric with respect to flipping each bit of the database $\mathbf{x}$. This implies two simplifications in the formula for inferential privacy, Equation (19). First, the odds ratio $\rho = \frac{\mu(x_a = 1)}{\mu(x_a = 0)}$ is equal to 1. Second, both terms in the maximum on the right-hand side of the equation are equal, so ν_a is equal to $\ln \left(\frac{1 + \langle \sigma_a \rangle}{1 - \langle \sigma_a \rangle} \right)$, where the $\langle \cdot \rangle$ denotes averaging over the Gibbs measure (at inverse temperature 1) of the Ising model with interaction strength J and external field $\epsilon/2$. Applying Lemma A.5 we obtain (24) as a direct consequence. $\quad\square$

20

Proof of Theorem A.2. For part (a) of the theorem, Corollary A.6 justifies focusing our attention on the function $\ln x(J, h)$ where $h = \frac{\epsilon}{2}$ and $\epsilon > 0$ varies. In particular, when $\tanh(J) > \frac{1}{d}$, we have

$$\lim_{\epsilon \searrow 0} \nu(\epsilon, \mathscr{D}) = \frac{\Delta}{\Delta - 1} \lim_{h \searrow 0} \ln x(J, h/2) > 0$$

by Lemma A.3. This implies that the set of differentially enforceable ν has a strictly positive infimum, as claimed in part (a) of Theorem A.2.

When $\tanh(J) = \frac{1-\delta}{d}$, Eq. (22) for the partial derivative $\frac{\partial x}{\partial h}$ implies that $\frac{\partial x}{\partial h} = \frac{2}{\delta}$ at $h = 0, x = 1$. We now find that

$$
\begin{aligned}
\frac{d\nu(\epsilon, \mathscr{D})}{d\epsilon}\bigg|_{\epsilon=0} &= \left(\frac{\Delta}{\Delta-1}\right) \frac{\partial}{\partial \epsilon} \left[\ln x\left(J, \frac{\epsilon}{2}\right)\right]_{\epsilon=0} - \left(\frac{1}{\Delta-1}\right) \\
&= \frac{\Delta}{\Delta-1} \cdot \frac{1}{x(J,0)} \cdot \frac{1}{2} \cdot \left[\frac{\partial x}{\partial h}\right]_{h=0} - \left(\frac{1}{\Delta-1}\right) \\
&= \frac{\Delta}{\Delta-1} \cdot 1 \cdot \frac{1}{2} \cdot \frac{2}{\delta} - \left(\frac{1}{\Delta-1}\right) \\
&= \left(\frac{\Delta}{\Delta-1}\right) \frac{1}{\delta} - \left(\frac{1}{\Delta-1}\right) > \frac{1}{\delta}.
\end{aligned}
$$

Thus, for sufficiently small $\epsilon > 0$, we have $\nu(\epsilon, \mathscr{D}) > \epsilon/\delta$, which completes the proof of part (a) of the theorem.

To prove part (b) we consider rooted d-ary trees for some fixed $d \geq 2$. As in the proof of Lemma A.5 let T_n denote the complete rooted d-ary tree of depth n, with root node denoted by a. For $J > 0, h \in \mathbb{R}$ define

$$w_n(J, h) = \ln\left(\frac{1 + \langle \sigma_a \rangle}{1 - \langle \sigma_a \rangle}\right)$$

where $\langle \sigma_a \rangle$ denotes the expectation of σ_a under the Ising model on T_n with interaction strength J and external field h. In the proof of Lemma A.5 we denoted $\exp(w_n(J, h))$ by x_n and proved that $x_n \to x(J, h)$ from below as $n \to \infty$.

Now consider an adversary whose prior μ is the Ising model with interaction strength J and external field h_0. Note that the prior odds ratio $\rho = \frac{\mu(x_a=1)}{\mu(x_a=0)}$ satisfies

$$\ln \rho = \ln\left(\frac{\mu(\sigma_a = -1)}{\mu(\sigma_a = 1)}\right) = \ln\left(\frac{1 - \langle \sigma_a \rangle}{1 + \langle \sigma_a \rangle}\right) = -w_n(J, h_0).$$

Substituting this into Eq. (19), we see that for a given differential privacy parameter ε the corresponding inferential privacy guarantee is

$$\nu(\varepsilon) = \max\{w_n(J, h_0 + \tfrac{\varepsilon}{2}) - w_n(J, h_0), \ w_n(J, h_0) - w_n(J, h_0 - \tfrac{\varepsilon}{2})\}. \tag{25}$$

To prove Theorem A.2(b) consider any $0 < \varepsilon_0 < \varepsilon_1$ and $1 < r < R$. In setting up the adversary's prior, choose a value of h_0 that satisfies $2h_0 - \varepsilon_1 < 0 < 2h_0 - \varepsilon_0$. We aim to show that for all sufficiently large J and all $n > n_0(J)$, we have $\nu(\varepsilon_0) < r \cdot \varepsilon_0$ but $\nu(\varepsilon_1) > R \cdot \varepsilon_1$.

Let $w(J, h) = \lim_{n \to \infty} w_n(J, h) = \ln x(J, h)$. To prove that $\nu(\varepsilon_0) < r \cdot \varepsilon_0$ but $\nu(\varepsilon_1) > R \cdot \varepsilon_1$ for all sufficiently large n, it is sufficient to prove that

$$\frac{w(J, h_0 + \frac{1}{2}\varepsilon_0) - w(J, h_0)}{\varepsilon_0} < r \tag{26}$$

$$\frac{w(J, h_0) - w(J, h_0 - \frac{1}{2}\varepsilon_0)}{\varepsilon_0} < r \tag{27}$$

$$\frac{w(J, h_0) - w(J, h_0 - \frac{1}{2}\varepsilon_1)}{\varepsilon_1} > R. \tag{28}$$

To prove (26)-(27) we will show that $|\partial w / \partial h|$ is bounded above by $2r$ on the interval $[h_0 - \frac{1}{2}\varepsilon_0, h_0 + \frac{1}{2}\varepsilon_0]$ and apply the mean value theorem. Since $w(J, h) = \ln x(J, h)$ we have

$$\frac{\partial w}{\partial h} = \frac{\partial x / \partial h}{x} = \frac{(\partial y / \partial h)/x}{1 - \partial y / \partial x} = \frac{2}{1 - \partial y / \partial x}, \tag{29}$$

where we have used Eq. (22) and the facts that $\partial y / \partial h = 2y$ and that $y(x(J, h)) = x(J, h)$. Now, recalling the definition of $y(x)$ in Lemma A.3, we differentiate with respect to x and find that

$$\begin{aligned}
\frac{\partial y}{\partial x} &= e^{2h} \cdot d \left(\frac{e^J x + e^{-J}}{e^J + e^{-J} x} \right)^{d-1} \cdot \left(\frac{1 - e^{-4J}}{(e^J + e^{-J} x)^2} \right) \\
&= y(x) \cdot \left(\frac{d \cdot \left(1 - e^{-4J} \right)}{\left(e^J x + e^{-J} \right) \left(e^J + e^{-J} x \right)} \right) \\
&< \frac{y(x)}{x} \cdot \frac{d}{e^{2J}}.
\end{aligned} \tag{30}$$

Since $y(x)/x = 1$ when $x = x(J, h)$, we may combine (29) with (30) to conclude that whenever J is large enough that $de^{-2J} < 1 - \frac{1}{r}$, then the value of $\partial y / \partial x$ at $x(J, h)$ is less than $1 - 1/r$ for all $h > 0$, and consequently $\partial w / \partial h$ is bounded above uniformly by r, as desired.

Finally, to prove (28) we note that for $x = e^{2h+Jd}$ we have

$$\frac{y(x)}{x} = e^{2h} \left(\frac{e^J x + e^{-J}}{e^J + e^{-J} x} \right)^d e^{-2h-Jd} = \left(\frac{e^{2h+Jd} + e^{-2J}}{e^{2h+Jd-J} + e^J} \right)^d > 1 \tag{31}$$

for J sufficiently large. Recalling from the proof of Lemma A.3 that for $h > 0$ we have $y(x) > x$ when $1 < x < x(J, h)$ and $y(x) < x$ when $x > x(J, h)$, we see that $x(J, h) > e^{2h+Jd}$ provided that $h > 0$ and J is sufficiently large. An analogous argument applying the $h < 0$ case of Lemma A.3 shows that $x(J, h) < e^{2h-Jd}$ for $h < 0$ and J sufficiently large. Recalling that $w(J, h) = \ln x(J, h)$ and that $h_0 > 0 > h_0 - \varepsilon_1/2$ we find that

$$\begin{aligned}
w(J, h_0) &> 2h_0 + Jd > Jd \\
w(J, h_0 - \varepsilon_1/2) &< 2h_0 - \varepsilon_1 - Jd < -Jd \\
\tfrac{w(J,h_0)-w(J,h_0-\frac{1}{2}\varepsilon_1)}{\varepsilon_1} &> \tfrac{2Jd}{\varepsilon_1} > R
\end{aligned}$$

provided J is sufficiently large. This establishes (31) and concludes the proof of Theorem A.2(b). $\qquad \square$

B   Appendix to §4: Bounded Affiliation Distributions

This appendix contains a full proof of Theorem 4.2. The proof requires developing a theory of "multiplicative estimates" that is the multiplicative analogue of the notion of "estimate" used by Dobrushin [6], Föllmer [11], and Künsch [21] in their proofs of the so-called Dobrushin Comparison Theorem. We define multiplicative estimates and build up the necessary machinery for dealing with them in §B.1. Then, in §B.2 we prove Theorem 4.2.

B.1   Multiplicative estimates

Let S be any set of potential outcomes of the mechanism $\mathcal{M}$ such that $\Pr(\mathcal{M}(\mathbf{x}) \in S) > 0$. Let π^1 denote the conditional distribution on databases $\mathbf{x} \in X^n$, given that $\mathcal{M}(\mathbf{x}) \in S$, and let π^2 denote the unconditional distribution μ, respectively.

For $a \in \{1, 2\}$ and for any function $f : X^n \to \mathbb{R}$, let $\pi^a(f)$ denote the expected value of f under distribution π^a. Also define the Lipschitz constants

$$\rho_i(f) = \max\{f(\mathbf{x}) - f(\mathbf{x}') \mid \mathbf{x} \sim_i \mathbf{x}'\}. \tag{32}$$

Let us say that a vector $\boldsymbol{\kappa} = (\kappa_i)$ is a *multiplicative estimate* if for every function $f : X^n \to \mathbb{R}_+$ we have

$$|\ln \pi^1(f) - \ln \pi^2(f)| \le \sum_{i=1}^n \kappa_i \rho_i(\ln f). \tag{33}$$

This section is devoted to proving some basic facts about multiplicative estimates that underpin the proof of Theorem 4.2. To start, we need the following lemma.

Lemma B.1. *Consider a probability space with two functions A, B taking values in the positive real numbers. If $(\sup A)/(\inf A) \le e^{2a}$ and $(\sup B)/(\inf B) \le e^{2b}$ then*

$$\frac{\mathbb{E}[AB]}{\mathbb{E}[A]\,\mathbb{E}[B]} \le 1 + \frac{(e^{2a}-1)(e^{2b}-1)}{(e^a+e^b)^2} \le e^{ab}. \tag{34}$$

Proof. The hypotheses and conclusion of the lemma are invariant under rescaling each of A and B, so we may assume without loss of generality that A is supported in the interval $[e^{-a}, e^a]$ and that B is supported in the interval $[e^{-b}, e^b]$. At each sample point ω, the following two equations hold:

$$\begin{bmatrix} \frac{e^a - A(\omega)}{e^a - e^{-a}} & \frac{A(\omega) - e^{-a}}{e^a - e^{-a}} \end{bmatrix} \begin{bmatrix} 1 & e^{-a} \\ 1 & e^a \end{bmatrix} = \begin{bmatrix} 1 & A(\omega) \end{bmatrix} \tag{35}$$

$$\begin{bmatrix} \frac{e^b - B(\omega)}{e^a - e^{-b}} & \frac{B(\omega) - e^{-b}}{e^a - e^{-b}} \end{bmatrix} \begin{bmatrix} 1 & e^{-b} \\ 1 & e^b \end{bmatrix} = \begin{bmatrix} 1 & B(\omega) \end{bmatrix} \tag{36}$$

$$\tag{37}$$

Therefore, if we define the matrix-valued random variable

$$M(\omega) = \begin{bmatrix} \frac{e^a - A(\omega)}{e^a - e^{-a}} & \frac{A(\omega) - e^{-a}}{e^a - e^{-a}} \end{bmatrix}^\mathsf{T} \begin{bmatrix} \frac{e^b - B(\omega)}{e^b - e^{-b}} & \frac{B(\omega) - e^{-b}}{e^b - e^{-b}} \end{bmatrix}$$

we have

$$\begin{bmatrix} 1 & 1 \\ e^{-a} & e^a \end{bmatrix} M(\omega) \begin{bmatrix} 1 & e^{-b} \\ 1 & e^b \end{bmatrix} = \begin{bmatrix} 1 \\ A(\omega) \end{bmatrix} \begin{bmatrix} 1 & B(\omega) \end{bmatrix} = \begin{bmatrix} 1 & B(\omega) \\ A(\omega) & A(\omega)B(\omega) \end{bmatrix}.$$

Integrating over ω we obtain

$$\begin{bmatrix} 1 & 1 \\ e^{-a} & e^a \end{bmatrix} \mathbb{E}[M] \begin{bmatrix} 1 & e^{-b} \\ 1 & e^b \end{bmatrix} = \begin{bmatrix} 1 & \mathbb{E}[B] \\ \mathbb{E}[A] & \mathbb{E}[AB] \end{bmatrix}$$

$$\mathbb{E}[M] = \begin{bmatrix} 1 & 1 \\ e^{-a} & e^a \end{bmatrix}^{-1} \begin{bmatrix} 1 & \mathbb{E}[B] \\ \mathbb{E}[A] & \mathbb{E}[AB] \end{bmatrix} \begin{bmatrix} 1 & e^{-b} \\ 1 & e^b \end{bmatrix}^{-1}$$

$$\left(e^a - e^{-a}\right)\left(e^b - e^{-b}\right) \mathbb{E}[M] = \begin{bmatrix} e^a & -1 \\ -e^{-a} & 1 \end{bmatrix} \begin{bmatrix} 1 & \mathbb{E}[B] \\ \mathbb{E}[A] & \mathbb{E}[AB] \end{bmatrix} \begin{bmatrix} e^b & -e^{-b} \\ -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} e^{a+b} - e^a\mathbb{E}[B] - e^b\mathbb{E}[A] + \mathbb{E}[AB] & -e^{a-b} + e^a\mathbb{E}[B] + e^{-b}\mathbb{E}[A] - \mathbb{E}[AB] \\ -e^{b-a} + e^{-a}\mathbb{E}[B] + e^b\mathbb{E}[A] - \mathbb{E}[AB] & e^{-a-b} - e^{-a}\mathbb{E}[B] - e^{-b}\mathbb{E}[A] + \mathbb{E}[AB] \end{bmatrix}.$$

23

Each entry of the matrix on the left side is non-negative, hence the entries on the right side are non-negative as well. This tells us that

$$\mathbb{E}[AB] \le \min\left\{e^a\mathbb{E}[B] + e^{-b}\mathbb{E}[A] - e^{a-b}, e^b\mathbb{E}[A] + e^{-a}\mathbb{E}[B] - e^{b-a}\right\}$$

$$= \mathbb{E}[A]\mathbb{E}[B] + \min\left\{(e^a - \mathbb{E}[A])(\mathbb{E}[B] - e^{-b}), (e^b - \mathbb{E}[B])(\mathbb{E}[A] - e^{-a})\right\}. \tag{38}$$

Letting

$$\alpha = \tfrac{1}{\mathbb{E}[A]}, \quad \beta = \tfrac{1}{\mathbb{E}[B]},$$

we can multiply both sides of (38) by $\alpha\beta$ to obtain

$$\frac{\mathbb{E}[AB]}{\mathbb{E}[A]\,\mathbb{E}[B]} \le 1 + \min\{(e^a\alpha - 1)(1 - e^{-b}\beta), (e^b\beta - 1)(1 - e^{-a}\alpha)\}. \tag{39}$$

Denote the right side of (39) by $G(\alpha, \beta)$. We aim to find the maximum value of $G(\alpha, \beta)$ as (α, β) ranges over the rectangle $[e^{-a}, e^a] \times [e^{-b}, e^b]$. Note that $G \equiv 1$ on the boundary of this rectangle, whereas $G > 1$ on the interior of the rectangle. At any point of the interior where $(e^a\alpha - 1)(1 - e^{-b}\beta) > (e^b\beta - 1)(1 - e^{-a}\alpha)$ we have $\frac{\partial G}{\partial \beta} = e^b(1 - e^{-a}\alpha) > 0$, and similarly at any point of the interior where $(e^a\alpha - 1)(1 - e^{-b}\beta) < (e^b\beta - 1)(1 - e^{-a}\alpha)$ we have $\frac{\partial G}{\partial \alpha} > 0$. Therefore if (α, β) is a global maximum of G we must have $(e^a\alpha - 1)(1 - e^{-b}\beta) = (e^b\beta - 1)(1 - e^{-a}\alpha)$. Let

$$r = \frac{e^a\alpha - 1}{1 - e^{-a}\alpha} = \frac{e^b\beta - 1}{1 - e^{-b}\beta}. \tag{40}$$

A manipulation using (40) yields

$$(e^a\alpha - 1)(1 - e^{-b}\beta) = \left(e^{2a} - 1\right)\left(e^{2b} - 1\right)\frac{r}{\left(r + e^{2a}\right)\left(r + e^{2b}\right)} \tag{41}$$

and by setting the derivative of the right side to zero we find that it is maximized at $r = e^{a+b}$, when it equates to $(e^{2a-1})(e^{2b} - 1)(e^a + e^b)^{-2}$. Therefore,

$$\frac{\mathbb{E}[AB]}{\mathbb{E}[A]\,\mathbb{E}[B]} \le 1 + \frac{\left(e^{2a} - 1\right)\left(e^{2b} - 1\right)}{\left(e^a + e^b\right)^2},$$

which establishes the first inequality in (34). The prove the second inequality, we consider how $1 + (e^{2a} - 1)(e^{2b} - 1)(e^a + e^b)^{-2}$ varies as we vary a and b while holding their product fixed at some value, x^2. To begin we compute the gradient of $1 + (e^{2a} - 1)(e^{2b} - 1)(e^a + e^b)^{-2}$.

$$\nabla\left[1 + (e^{2a} - 1)(e^{2b} - 1)(e^a + e^b)^{-2}\right] = \left[\left(\frac{2e^{2a}(e^{2b}-1)}{(e^a+e^b)^2} - \frac{2e^a(e^{2a}-1)(e^{2b}-1)}{(e^a+e^b)^3}\right) \quad \left(\frac{2e^{2b}(e^{2a}-1)}{(e^a+e^b)^2} - \frac{2e^b(e^{2a}-1)(e^{2b}-1)}{(e^a+e^b)^3}\right)\right]$$

$$= \frac{2(e^{a+b} - 1)}{(e^a + e^b)^3}\left[e^a(e^{2b} - 1) \quad e^b(e^{2a} - 1)\right]$$

$$= \frac{8\left(1 - e^{-a-b}\right)}{(e^a + e^b)^3}\left[\sinh b \quad \sinh a\right]$$

Parameterizing the curve $ab = x^2$ by $a(t) = xt, b(t) = x/t$, we have $\dot{a}(t) = a/t$ and $\dot{b}(t) = -b/t$, so

$$\frac{d}{dt}\left[1 + (e^{2a} - 1)(e^{2b} - 1)(e^a + e^b)^{-2}\right] = \frac{8\left(1 - e^{-a-b}\right)}{(e^a + e^b)^3}\left[\sinh b \quad \sinh a\right]\begin{bmatrix} a/t \\ -b/t \end{bmatrix}$$

$$= \frac{8ab\left(1 - e^{-a-b}\right)}{(e^a + e^b)^3 t}\left(\frac{\sinh b}{b} - \frac{\sinh a}{a}\right).$$

24

From the Taylor series $\frac{\sinh y}{y} = \sum_{i=0}^{\infty} \frac{1}{(2i+1)!} y^{2i}$ we see that $\frac{\sinh y}{y}$ is an increasing function of $y \geq 0$, so along the curve $a(t) = xt, b(t) = x/t$, the function $1 + (e^{2a} - 1)(e^{2b} - 1)(e^a + e^b)^{-2}$ increases when $a < b$ (corresponding to $t < 1$) and decreases when $a > b$ (corresponding to $t > 1$), reaching its maximum when $t = 1$ and $a = b = x$. Hence

$$1 + \frac{(e^{2a} - 1)(e^{2b} - 1)}{(e^a + e^b)^2} \leq 1 + \left(\frac{e^{2x} - 1}{2e^x}\right)^2 = 1 + \sinh^2 x = \cosh^2 x. \tag{42}$$

Finally, by comparing Taylor series coefficients we can see that $\cosh x \leq e^{x^2/2}$ for all $x \geq 0$, and squaring both sides of this relation we obtain

$$\cosh^2 x \leq e^{x^2} = e^{ab}. \tag{43}$$

The second inequality in (34) follows by combining (42) with (43). $\qquad\square$

Lemma B.2. *If κ is a multiplicative estimate, then for any $i \in \{1, \ldots, n\}$, the vector $T_i(\kappa)$ defined by*

$$(T_i(\kappa))_\ell = \begin{cases} \kappa_\ell & \text{if } \ell \neq i \\ \frac{\epsilon_i}{2} + \sum_{j=1}^n \gamma_{ij} \kappa_j & \text{if } \ell = i \end{cases} \tag{44}$$

is also a multiplicative estimate.

Proof. For a distribution π on X^n, a database $\mathbf{x} \in X^n$, and an individual i, let $\pi(\cdot \mid \mathbf{x}_{-i})$ denote the conditional distribution of x_i given $\mathbf{x}_{-i}$. In other words, $\pi(\cdot \mid \mathbf{x}_{-i})$ is the probability distribution on X given by

$$\pi(x \mid \mathbf{x}_{-i}) = \frac{\pi(x, \mathbf{x}_{-i})}{\sum_{y \in X} \pi(y, \mathbf{x}_{-i})}. \tag{45}$$

Letting W denote the vector space of real-valued functions on X^n, we define an averaging operator $\tau_i : W \to W$ which maps a function f to the function

$$\tau_i f(\mathbf{x}) = \sum_{y \in X} f(y, \mathbf{x}_{-i}) \pi(y \mid \mathbf{x}_{-i}).$$

Equivalently, $\tau_i f$ is the unique function satisfying:

1. The value $\tau_i f(\mathbf{x})$ depends only on $\mathbf{x}_{-i}$.

2. For any other function g whose value $g(\mathbf{x})$ depends only on $\mathbf{x}_{-i}$, we have

$$\pi(fg) = \pi((\tau_i f)g). \tag{46}$$

Note that $\pi(f) = \pi(\tau_i f)$, as can be seen from applying (46) to the constant function $g(\mathbf{x}) = 1$.

For the distributions π^1 and π^2 defined earlier, let us denote the corresponding averaging operators by τ_i^1 and τ_i^2. Using the identities $\pi^1(f) = \pi^1(\tau_i^1 f)$ and $\pi^2(f) = \pi^2(\tau_i^2 f)$, we find that

$$|\ln \pi^1(f) - \ln \pi^2(f)| \leq |\ln \pi^1(\tau_i^1 f) - \ln \pi^1(\tau_i^2 f)| + |\ln \pi^1(\tau_i^2 f) - \ln \pi^2(\tau_i^2 f)|. \tag{47}$$

We bound the two terms on the right side separately. For the first term, we write

$$\left|\ln\left(\frac{\pi^1(\tau_i^1 f)}{\pi^1(\tau_i^2 f)}\right)\right| = \left|\ln\left(\frac{\sum_{\mathbf{x}} \pi^1(\mathbf{x})\tau_i^1 f(\mathbf{x})}{\sum_{\mathbf{x}} \pi^1(\mathbf{x})\tau_i^2 f(\mathbf{x})}\right)\right| \leq \max_{\mathbf{x}} \left|\ln\left(\frac{\tau_i^1 f(\mathbf{x})}{\tau_i^2 f(\mathbf{x})}\right)\right|. \tag{48}$$

For any particular $\mathbf{x}$, we can bound the ratio $\frac{\tau_i^1 f(\mathbf{x})}{\tau_i^2 f(\mathbf{x})}$ from above using Lemma B.1 applied to the probability space X, under the distribution $\pi^2(\cdot \mid \mathbf{x}_{-i})$. Letting $A(x) = f(x, \mathbf{x}_{-i}), B(x) = \pi^1(x \mid \mathbf{x}_{-i})/\pi^2(x \mid \mathbf{x}_{-i})$ we have

$$\mathbb{E}[A(x)B(x)] = \tau_i^1 f(\mathbf{x}) \tag{49}$$

$$\mathbb{E}[A(x)] = \tau_i^2 f(\mathbf{x}) \tag{50}$$

$$\mathbb{E}[B(x)] = 1 \tag{51}$$

$$\frac{\sup A}{\inf A} \le e^{\rho_i(\ln f)} \tag{52}$$

$$\frac{\sup B}{\inf B} \le e^{2\epsilon_i}. \tag{53}$$

The first four of these relations are straightforward, and the last requires some justification. In the following calculation we use the operator $\Pr(\cdot)$ to denote probabilities of events in the sample space where $\mathbf{x}$ is sampled from the original joint distribution μ, and randomized mechanism $\mathcal{M}$ is applied to $\mathbf{x}$. Starting from the definitions of π^1 and π^2, an application of Bayes' Law yields the following calculation.

$$\pi^1(x \mid \mathbf{x}_{-i}) = \frac{\Pr(\mathbf{x} = (x, \mathbf{x}_{-i}) \mid \mathcal{M}(\mathbf{x}) \in S)}{\Pr(\mathbf{x} \in X \times \{\mathbf{x}_{-i}\} \mid \mathcal{M}(\mathbf{x}) \in S)}$$

$$= \frac{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid \mathbf{x} = (x, \mathbf{x}_{-i}))}{\Pr(\mathcal{M}(\mathbf{x}) \in S \mid \mathbf{x} \in X \times \{\mathbf{x}_{-i}\})} \cdot \frac{\Pr(\mathbf{x} = (x, \mathbf{x}_{-i}))}{\Pr(\mathbf{x} \in X \times \{\mathbf{x}_{-i}\})}.$$

The first factor on the right-hand side is between $e^{-\epsilon_i}$ and e^{ϵ_i}, while the second factor is equal to $\pi^2(x \mid \mathbf{x}_{-i})$. This completes the proof of (53). By combining Lemma B.1 with the contents of (48)-(53) we obtain the bound

$$|\ln \pi^1(\tau_i^1 f) - \ln \pi^1(\tau_i^2 f)| \le \tfrac{1}{2}\epsilon_i \rho_i(\ln f). \tag{54}$$

To bound the second term in (47), we will make use of the following inequality, a multiplicative analogue of inequality (3.5) in [14].

$$\forall i, j \quad \rho_j(\ln \tau_i^2 f) \le \begin{cases} 0 & \text{if } i = j \\ \rho_j(\ln f) + \gamma_{ij}\, \rho_i(\ln f) & \text{if } i \neq j. \end{cases} \tag{55}$$

The validity of (55) is evident when $i = j$, since the value $\tau_i^2 f(\mathbf{x})$ does not depend on x_i. To prove (55) when $i \neq j$, we use the definition of $\rho_j(\cdot)$ to choose $\mathbf{x}, \mathbf{x}' \in X^n$ such that $\mathbf{x} \sim_j \mathbf{x}'$ and

$$\rho_j(\ln \tau_i^2 f) = \ln \tau_i^2 f(\mathbf{x}') - \ln \tau_i^2 f(\mathbf{x})$$

$$= \ln \left(\sum_{x \in X} \pi^2(x \mid \mathbf{x}'_{-i}) f(x, \mathbf{x}'_{-i}) \right) - \ln \left(\sum_{x \in X} \pi^2(x \mid \mathbf{x}_{-i}) f(x, \mathbf{x}_{-i}) \right)$$

$$\le \left| \ln \left(\frac{\sum_{x \in X} \pi^2(x \mid \mathbf{x}'_{-i}) f(x, \mathbf{x}'_{-i})}{\sum_{x \in X} \pi^2(x \mid \mathbf{x}'_{-i}) f(x, \mathbf{x}_{-i})} \right) \right| + \left| \ln \left(\frac{\sum_{x \in X} \pi^2(x \mid \mathbf{x}'_{-i}) f(x, \mathbf{x}_{-i})}{\sum_{x \in X} \pi^2(x \mid \mathbf{x}_{-i}) f(x, \mathbf{x}_{-i})} \right) \right|. \tag{56}$$

Using the fact that $\frac{f(x, \mathbf{x}'_{-i})}{f(x, \mathbf{x}_{-i})} \le e^{\rho_j(\ln f)}$ for all $x \in X$, we see that the first term on the right side of (56) is bounded above by $\rho_j(\ln f)$. To bound the second term, we again make use of Lemma B.1, this time substituting $A(x) = f(x, \mathbf{x}_{-i})$ and $B(x) = \frac{\pi^2(x \mid \mathbf{x}'_{-i})}{\pi^2(x \mid \mathbf{x}_{-i})}$. Taking expectations under the probability distribution

$\pi^2(x \mid \mathbf{x}_{-i})$ we have

$$\mathbb{E}[A(x)B(x)] = \sum_{x \in X} \pi^2(x \mid \mathbf{x}'_{-i}) f(x, \mathbf{x}_{-i})$$

$$\mathbb{E}[A(x)] = \sum_{x \in X} \pi^2(x \mid \mathbf{x}_{-i}) f(x, \mathbf{x}_{-i})$$

$$\mathbb{E}[B(x)] = 1$$

$$\frac{\sup A}{\inf A} \leq e^{\rho_i(\ln f)}$$

$$\frac{\sup B}{\inf B} \leq e^{4\gamma_{ij}},$$

where the last line is justified by observing that the definition of γ_{ij} ensures that $\sup B \leq e^{2\gamma_{ij}}$ and $\inf B \geq e^{-2\gamma_{ij}}$. An application of Lemma B.1 immediately implies that the second term on the right side of (56) is bounded above by $\gamma_{ij}\rho_i(\ln f)$. This completes the proof of (55).

Now, our hypothesis that $\boldsymbol{\kappa}$ is a multiplicative estimate implies, by definition, that

$$|\ln \pi^1(\tau_i^2 f) - \ln \pi^2(\tau_i^2 f)| \leq \sum_{j=1}^n \kappa_j \rho_j(\ln \tau_i^2 f) \leq \sum_{j \neq i} \kappa_j \rho_j(\ln f) + \sum_{j=1}^n \kappa_j \gamma_{ij} \, \rho_i(\ln f) \qquad (57)$$

where we have applied (55) to derive the second inequality. Combining (47), (54), and (57) we now have

$$|\ln \pi^1(f) - \ln \pi^2(f)| \leq \sum_{j \neq i} \kappa_j \rho_j(\ln f) + \left(\tfrac{1}{2}\epsilon_i + \sum_{j=1}^n \gamma_{ij}\kappa_j \right) \rho_i(\ln f).$$

$\square$

Lemma B.3. *If $\boldsymbol{\kappa}$ is a multiplicative estimate, then the vector $T(\boldsymbol{\kappa}) = \frac{1}{2n}\boldsymbol{\epsilon} + \left(1 - \frac{1}{n}\right)\boldsymbol{\kappa} + \frac{1}{n}\Gamma\boldsymbol{\kappa}$ is also a multiplicative estimate.*

Proof. Averaging (44) over $i = 1, \ldots, n$, we find that $T(\boldsymbol{\kappa}) = \frac{1}{n}T_i(\boldsymbol{\kappa})$. The lemma follows because the set of multiplicative estimates is closed under convex combinations, as is evident from the definition of a multiplicative estimate. $\square$

Lemma B.4. *If the influence matrix Γ has spectral norm strictly less than 1, then $\frac{1}{2}\Phi\boldsymbol{\epsilon}$ is an estimate.*

Proof. To begin, let us prove that the set of multiplicative estimates is non-empty; in fact, it contains the vector $\mathbf{1} = (1, \ldots, 1)$. To see this, consider any $f : X^n = \mathbb{R}_+$ and choose $\mathbf{x} \in \arg\max(f), \mathbf{x}' \in \arg\min(f)$. Define a sequence $(\mathbf{x}^{(k)})_{k=0}^n$ by the formula

$$\mathbf{x}_i^{(k)} = \begin{cases} x_i & \text{if } i > k \\ x_i' & \text{if } i \leq k \end{cases}.$$

Note that $\mathbf{x}^{(0)} = \mathbf{x}$, $\mathbf{x}^{(n)} = \mathbf{x}'$, and $\mathbf{x}^{(k-1)} \sim_k \mathbf{x}^{(k)}$ for $k = 1, \ldots, n$. Therefore,

$$|\ln \pi^1(f) - \ln \pi^2(f)| \leq |\max(\ln f) - \min(\ln f)| = |\ln f(\mathbf{x}) - \ln f(\mathbf{x}')|$$

$$\leq \sum_{k=1}^n |\ln f(\mathbf{x}^{(k-1)}) - \ln f(\mathbf{x}^{(k)})| \leq \sum_{k=1}^n \rho_k(\ln f),$$

so $\mathbf{1}$ is an estimate as claimed.

Now let $\Upsilon = \left(1 - \frac{1}{n}\right)I + \frac{1}{n}\Gamma$. Applying Theorem B.3 inductively, each element of the sequence

$$T^m(\mathbf{1}) = \frac{1}{2n}\left(\sum_{k=0}^{m-1}\Upsilon^k\right)\boldsymbol{\epsilon} + \Upsilon^m\mathbf{1}$$

is a multiplicative estimate. If $\|\Gamma\| < 1$ (where $\|\cdot\|$ denotes spectral norm) then Υ also has spectral norm less than 1 because it is a convex combination of Γ and I. This implies that the sequence $(T^m(\boldsymbol{\kappa}))_{m=0}^\infty$ converges to $\frac{1}{2n}(I - \Upsilon)^{-1}\boldsymbol{\epsilon}$. Now,

$$(I - \Upsilon)^{-1} = (\tfrac{1}{n}I - \tfrac{1}{n}\Gamma)^{-1} = n(I - \Gamma)^{-1},$$

so the sequence $(T^m(\boldsymbol{\kappa}))_{m=0}^\infty$ converges to $\frac{1}{2}\Phi\boldsymbol{\epsilon}$. The proof concludes with the observation that a limit of multiplicative estimates is again a multiplicative estimate. $\qquad\square$

B.2   Proof of Theorem 4.2

Let us begin by restating Theorem 4.2.

Theorem B.5. *Suppose that the joint distribution* $\mathbf{x}$ *has a multiplicative influence matrix* Γ *whose spectral norm is strictly less than 1. Let* $\Phi = (\phi_{ij})$ *denote the matrix inverse of* $I - \Gamma$. *Then for any mechanism with individual privacy parameters* $\boldsymbol{\epsilon} = (\epsilon_i)$, *the networked differential privacy guarantee satisfies*

$$\forall i \ \ \nu_i \leq 2\sum_{j=1}^n \phi_{ij}\epsilon_j. \tag{58}$$

If the matrix of multiplicative influences satisfies

$$\forall i \ \ \sum_{j=1}^n \gamma_{ij}\epsilon_j \leq (1 - \delta)\epsilon_i \tag{59}$$

for some $\delta > 0$, *then*

$$\forall i \ \ \nu_i \leq 2\epsilon_i/\delta. \tag{60}$$

Proof. Above, in Lemma B.4, we proved that $\frac{1}{2}\Phi\boldsymbol{\epsilon}$ is a multiplicative estimate. In other words, for any $f : X^n \to \mathbb{R}_+$ it holds that

$$\left|\ln \pi^1(f) - \ln \pi^2(f)\right| \leq \tfrac{1}{2}\sum_{i,j=1}^n \Phi_{ij}\epsilon_j\rho_i(\ln f). \tag{61}$$

To prove (58), we are required to show the following: if z_0, z_1 are any two distinct elements of X such that $\Pr(x_i = z_0)$ and $\Pr(x_i = z_1)$ are both positive, then

$$\left|\ln\left(\frac{\Pr(x_i = z_1 \mid \mathcal{M}(\mathbf{x}) \in S)\,/\,\Pr(x_i = z_0 \mid \mathcal{M}(\mathbf{x}) \in S)}{\Pr(x_i = z_1)\,/\,\Pr(x_i = z_0)}\right)\right| \leq 2\sum_{j=1}^n \Phi_{ij}\epsilon_j. \tag{62}$$

We will do this by setting f and g to be the indicator functions of the events $x_i = z_0$ and $x_i = z_1$, respectively. Then (62) can be rewritten in the form

$$\left|\ln\left(\frac{\pi^1(f)/\pi^1(g)}{\pi^2(f)/\pi^2(g)}\right)\right| \leq 2\sum_{j=1}^n \Phi_{ij}\epsilon_j. \tag{63}$$

28

If the Lipschitz constants of f and g satisfied $\rho_j(\ln f) = \rho_j(\ln g) = 0$ for $j \neq i$ and $\rho_i(\ln f), \rho_i(\ln g) \leq 1$, then (63) would follow immediately by applying (61) to f and g separately. Instead $\rho_i(f) = \rho_i(g) = \infty$ so we will have to be more indirect, applying (61) to τf and τg where τ is an averaging operator designed to smooth out f and g, thereby improving their Lipschitz constants. Specifically, define

$$\tau f(\mathbf{x}) = \pi^2(z_0, \mathbf{x}_{-i}), \qquad \tau g(\mathbf{x}) = \pi^2(z_1, \mathbf{x}_{-i}).$$

It is useful to describe τf and τg in terms of the following sampling process: generate a coupled pair of samples $(\mathbf{x}', \mathbf{x}'')$ by sampling $\mathbf{x}'$ from π^2, then resampling x_i'' from the conditional distribution $\pi^2(\cdot \mid \mathbf{x}'_{-i})$, and then assembling the database $\mathbf{x}'' = (x_i'', \mathbf{x}'_{-i})$. Then $\tau f(\mathbf{x})$ is the conditional probability that $x_i'' = z_0$ given that $\mathbf{x}' = \mathbf{x}$, and τg is defined similarly using z_1 instead of z_0. An important observation is that the distribution of $(\mathbf{x}', \mathbf{x}'')$ is exchangeable, i.e. $(\mathbf{x}', \mathbf{x}'')$ and $(\mathbf{x}'', \mathbf{x}')$ have the same probability. From this observation we can immediately conclude that

$$\pi^2(\tau f) = \pi^2(f), \quad \pi^2(\tau g) = \pi^2(g), \tag{64}$$

because $\pi^2(\tau f)$ is the probability that $x_i'' = z_0$ whereas $\pi^2(f)$ is the probability that $x_i' = z_0$, and similarly for g and z_1. Our strategy for proving (63) will be to bound the left side using

$$\left| \ln \left(\frac{\pi^1(f)/\pi^1(g)}{\pi^2(f)/\pi^2(g)} \right) \right| = \left| \ln \left(\frac{\pi^1(f)/\pi^1(g)}{\pi^2(\tau f)/\pi^2(\tau g)} \right) \right|$$

$$\leq \left| \ln \left(\frac{\pi^1(f)/\pi^1(g)}{\pi^1(\tau f)/\pi^1(\tau g)} \right) \right| + \left| \ln \left(\frac{\pi^1(\tau f)/\pi^1(\tau g)}{\pi^2(\tau f)/\pi^2(\tau g)} \right) \right| \tag{65}$$

and to bound the two terms on the last line separately. For the second term we will use (61) applied to τf and τg separately. This requires us to bound the Lipschitz constants $\rho_k(\ln \tau f)$ and $\rho_k(\ln \tau g)$. Since $\tau f(\mathbf{x})$ and $\tau g(\mathbf{x})$ do not depend on x_i, it is immediate that $\rho_f(\ln \tau f) = \rho_i(\ln \tau g) = 0$. For $k \neq i$, the definition of the multiplicative influence parameter γ_{ik} leads to the bounds

$$\rho_k(\ln \tau f), \, \rho_k(\ln \tau g) \leq 2\gamma_{ik}. \tag{66}$$

Note that (66) also holds when $k = i$ since $\gamma_{ii} = 0$. Applying (61) to τf and τg yields the bound

$$\left| \ln \left(\frac{\pi^1(\tau f)/\pi^1(\tau g)}{\pi^2(\tau f)/\pi^2(\tau g)} \right) \right| \leq \left| \ln \pi^1(\tau f) - \ln \pi^2(\tau f) \right| + \left| \ln \pi^1(\tau g) - \ln \pi^2(\tau g) \right|$$

$$\leq 2 \cdot \frac{1}{2} \cdot \sum_{k,j=1}^{n} (2\gamma_{ik}) \Phi_{kj} \epsilon_j = 2(\Gamma \Phi \boldsymbol{\epsilon})_i \tag{67}$$

Recalling that $\Phi = (I - \Gamma)^{-1}$, we have $(I - \Gamma)\Phi = I$ and hence $\Gamma\Phi = \Phi - I$. Thus, we can rewrite (67) as

$$\left| \ln \left(\frac{\pi^1(\tau f)/\pi^1(\tau g)}{\pi^2(\tau f)/\pi^2(\tau g)} \right) \right| \leq 2 \sum_{j=1}^{n} \Phi_{ij} \epsilon_j \; - \; 2\epsilon_i. \tag{68}$$

Now we turn to bounding the first term in (65). Letting o denote the random variable representing the mechanism's outcome, $\mathcal{M}(\mathbf{x})$. Bayes' Law tells us that

$$\pi^1(\mathbf{x}) = \frac{\pi^2(\mathbf{x}) \Pr(o \in S \mid \mathbf{x})}{\Pr(o \in S)}.$$

Therefore,

$$
\frac{\pi^1(\tau g)}{\pi^1(g)} = \frac{\sum_{\mathbf{x}} \pi^2(z_1 \mid \mathbf{x}_{-i})\pi^1(\mathbf{x})}{\sum_{\mathbf{x}} g(\mathbf{x})\pi^1(\mathbf{x})} = \frac{\sum_{\mathbf{x}} \pi^2(z_1 \mid \mathbf{x}_{-i})\pi^2(\mathbf{x})\Pr(o \in S \mid \mathbf{x})}{\sum_{\mathbf{x}} g(\mathbf{x})\pi^2(\mathbf{x})\Pr(o \in S \mid \mathbf{x})}
$$

$$
= \frac{\sum_{\mathbf{x}} \pi^2(z_1 \mid \mathbf{x}_{-i})\pi^2(\mathbf{x})\Pr(o \in S \mid \mathbf{x})}{\sum_{\mathbf{x}_{-i}} \pi^2(z_1, \mathbf{x}_{-i})\Pr(o \in S \mid (z_1, \mathbf{x}_{-i}))}
$$

$$
= \frac{\sum_{\mathbf{x}_{-i}} \pi^2(z_1 \mid \mathbf{x}_{-i})\sum_{z \in X} \pi^2(z, \mathbf{x}_{-i})\Pr(o \in S \mid (z, \mathbf{x}_{-i}))}{\sum_{\mathbf{x}_{-i}} \pi^2(z_1 \mid \mathbf{x}_{-i})\sum_{z \in X} \pi^2(z, \mathbf{x}_{-i})\Pr(o \in S \mid (z_1, \mathbf{x}_{-i}))}.
$$

The right side lies between $e^{-\epsilon_i}$ and e^{ϵ_i} because each ratio $\frac{\Pr(o \in S \mid (z, \mathbf{x}_{-i}))}{\Pr(o \in S \mid (z_1, \mathbf{x}_{-i}))}$ lies between $e^{-\epsilon_i}$ and e^{ϵ_i}. Thus,

$$
\left| \ln\left(\frac{\pi^1(\tau g)}{\pi^1(g)} \right) \right| \le \epsilon_i \tag{69}
$$

Similarly,

$$
\left| \ln\left(\frac{\pi^1(f)}{\pi^1(\tau f)} \right) \right| \le \epsilon_i. \tag{70}
$$

Combining (69) with (70) yields the bound

$$
\left| \ln\left(\frac{\pi^1(f)/\pi^1(g)}{\pi^1(\tau f)/\pi^1(\tau g)} \right) \right| \le 2\epsilon_i. \tag{71}
$$

Combining (71) with (68) we obtain the bound (63), which finishes the proof of the first inequality in the theorem statement, namely (58).

To prove inequality (60), we use the partial ordering on vectors defined by $\boldsymbol{a} \preceq \boldsymbol{b}$ if and only if $a_i \le b_i$ for all i. The matrix Γ has non-negative entries, so it preserves this ordering: if $\boldsymbol{a} \preceq \boldsymbol{b}$ then $\forall i\ \sum_j \gamma_{ij}a_j \le \sum_j \gamma_{ij}b_j$ and hence $\Gamma \boldsymbol{a} \preceq \Gamma \boldsymbol{b}$. Rewriting the relation (59) in the form $\Gamma\boldsymbol{\epsilon} \preceq (1-\delta)\boldsymbol{\epsilon}$ and applying induction, we find that for all $n \ge 0$, $\Gamma^n\boldsymbol{\epsilon} \preceq (1-\delta)^n\boldsymbol{\epsilon}$. Summing over n yields

$$
\Phi\boldsymbol{\epsilon} = \sum_{n=0}^{\infty} \Gamma^n\boldsymbol{\epsilon} \preceq \sum_{n=0}^{\infty} (1-\delta)^n\boldsymbol{\epsilon} = \tfrac{1}{\delta}\boldsymbol{\epsilon}
$$

which, when combined with (58), yields (60). $\qquad\square$