Differentially Private Maximal Information Coefficients

John Lazarsfeld [1] **Aaron Johnson** [2] **Emmanuel Adeniran** [1]

Abstract

The Maximal Information Coefficient (MIC) is a powerful statistic to identify dependencies between variables. However, it may be applied to sensitive data, and publishing it could leak private information. As a solution, we present algorithms to approximate MIC in a way that provides differential privacy. We show that the natural application of the classic Laplace mechanism yields insufficient accuracy. We therefore introduce the MICr statistic, which is a new MIC approximation that is more compatible with differential privacy. We prove MICr is a consistent estimator for MIC, and we provide two differentially private versions of it. We perform experiments on a variety of real and synthetic datasets. The results show that the private MICr statistics significantly outperform direct application of the Laplace mechanism. Moreover, experiments on real-world datasets show accuracy that is usable when the sample size is at least moderately large.

1. Introduction

The Maximal Information Coefficient (MIC) is a powerful and relatively new tool to detect correlations in data (Reshef et al., 2011; 2016). MIC uses mutual information to detect general dependencies between numeric attributes, in contrast to a more common statistic such as Pearsons correlation coefficient, which is only designed to detect linear relationships. MIC is thus particularly suited to identify novel relationships in complex data, as in that setting it is unknown which properties might be related and how.

However, many datasets that are valuable for such data mining (such as medical or economic data) contain sensitive personal information. Moreover, even publishing just the statistics that result from a correlation analysis can reveal private details of the individuals comprising the data (Homer et al., 2008; Wang et al., 2009). A scientist or government analyst must therefore consider not only the effectiveness of their statistical techniques, but they must also take into account if the resulting statistics can be published in a way that protects privacy.

Differential privacy (DP) (Dwork & Roth, 2014) has emerged as the leading method for privacy-preserving data publishing. Major companies and institutions use differential privacy to publish statistics about their sensitive data (Cormode et al., 2018; Dwork, 2019). MIC appears at first to be well-suited to being published in a differentially private way. At a high level, for a pair of numeric variables, it measures the maximum mutual information over possible *grids* partitioning their joint range. For any given grid, the effect of changing just one data point, which is the the main criterion for differential privacy, changes the distribution of points very little. This fact suggests that a differentially private MIC could be designed with high accuracy.

However, MIC is difficult to compute, and instead it is suggested to approximate it using the MICe statistic (Reshef et al., 2016). MICe is efficiently computable because it restricts its optimization to subgrids of a master *mass equipartition* of the dataset, that is, a grid in which each row and column contains the same number of data points. This master grid depends on the data, though, and we obtain a bound on the change in MICe from altering one data point that is significantly higher than would be expected for MIC.

We therefore investigate a new method to approximate MIC that is less sensitive to small changes in the dataset. We propose the MICr statistic, which optimizes over subgrids of a master *range equipartition*, that is, a grid in which the rows and columns divide the range equally. MICr is efficiently computable, and we obtain a bound on its sensitivity to input perturbations that is lower than our bound for MICe, both asymptotically and concretely. We prove that MICr converges in probability to MIC (or, more properly, to the analogous statistic defined over distributions).

We then present two differentially private versions of MICr, representing fundamentally different approaches to adding DP noise. MICr-Lap uses the classic Laplace mechanism, which adds random noise to the function *output* (in our case, the non-private MICr value). In contrast, MICr-Geom

[1]Department of Computer Science, Yale University [2]U.S. Naval Research Laboratory. Correspondence to: John Lazarsfeld <john.lazarsfeld@yale.edu>.

Proceedings of the 39^{th} *International Conference on Machine Learning*, Baltimore, Maryland, USA, PMLR 162, 2022.

perturbs the *input*, accomplished by adding a random count sampled from the geometric distribution to each grid cell, before then computing MICr. We prove that the error added to MICr by each of these mechanisms goes to zero as the size of the dataset increases.

Finally, we implement and experimentally analyze the two MICr mechanisms and the naive application of the Laplace mechanism to MICe (MICe-Lap). Our experiments use the synthetic and real datasets used to evaluate the original MIC statistics. The results show that the MICr mechanisms both significantly outperform MICe-Lap. Comparing the MICr mechanisms, we observe that MICr-Lap has lower bias but higher variance than MICr-Geom. Moreover, the experiments on real-world datasets show usable accuracy for a reasonable level of privacy when the sample size is at least moderately large. For example, with $\epsilon = 1$ we obtain average errors as low as 0.068 and 0.016 on datasets with 337 and 4381 datapoints, respectively, where MIC itself can range from 0 to 1 (low to high correlation).

2. Preliminaries

We begin by introducing notation and concepts that are used to define MIC and its related approximations.

2.1. Datasets, Grids, and Distributions

We define a *dataset* D as a sequence of n points in $\mathbb{R}^2$. To distinguish between x and y coordinates, we write $D = (D_x, D_y)$, where D_x, D_y are sequences of n points in $\mathbb{R}$.

For integers $k, \ell \geq 2$, a $k \times \ell$ *grid* $G = (P, Q)$ on $\mathbb{R}^2$ is comprised of a size-k partition $P = \{P_1, \ldots, P_k\}$ of the y-axis, and a size-ℓ partition $Q = \{Q_1, \ldots, Q_\ell\}$ of the x-axis. We say that a $k \times \ell$ grid G has $k\ell$ total *cells*, and we let $\mathcal{G}(k, \ell)$ denote the set of all grids with $k\ell$ total cells.

For $k \leq j$, a single-axis partition $P = \{P_1, \ldots, P_k\}$, is a *subpartition* of $C = \{C_1, \ldots, C_j\}$ (denoted by $P \subseteq C$) if every P_i is the union of adjacent intervals in C. For example, let P={[0, 0.2), [0.2, 0.5), [0.5, 0.8), [0.8, 1]} be a size-4 partition of the interval [0, 1]. Then Q={[0, 0.5), [0.5, 1]} is a size-2 subpartition of P because each element of Q is the union of two elements of P that are adjacent on $[0, 1]$.

We call P a size-k *range equipartition* of an interval $I = [x_0, x_1]$ when P is a size-k partition of I and all parts of P are intervals of length $(x_1 - x_0)/k$, We call a grid $G = (P, Q)$ a *subgrid* of $\Gamma = (C_x, C_y)$ (denoted by $G \subseteq \Gamma$) if P and Q are subpartitions of C_x and C_y respectively.

For a $k \times \ell$ grid $G = (P, Q)$, and for a point $d = (d_x, d_y) \in \mathbb{R}^2$, we define the point-mapping function ϕ, where $\phi(d, G) = (i, j)$ iff $d_x \in Q_j$ and $d_y \in P_i$. Then for a dataset D of n points and a $k \times \ell$ grid G, we define $\mathbf{A}_{D,G} \in \mathbb{Z}^{k \times \ell}$ as the *count matrix* for D and G. For all

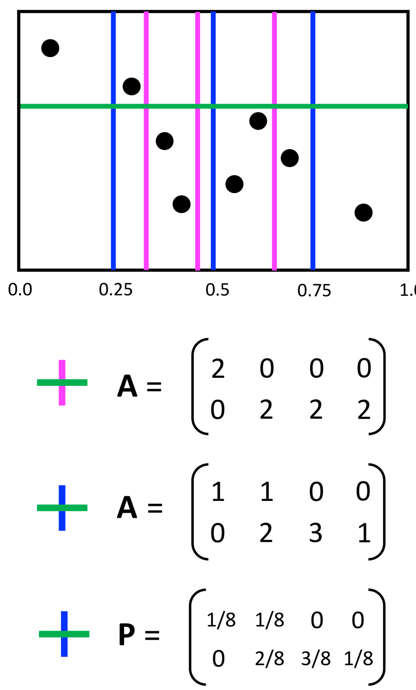

Figure 1. An example dataset of n=8 points (black dots). The pink vertical lines represent a size-4 *mass* equipartition, and the blue vertical lines represent a size-4 *range* equipartition of the interval $[0, 1]$. The first matrix shows the count matrix for the grid comprised of the pink column partition and green row partition. The second matrix shows the count matrix for the grid comprised of the blue column partition and green row partition. Notice how the count matrix can vary depending on whether the column partition is a mass (pink) or range (blue) equipartition. The final matrix is the normalized count matrix for the blue/green grid.

$(i, j) \in [k] \times [\ell]$, $\mathbf{A}_{D,G}$ has entries

$$a(i, j) = |\{d \in D \ : \ \phi(d, G) = (i, j)\}| \,,$$

meaning $\sum_{i,j} a(i, j) = n$. For a $k \times \ell$ grid $G = (P, Q)$, we call P (resp. Q) a *mass equipartition* if all row sums (resp. column sums) of $\mathbf{A}_{D,G}$ are equal.

If G is a subgrid of a grid Γ, observe that each entry of $\mathbf{A}_{D,G}$ can be "generated" by summing adjacent entries in $\mathbf{A}_{D,\Gamma}$, and we write $\psi(\mathbf{A}_{D,\Gamma}, \Gamma, G) = \mathbf{A}_{D,G}$ to denote the matrix whose entries are generated via this process.

Given $\mathbf{A}_{D,G}$, the matrix $\mathbf{P}_{D,G} \in \mathbb{R}^{k \times \ell}$ with entries $p(i, j)$ is the *normalized count matrix* for D and G, where

$$\mathbf{P}_{D,G} = (1/n) \cdot \mathbf{A}_{D,G}.$$

We use $p(i, *) = \sum_j p(i, j)$ and $p(*, j) = \sum_i p(i, j)$ to denote the row and column sums of $\mathbf{P}_{D,G}$ respectively.

Figure 1 shows an example dataset to illustrate the differences between mass and range-based equipartitions and their resulting count matrices.

For a fixed D and G of size $k \times \ell$, we write $D|_G$ (read as *D partitioned by G*) to denote the joint distribution over the space $[k] \times [\ell]$, with a probability mass function (PMF) given by the entries of $\mathbf{P}_{D,G}$. Then the discrete mutual information of $D|_G$ (equivalently, of $\mathbf{P}_{D,G}$) is computed by

$$I(D|_G) = I(\mathbf{P}_{D,G}) = \sum_{i,j} p(i,j) \log_2 \frac{p(i,j)}{p(i,*)p(*,j)}.$$

We then define the normalized $I^\star(D|_G)$ by

$$I^\star(D|_G) := \frac{I(D|_G)}{\log_2 \min\{k,\ell\}},$$

which ensures $I^\star(\cdot)$ is always in the range $[0,1]$.

For a jointly distributed pair of random variables $\Pi = (X,Y)$ and a $k \times \ell$ grid G, we define $\Pi|_G$ and $\mathbf{P}_{\Pi,G}$ similarly, where $\mathbf{P}_{\Pi,G}$ has entries $p(i,j) = \Pr(X \in Q_j, Y \in P_i)$. So $\Pi|_G$ is a discrete probability distribution over $[k] \times [\ell]$, and we compute $I^\star(\Pi|_G)$ as above.

2.2. The MIC, MIC*, and MICe Statistics

Given a dataset D, the MIC statistic measures correlation by finding the grid G where $I^\star(D|_G)$ is maximized. If the variables represented by D have some relationship, MIC will identify the grid containing the intervals where each variable gives the most information about the other.

For any D with n points, it is easy to see that (assuming no points with the same x or y value) there exists an $n \times n$ grid G that separates each point into its own cell, which means $I^\star(D|_G) = 1$. On the other hand, grids with too few total cells may not be able to capture more complex relationships. To negotiate this tradeoff, the MIC statistic requires a *maximum grid size* parameter $B := B(n)$, where only grids with at most $B(n)$ cells are considered, but where $B(n)$ is expected to grow with n. Defined formally:

Definition 2.1 (MIC statistic (Reshef et al., 2011))**.** For a dataset D of size n and $B := B(n)$, let $\mathbf{M}_D^{\mathcal{G}}$ denote its *characteristic matrix* with (k,ℓ) entries $\max_{G \in \mathcal{G}(k,\ell)} I^\star(D|_G)$. Then $\text{MIC}(D,B) = \max_{k,\ell:\ k\ell \le B(n)} \left(\mathbf{M}_D^{\mathcal{G}}\right)_{k,\ell}$.

When the dataset D can be modeled as a sample of points drawn from a joint distribution Π, then MIC can be viewed as an *estimator* of the analogous MIC* statistic defined for the distributional setting:

Definition 2.2 (MIC* statistic (Reshef et al., 2016))**.** For a jointly-distributed pair of random variables $\Pi = (X,Y)$, let $\mathbf{M}_\Pi^{\mathcal{G}}$ denote its characteristic matrix with (k,ℓ) entries $\max_{G \in \mathcal{G}(k,\ell)} I^\star(\Pi|_G)$. Then $\text{MIC*}(\Pi) = \sup \mathbf{M}_\Pi^{\mathcal{G}}$.

When $B(n) = O(n^\alpha)$ for some $\alpha \in (0, 0.5)$, MIC is also a *consistent* estimator of MIC*, meaning that for a dataset D_n of n points drawn i.i.d. from Π, MIC(D_n, B) converges in probability to MIC*(Π) as $n \to \infty$ (Reshef et al., 2016; Lazarsfeld & Johnson, 2021). However, for every k, ℓ with $k\ell \le B(n)$, MIC is a maximization over the set $\mathcal{G}(k,\ell)$ of *all* grids with $k\ell$ cells, which for datasets of size n makes computing MIC infeasible in practice.

To this end, Reshef et al. (2016) introduced an efficiently-computable *approximation* of MIC called MICe. This newer statistic approximates MIC by defining its characteristic matrix entries as maximizations over *dataset-dependent subsets* of $\mathcal{G}(k,\ell)$. Specifically, for a dataset $D = (D_x, D_y)$, a constant $c > 0$, and for all $2 \le k \le \ell$, the set $\mathcal{E}(D,c,k,\ell)$ contains all grids $G = (P,Q)$ where Q is a size-ℓ mass equipartition of D_x and P is a size-k subpartition of a size-$c\ell$ mass equipartition of D_y. The set is defined symmetrically for $k > \ell$. Then MICe is defined similarly to MIC but replaces the maximizations over $\mathcal{G}(k,\ell)$ with $\mathcal{E}(D,c,k,\ell)$:

Definition 2.3 (MICe statistic, Reshef et al. (2016))**.** For any dataset D of n points, constant $c > 0$, and $B := B(n)$, let $\mathbf{M}_D^{\mathcal{E}(D,c)}$ be the *equicharacteristic* matrix with (k,ℓ) entries $\max_{G \in \mathcal{E}(D,c,k,\ell)} I^\star(D|_G)$. Then

$$\text{MICe}(D,B,c) = \max_{k,\ell:\ k\ell \le B(n)} \left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell}.$$

Each (k,ℓ) entry of the equicharacteristic matrix can be reduced to a maximization defined only over subpartitions on a *single* axis. This maximization can be computed efficiently using the dynamic-programming algorithm OptimizeAxis of Reshef et al. (2011). Then in total, MICe(D,B,c) can be computed in $O(c^2B^4)$ time for any B and c, and this becomes $O(c^2 n^{4\alpha})$ when $B(n) = O(n^\alpha)$. Here, the constant c can be viewed as an approximation parameter, where a larger value leads to a better approximation of MIC, but at the expense of slower computation[1]. More details of computing MICe are given in Appendix A.1.

In addition to its efficiency, MICe is still a consistent estimator of MIC* when $B(n) = O(n^\alpha)$ for $\alpha \in (0, 0.5)$ (Reshef et al., 2016; Lazarsfeld & Johnson, 2021). In practice, Reshef et al. (2018; 2016) suggested using α=0.6 and c=15 after evaluating the statistic on synthetic and real data, and so we treat these as its default settings. Given its computational efficiency *and* consistency, we consider MICe as a starting point for developing private approximations of MIC, but we first briefly recall a few definitions and tools related to differential privacy.

[1] Note that Reshef et al. (2016) originally defined MICe with a more complicated dependence on the parameter B to constrain the maximization space. The result is faster computation but with lower accuracy for a given n, though in practice the two variants are similar. We discuss this further in Appendix E.1.

2.3. Differential Privacy

For two datasets $D = (d_1, \ldots, d_n)$ and $D' = (d'_1, \ldots, d'_m)$, we say D and D' are *neighboring* (denoted $D \sim D'$) if $n = m$ and there exists at most one index i where $d_i \neq d'_i$. Differential privacy ensures that the output distributions of a (randomized) algorithm are similar when run on $D \sim D'$:

Definition 2.4 (Dwork & Roth (2014))**.** For $\epsilon > 0$, a randomized algorithm $\mathcal{A} : \mathbb{R}^{2\times n} \to \mathbb{R}$ is ϵ-differentially private (ϵ-DP) if, for every $D \sim D'$ and every $I \subseteq \mathbb{R}$, $\Pr(\mathcal{A}(D) \in I) \leq \exp(\epsilon) \Pr(\mathcal{A}(D') \in I)$.

In our setting, using an ϵ-DP mechanism to estimate MIC implies that the output leaks little information about any $d_i \in D$ (see § 7 for more on the privacy semantics).

One common approach for designing ϵ-DP mechanisms is to add random noise to the output of a non-private function. When doing so, an important consideration is the *sensitivity* of the function, which is the maximum possible change in function value over neighboring $D \sim D'$:

Definition 2.5 (Sensitivity)**.** The (ℓ_1) *sensitivity* of a function $f : \mathbb{R}^{2\times n} \to \mathbb{R}$ is $\max_{D\sim D'} |f(D) - f(D')|$.

For a function f with sensitivity Δ, the Laplace mechanism of Dwork et al. (2006) adds noise from a zero-mean Laplace distribution with parameter Δ/ϵ. The result is ϵ-DP.

For convenience, we will write $\text{Lap}(b)$ to denote a random variable with a zero-mean Laplace distribution with parameter b, which has density $f(x) = e^{-|x|/b}/(2b)$. Also, for $x \in \mathbb{R}$, $[x]_{0,1}$ denotes that x is tructated to the range $[0, 1]$.

3. MICe-Lap Mechanism

We begin by considering the compatability of MICe with the standard Laplace mechanism. Doing so requires obtaining a bound on the sensitivitiy of MICe, which for datasets of size n, $B(n)$, and $c > 0$ we denote by $\Delta_n(\text{MICe}, B, c)$. Our first result gives an upper bound on this quantity.

Theorem 3.1. *For any $B := B(n)$, $c > 0$, and $n \geq 6$,* $\Delta_n(\text{MICe}, B, c) \;\leq\; B \cdot ((2\log_2 n)/n + 4.8/n)$.

Using the suggested setting $B(n) = n^\alpha$ (Reshef et al., 2011; 2016), $\Delta_n(\text{MICe}, B, c) = O((\log_2 n)/n^{1-\alpha})$, which is asymptotically larger than the $O(\log_2 n/n)$ sensitivity of MIC for $\alpha \in (0, 1)$ (this MIC sensitivity bound can be obtained from the proof of Theorem 4.3). Theorem 3.1 is proved in Appendix A.2, and the main intuition is that because MICe maximizes over a dataset-dependent set of grids, for $D \sim D'$, a worst-case change in $I^\star$ can occur when the grids for D have no cells in common with those of D', which changes $I^\star$ a bit for each of the at most B cells.

This sensitivity bound then yields the ϵ-DP MICr-Lap mechanism by adding Laplace noise and truncating:

Mechanism 1 (MICe-Lap)**.** For any dataset D of size $n \geq 6$, $B := B(n)$, $c > 0$, and $\epsilon > 0$, let $\Delta := B \cdot ((2\log_2 n/n) + 4.8/n)$. Then $\text{MICe-Lap}(D, B, c, \epsilon) := [\text{MICe}(D, B, c) + \text{Lap}(\Delta/\epsilon)]_{0,1}$.

Theorem 3.2. $\text{MICe-Lap}(\cdot, B, c, \epsilon)$ *is ϵ-DP.*

Theorem 3.2 is proved in Appendix A. Given that the Laplace sampling can be done in constant time, the runtime of computing MICe-Lap is asymptotically equivalent to that of MICe because only output noise is added.

Additionally, due to the $[0, 1]$ truncation, the standard deviation of the mechanism is bounded by that of $\text{Lap}(\Delta/\epsilon)$. So for $B = n^\alpha$, the mechanism's standard deviation is at most $(\sqrt{2}/\epsilon) \cdot ((2\log_2 n/(n^{1-\alpha})) + (4.8/n^{1-\alpha}))$. While for a fixed ϵ this value decreases with n, using the suggested α=0.6, this quantity is 1.38 when ϵ=1 and n=5000. Given that the MICe value is in $[0, 1]$, the outputs of MICe-Lap have intolerably high error, even for moderately large values of ϵ and n. This result motivates designing an alternative MIC approximation with lower-error private variants.

4. MICr Mechanisms

4.1. Non-private MICr Statistic

Despite being efficiently computable and a consistent estimator, the dataset dependence and resulting high sensitivity of MICe precludes a straightforward, low-error private variant. To that end, we introduce the MICr approximation for MIC. It remains both efficiently computable and a consistent estimator of MIC*, but it optimizes over dataset-*independent* sets of grids, yielding a lower sensitivity. Indeed, its sensitivity matches that of MIC, making it more compatible with designing differentially private variants.

We begin by describing the (non-private) MICr statistic before introducing two differentially private variants. The key difference between MICe and MICr is that the former maximizes over mass equipartitions while the latter maximizes over *range* equipartitions. Therefore, in addition to a maximum grid size parameter B and a finite $c > 0$, MICr also takes as a parameter a range $L \subset \mathbb{R}^2$ of the form $L = [x_0, x_1] \times [y_0, y_1]$. When computing MICr, we require that the coordinates of all points of D lie within L. This requirement can be satisfied by most types of data, for example those with a defined range or using conservative maxima and minima estimates. L must be dataset-independent.

For any D restricted to range $L = L_x \times L_y$ and $c > 0$, the dataset-independent sets of grids used to define the entries of the analogous equicharacteristic matrix for MICr are defined as follows. First, for an interval $I \subset \mathbb{R}$, let $R_{I,\ell}$ denote the size-ℓ range equipartition of I, and let $\mathcal{P}(I, k, [j])$ denote the set of all size-k subpartitions of a size-j range equipartition of I. Then for $k \leq \ell$, we define $\mathcal{R}(L, c, k, \ell)$ as the set of all grids $G = (P, Q)$ where $Q = R_{L_x,\ell}$ and $P \in \mathcal{P}(L_y, k, [c\ell])$. When $k \leq \ell$, we call the grid $\Gamma_{L,c,\ell} = (R_{L_y,c\ell}, R_{L_x,\ell})$ the *master* range equipartition for $\mathcal{R}(L, c, k, \ell)$. When $k > \ell$, the set $\mathcal{R}(L, c, k, \ell)$ and the master grid $\Gamma_{L,c,k}$ are defined symmetrically. The full definition of MICr then follows similarly to MICe:

Definition 4.1 (MICr statistic)**.** For any dataset D of n points with range restricted to L, $c > 0$, and $B := B(n)$, let $\mathbf{M}_D^{\mathcal{R}(L,c)}$ denote the range equicharacteristic matrix with (k, ℓ)'th entry $\max_{G \in \mathcal{R}(L,c,k,\ell)} I^\star(D|_G)$. Then

$$\text{MICr}(D, L, B, c) = \max_{k,\ell:\ k\ell \leq B(n)} \left(\mathbf{M}_D^{\mathcal{R}(L,c)}\right)_{k,\ell}.$$

As with MICe, the MICr statistic can be computed in $O(c^2 B^4)$ time using the OptimizeAxis routine, and c can again be viewed as an approximation parameter. We also prove that MICr is still a consistent estimator of MIC* when $B(n) = O(n^\alpha)$ for $\alpha \in (0, 0.5)$.

Theorem 4.2. MICr *is a consistent estimator of* MIC*.

Details of the computation and runtime of MICr appear in Appendix B.1, and a more formal statement of the consistency result is given by Theorem B.5 in Appendix B.2.

In addition to consistency, we leverage the fact that the sets $\mathcal{R}(L, c, k, \ell)$ are dataset-independent to prove an upper bound on the ℓ_1 sensitivity of $\text{MICr}(\cdot, L, B, c)$, which for datasets of size n we denote by $\Delta_n(\text{MICr}, L, B, c)$.

Theorem 4.3. *For any L, $c > 0$, $B := B(n)$, and $n \geq 4$, $\Delta_n(\text{MICr}, L, B, c) \leq (4 \log_2 n)/n + 6/n$.*

Compared to the sensitivity bound for MICe, the bound here loses the multiplicative B factor. This is because the sets of grids considered by MICr are *fixed* for all datasets of the same size, and for any G and $D \sim D'$, the count matrices $\mathbf{A}_{D,G}$ and $\mathbf{A}_{D',G}$ can differ by at most 1 at exactly 2 entries. The proof of the theorem is developed in Appendix B.3.

Equipped with the fact that MICr is both a consistent estimator of MIC* *and* has low sensivity, MICr is a more suitable base statistic for designing high-utility private variants. To this end, because two classical approaches to designing differentially private mechanisms (*output perturbation*, as with the Laplace mechanism, and *input perturbation*, as with private histograms (Dwork & Roth, 2014)) appear to have equal potential for achieving this task, we designed two private variants, each one following a different approach.

4.2. MICr-Lap Mechanism

The first private variant is analogous to MICe-Lap, but using the smaller MICr sensitivity bound from Theorem 4.3.

Mechanism 2 (MICr-Lap)**.** For any dataset D of size $n \geq 4$ restricted to L, any $B := B(n)$, any $c > 0$ and any $\epsilon > 0$, let $\Delta := (4 \log_2 n)/n + 6/n$. Then

$$\text{MICr-Lap}(D, L, B, c, \epsilon) := \left[\text{MICr}(D, L, B, c) + \text{Lap}(\Delta/\epsilon)\right]_{0,1}.$$

Theorem 4.4. $\text{MICr-Lap}(\cdot, L, B, c, \epsilon)$ *is ϵ-DP.*

Again using the standard deviation of the Laplace mechanism, for any B, the standard deviation of MICr-Lap is at most $(\sqrt{2}/\epsilon) \cdot ((4 \log_2 n)/n + 6/n)$. When n=5000 and ϵ=1 this value is 0.02, which means (as we show in Section 5) that the error of the private output is likely small enough for the mechanism to be useable in practice.

Also, because for fixed ϵ the standard deviation is decreasing with n, and using the consistency of MICr, it is straightforward to show that MICr-Lap is still a consistent estimator of MIC*, and we prove this in Theorem C.1 in Appendix C.

4.3. MICr-Geom Mechanism

Our second private variant, MICr-Geom, adds noise during the computation of each entry in the MICr-Geom range-equicharacteristic matrix.

Recall for subpartitions $G \subseteq \Gamma$ that the count matrix $\mathbf{A}_{D,G}$ (and thus $\mathbf{P}_{D,G}$) can be generated via the function $\psi(\mathbf{A}_{D,\Gamma}, \Gamma, G)$, which doesn't depend directly on the coordinates of D. Then the k, ℓ entry of the range-equicharacteristic matrix for MICr-Geom is a maximization of $I^\star$ over grids $G \in \mathcal{R}(L, c, k, \ell)$, but where the count matrix for each grid is generated via $\psi(\widehat{\mathbf{A}}, \Gamma, G)$, where $\widehat{\mathbf{A}}$ is some *noisy* approximation of $\mathbf{A}_{D,\Gamma}$ and $\Gamma := \Gamma_{L,c,\ell}$ is the master range-equipartition for $\mathcal{R}(L, c, k, \ell)$.

Specifically, as established earlier, for any neighboring $D \sim D'$ of size n, at most two corresponding entries of $\mathbf{A}_{D,\Gamma}$ and $\mathbf{A}_{D',\Gamma}$ can differ, and they differ by at most 1. Thus we generate a noisy version of $\mathbf{A}_{D,\Gamma}$ by producing independent, noisy estimates for each of its entries using the ϵ-DP Truncated Geometric mechanism of Ghosh et al. (2012), which applies to counts with sensitivity at most 1. Intuitively, the $\text{TruncGeom}(\epsilon, n, f)$ is a doubly-geometric distribution centered at f with parameter $e^{-\epsilon}$, and with truncation at f and $n - f$. We summarize the distribution here:

Definition 4.5 (TruncGeom, (Ghosh et al., 2012))**.** For any $\epsilon > 0$, n, and $0 \leq f \leq n$, let $\text{TruncGeom}(\epsilon, n, f)$ be a discrete distribution over $\{0, \ldots, n\}$. Set $\rho := e^{-\epsilon}$. Then:

- $\text{TruncGeom}(\epsilon, n, f) = 0$ w.p. $\rho^f/(1 + \rho)$.
- $\text{TruncGeom}(\epsilon, n, f) = i$ w.p. $((1 - \rho)/(1 + \rho))\rho^{|f-i|}$ for all $1 \leq i \leq n - 1$.
- $\text{TruncGeom}(\epsilon, n, f) = n$ w.p. $\rho^{(n-f)}/(1 + \rho)$.

We then define the MICr-Geom mechanism formally and give its privacy guarantee as follows:

Mechanism 3 (MICr-Geom)**.** Fix any dataset D of size n restricted to L, any $c > 0$, any $B := B(n)$, and any $\epsilon > 0$.

1. For every $k, \ell \geq 2$, let Γ denote the master range-equipartition grid for $\mathcal{R}(L, c, k, \ell)$. Let $\widehat{\mathbf{A}} := \widehat{\mathbf{A}}^{\epsilon}_{D,\Gamma}$ be the *noisy* count matrix whose (i, j) entry is given by TruncGeom$(\epsilon/2, n, a(i, j))$ (where $a(i, j)$ is the corresponding entry of $\mathbf{A}_{D,G}$).

   Let $\widehat{n}$ be the sum of entries in $\widehat{\mathbf{A}}$, and let $\widehat{\mathbf{P}} = (1/\widehat{n}) \cdot \widehat{\mathbf{A}}$.
2. For every $G \in \mathcal{R}(L, c, k, \ell)$, let $\widehat{\mathbf{P}}_G := \psi(\widehat{\mathbf{P}}, \Gamma, G)$.
3. Let $\widehat{\mathbf{M}}^{\mathcal{R}(L,c)}_{D,\epsilon}$ be the *noisy* range-equicharacteristic matrix with (k, ℓ) entry $\max_{G\in\mathcal{R}(L,c,k,\ell)} I^{\star}(\widehat{\mathbf{P}}_G)$. Then:

$$\text{MICr-Geom}(D, L, B, c, \epsilon) = \max_{k,\ell:\ k\ell\leq B(n)} \left(\widehat{\mathbf{M}}^{\mathcal{R}(L,c)}_{D,\epsilon}\right)_{k,\ell}.$$

Theorem 4.6. MICr-Geom(D, L, B, c, ϵ) *is ϵ-DP.*

With a linear-time additive preprocessing step, samples from the TruncGeom distribution can be done in constant time, and thus the running time of computing MICr-Geom is asympotically equivalent to MICr and MICr-Lap. The details of both the privacy statement and the runtime analysis are given in Appendices D.1 and D.2 respectively.

Additionally, we prove that for a fixed ϵ and c, the error introduced from using the noisy count matrices $\mathbf{A}^{\epsilon}_{D,\Gamma}$ goes to 0 as n grows:

Theorem 4.7 (Added error of MICr-Geom)**.** *Fix any $\alpha \in (0, 0.5)$, finite $c > 0$, $\epsilon > 0$, and dataset D of size n. For sufficiently large n, there exists some $a > 0$ such that*

$$\left|\left(\widehat{\mathbf{M}}^{\mathcal{R}(L,c)}_{D,\epsilon}\right)_{k,\ell} - \left(\mathbf{M}^{\mathcal{R}(L,c)}_{D}\right)_{k,\ell}\right| = O((c/\epsilon)n^{-a})$$

for all $k\ell \leq B(n) = O(n^{\alpha})$ simultaneously with probability at least $1 - O(n^{-2})$.

Intuitively, the dependence on c in the error bound is a result of having a master range-equiparitition size of $c\ell^2$ (wlog when $k \leq \ell$). So choosing larger c yields a better approximation of MIC but results in a "noisier" $\widehat{\mathbf{A}}$ (and thus a slower convergence to the non-private MICr). We explore this tradeoff more experimentally in Section 5. Together with the fact that MICr is a consistent estimator of MIC* (Theorem B.5), it is straightforward to prove that MICr-Geom is also a consistent estimator of MIC*. We prove this in Theorem D.4 in Appendix D.3.

5. Experimental Evaluation

To investigate their utility, we evaluated the three private mechanisms on both synthetic and real datasets. We used synthetic data (following the methodology of Reshef et al. (2011; 2016)) to help better understand the error of our mechanisms at growing sample sizes over a *variety* of distributions designed to capture different types of relationships. Using real data helps to additionally verify the utility of our mechanisms in practice at specific, fixed sample sizes.

The code and data used to obtain our experimental results can be accessed at https://github.com/jlazarsfeld/dp-mic, and more implementation details are given in Appendix E.1.

5.1. Synthetic Data

We considered the family of 21 functional relationships introduced by Reshef et al. (2011). Similar to their later work (Reshef et al., 2016), for every relationship we defined 9 joint distributions, each generated by placing k=100 independent bivariate Gaussian distributions (with zero correlation and identical variances) centered at points evenly-spaced along the function graph. The 9 distributions were parameterized by an R^2 value in $\{0.1, 0.2, \ldots, 0.9\}$ that determined the variances of each Gaussian. The result is a set $\mathcal{Q}$ of 189 joint distributions bounded in range by $[0, 1] \times [0, 1]$, each representing a functional relationship with varying levels of noise. For each $\Pi \in \mathcal{Q}$, we computed an approximation of MIC*(Π) using the (provably convergent) method of Reshef et al. (2016). For n ranging from 25 to 10,000 and $\epsilon \in \{0.1, 1.0\}$, we measured the accuracy of each mechanism (wrt MIC*(Π)) on datasets of n points sampled i.i.d. from Π. The full details of this dataset generation process are included in Appendix E.2.

Given that MICe-Lap simply adds noise to MICe, we set its parameters to $B(n) = n^{0.6}$ and $c = 15$, matching the suggested settings for MICe. However, because MICr-Lap and MICr-Geom use a different base statistic, we evaluated these mechanisms with varying B and c to better determine optimal parameter settings for a fixed n and ϵ.

Parameter Tuning for MICr-Geom, MICr-Lap: We considered sample sizes of $n \in$ {25, 250, 500, 1000, 5000, 10000}, $\epsilon \in \{0.1, 1.0\}$, and various values of B between 4 and 150. Although the consistency guarantees for MICr-Geom and MICr-Lap are phrased in terms of $B := B(n) = O(n^{\alpha})$, defining B in absolute terms helps us better determine optimal values of B for each mechansim. For computational considerations, we fixed $c = 5$ for the MICr-Lap mechanism (we found the mechanism to be insensitive to larger values of c), and for the MICr-Geom mechanism we considered $c \in \{1, 2\}$ (under the consideration that larger c could worsen accuracy).

For each combination of (n, ϵ, B, c), distribution $\Pi \in \mathcal{Q}$, and mechanism, we ran 50 iterations of the following process: (1) construct a dataset D by sampling n points i.i.d. from Π, and (2) run the private mechanism on D. For each mechanism, n, and ϵ, we minimized an objective function

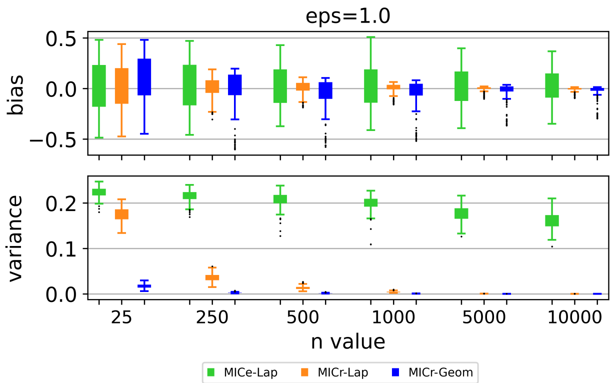

Figure 2. Boxplots of the bias and variance of each mechanism (over 50 iterations) over all $\Pi \in \mathcal{Q}$ for ϵ=1 and varying n.

over the (B, c) parameters that involved a weighted sum of the mechanism's average absolute error (wrt MIC*) across all distributions in $\mathcal{Q}$. The objective function was designed to choose parameters that could ensure parity in error across both low and high-correlation distributions, and the exact description is given in Appendix E.2.

The optimized B and c parameters for MICr-Lap and MICr-Geom are summarized in Table 2 in Appendix E.2. For both mechanisms and ϵ, the optimal B values are generally increasing with n, which aligns with the intuition that the mechanisms converge toward MIC* with larger n. The optimal values ranged from 12 to 150 for B and 1 to 5 for c.

Bias/Variance Evaluation: Using the parameters from Table 2 for MICr-Lap and MICr-Geom and $(B(n), c) = (n^{0.6}, 15)$ for MICe-Lap, for each mechanism we compared the bias (average signed error wrt to MIC*) and variance (of the 50 private iterations for a fixed distribution) over *all* 189 distributions in $\mathcal{Q}$ as n grows. The results for ϵ=1 are summarized in Figure 2, and analogous plots for ϵ=0.1 are given in Figure 4 of Appendix E.2. In both subplots in Figure 2, for every value of n, we show boxplots of the bias (resp. variance) for each mechanism over *all* $\Pi \in \mathcal{Q}$. Recall each box represents the interquartile range (IQR, 25'th to 75'th quantiles) of the data, the whiskers appear above and below these quantiles by an additional 1.5x of the IQR, and outlier points beyond the whiskers are plotted individually.

For ϵ=1, the IQR of the bias of each mechanism generally drops with n, and this is most prominent for MICr-Lap and MICr-Geom. In particular, when n=5000, the median bias of MICr-Lap is 0.01 with min and max biases of -0.1 and 0.02, which we consider tolerably low. On the other hand, while the IQR of the bias of MICr-Geom at n=5000 is also small, the outlier points indicate that the mechanism has large *negative* bias on a subset of distributions in $\mathcal{Q}$. Specifically, while the median and max bias of MICr-Geom at n=5000 are 0.01 and and 0.04 respectively, the min bias is -0.37. Because the MICr-Geom mechanism generates noisy counts for *every* cell of a master grid, we found this negative bias to occur on datasets for which the non-noisy master count matrix contains a large submatrix with mostly zero entries, and this corresponds to datasets restricted to L that have large sub-regions with no points. When ϵ=0.1, Figure 4 (Appendix E.2) shows similar trends for the bias of each mechanism, but where decreases with n are slower.

Additionally, the variances of MICr-Lap and MICr-Geom are significantly smaller than for MICe-Lap. For MICr-Geom this variance is particularly low, even for small n. For example, for n=250 and ϵ=1, the median variance of MICr-Geom is 0.003 and these medians decrease for larger n. In general, we observe (especially for smaller n) a bias/variance tradeoff between MICr-Lap (less bias, more variance) and MICr-Geom.

5.2. Real Data

We also evaluated the utility of the three private mechanisms on two sets of data used in the experiments of Reshef et al. (2011): the Spellman data and the Baseball data. Note that these datasets do not necessarily contain sensitive information, and these sources were chosen mainly because of their previous use by Reshef et al. (2011). Because both sets contain multiple columns, we constructed for each source a *collection* of datasets corresponding to different pairs of columns, and we first describe this process in more detail.

Dataset Description: The Spellman data (Spellman et al., 1998; Reshef et al., 2011) contains gene expression measurements for 4381 genes in the yeast organism, where each gene has a timeseries (at common, fixed time points) of n=23 measurments. Modeling the methodology of Reshef et al. (2011), for each gene, we consider the dataset $D = ([23], T_i)$, where T_i is the timeseries for the i'th gene. For each i, we let $l_i = |\max T_i - \min T_i|$, $y_0 = (\min T_i) - l_i/100$, $y_1 = (\max T_i) + l_i/100$, and we use the range bounds $L_i = [0, 24] \times [y_0, y_1]$. This results in a collection of $m = 4381$ datasets D, each of size $n = 23$, and we refer to this as the Spellman23 collection.

Because the size of each dataset in Spellman23 is small (n=23), and given that we expect the error of our mechanisms to decrease with larger n, we also constructed a collection of higher-dimensional datasets from the Spellman source as follows: for each fixed time index $t \in [23]$, let C_t denote the set of measurements for all 4381 genes at time index t. Then for each unique pair $t, v \in [23]$, we constructed the dataset $D = (C_t, C_v)$. We set global range bounds by considering the maximum and minimum values across all T_i, denoted by ℓ_0 and ℓ_1 respectively, and by setting $x_0 = \ell_0 - |\ell_1 - \ell_0|/100$, $x_1 = \ell_1 + |\ell_1 - \ell_0|/100$ and $L = [x_0, x_1] \times [x_0, x_1]$ for all i. The result is a collection

	MICe-Lap	MICr-Lap	MICr-Geom
Spellman23	0.18 (0.23)	0.14 (0.17)	0.24 (0.01)
Baseball	0.30 (0.21)	0.02 (0.02)	0.06 (9e-4)
Spellman4381	0.26 (0.17)	-0.01 (4e-4)	0.02 (1e-4)

Table 1. The median bias (average signed error wrt MICe over 100 runs) and median variance (over 100 iterations) of each private mechanism across all datasets of each collection for ϵ=1.

of $m = 253$ datasets, each of size $n = 4381$. We refer to this as the Spellman4381 collection.

Finally, the Baseball dataset (Reshef et al., 2011; Prospectus, Accessed March 2020) contains values of 129 performance-related statistics for 337 players from the 2008 MLB season. Again following Reshef et al. (2011), for the i'th statistic, we let C_i denote the set of corresponding values across all 337 players, and for each unique pair $i, j \in [129]$, we constructed the dataset $D = (C_i, C_j)$. We generated range bounds $[x_0, x_1]$ for C_i and $[y_0, y_1]$ for C_j using the same process as for each T_i in Spellman23. We then set $L = [x_0, x_1] \times [y_0, y_1]$. The result is a collection of $m = 8256$ datasets of size n=337, which we refer to as the Baseball collection.

For the purposes of these experiments, the range bounds L for each collection of datasets were constructed manually. However, we assume in general that practitioners with domain-specific knowledge can set reasonable bounds for L *without* observing a particular dataset. For example, a practitioner with baseball acumen could set the range for the statistic "Number of Games Played" to $[-1, 183]$, since there are 182 games in an MLB season.

Evaluation of Private Error: For each dataset in the three collections described above, we measured the error of all private mechanisms with respect to the dataset's non-private MICe score using parameters $B(n) = n^{0.6}$ and $c = 15$. Although MICr shares similar theoretical properties as MICe, we primarily view MICr as a conduit for private approximations of MIC. Thus measuring private error with respect to the exisiting MICe statistic allows for a more realistic evaluation of our mechanisms' utility.

For every dataset, we ran 100 computations of each private mechanism for ϵ=1 and ϵ=0.1 (we mainly describe here the results for ϵ=1 but the corresponding tables and figures for ϵ=0.1 can be found in Appendix E.3). For MICr-Geom and MICr-Lap, we set the B and c parameters by linearly interpolating values from Table 2. For MICe-Lap, we set $B(n) = n^{0.6}$ and $c = 15$ (matching the settings from the synthetic data evaluation).

For each collection of datasets, Table 1 lists the median bias (average signed error wrt MICe) and median variance (over 100 iterations) of each private mechanism over *all* datasets

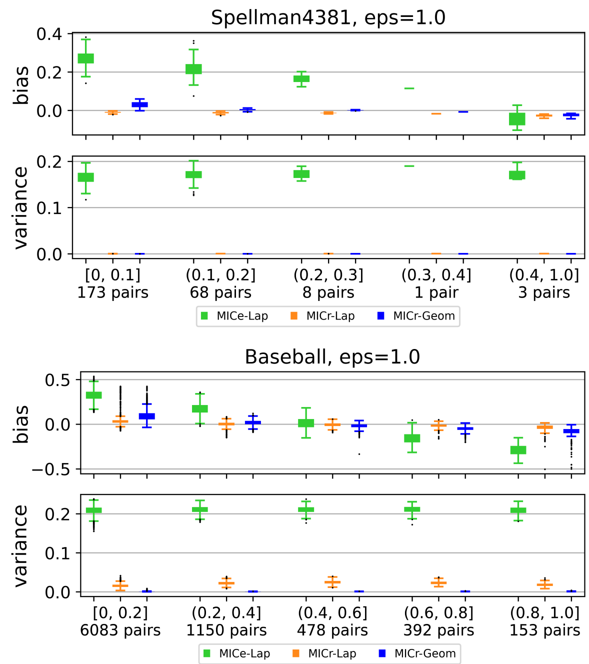

Figure 3. Bias and variance boxplots for each mechanism over datasets (pairs) in the Spellman4381 collection (top) and Baseball collection (bottom) binned by non-private MICe score for ϵ=1.

in the collection. For both MICr-Lap and MICr-Geom, the median bias drops significantly for the Baseball and Spellman4381 datasets compared to Spellman23. Similar to our evaluation on synthetic data, this supports the intutition that the error incurred due to the privatization of MICr (in both mechansims) decreases with n. Additionally, although MICe-Lap has relatively small median bias for Spellman23, its median average *unsigned* error was the largest (0.46 compared to 0.39 and 0.25 for MICr-Lap and MICr-Geom, respectively). While the median average unsigned errors for MICr-Lap and MICr-Geom on Spellman4381 (0.016 and 0.019) and on Baseball (0.097 and 0.068) are low, Table 1 shows these mechanisms still incur large error on Spellman23, which indicates that none of the private mechanisms are accurate enough for datasets as small as n=23.

On the other hand, to better understand the practicality of the mechansims in the higher-dimensional regime, we further analyze their error on datasets in the Baseball and Spellman4381 collections, with a particular focus on how error varies for low and high-correlation datasets (using low and high MICe as a proxy). To this end, each dataset of the Spellman4381 collection is *binned* according to its non-private MICe score using interval endpoints $[0, 0.1, 0.2, 0.3, 0.4, 1]$. Similarly, each dataset in the Baseball collection is binned by its MICe score using interval endpoints $[0, 0.2, 0.4, 0.6, 0.8, 1]$. For both collections, we analyzed the bias and variance of each mechanism across all pairs in a *fixed bin*. Figure 3 displays bias and variance boxplots for each mechanism using ϵ=1 for both collections, and similar plots for ϵ=0.1 are provided in Appendix E.3.

For both collections, the IQRs for bias for both MICr-Lap and MICr-Geom in every bin are smaller and with medians closer to zero than those of MICe-Lap. In addition, in both collections the variances for MICr-Lap and MICr-Geom are significantly smaller across all bins compared to MICe-Lap, and in general we see a bias/variance tradeoff between MICr-Lap and MICr-Geom similar to the synthetic data.

For MICr-Geom, we also notice on both collections that for datasets with lowest MICe scores, the mechanism tends to have a slightly more *positive* bias (medians of 0.03 and 0.09 for Spellman4381 and Baseball) compared to MICr-Lap (medians of -0.01 and 0.03). This positive bias of MICr-Geom likely occurs because generating private counts using TruncGeom on datasets with low MICe scores (which have more "uniform" scatterplots), can *reduce* some of the noise in the dataset. In contrast, the slight *negative* bias of MICr-Geom and MICr-Lap on datasets with the largest MICe values (in both collections, but particularly for Baseball) is likely due to underlying approximation differences between (non-private) MICe and MICr.

Also, while some outlier datasets with magnitudes of bias as large as 0.5 exist in the Baseball collection for both MICr-Lap and MICr-Geom, they are most prevalent for datasets with the smallest and largest MICe scores (left and right-most bins). This means it is less likely that the output of either mechanism on one of these datasets would suggest a highly-correlated pair (say, output above 0.5) when the true MICe score is low (or vice versa). Moreover, for the Spellman4381 collection, these slight positive and negative biases are less apparent, and the min and max biases of -0.04 and -0.002 for MICr-Lap and -0.04 and 0.06 for MICr-Geom across *all* bins indicates that for n as small as 4381, the error of these two mechanisms is sufficiently small to be useable in practice.

6. Related Work

MIC is one of several recently proposed statistics for detecting non-linear dependencies between variables in the non-private setting. Other works that propose measures like the Randomized Dependence Coefficient (Lopez-Paz et al., 2013) and the ξ_n coefficient (Chatterjee, 2021) compare results directly with MIC and claim slightly improved computational properties and more intuitive dependence measurements for certain classes of distributions. However, MIC and MICe have been more thoroughly investigated (Reshef et al., 2016; 2018; 2020) and analyzed (Kinney & Atwal, 2014; Reshef et al., 2014) and has found wide use in practice, particularly in bioinformatics settings (Albanese et al., 2018; Cao et al., 2021).

While no private variants of competing measures of dependence have been introduced, much work has been devoted to differentially private estimation of other statistical properties, for example mean estimation (Kamath et al., 2020), hypothesis selection (Bun et al., 2019), quantiles (Gillenwater et al., 2021), and false discovery rate control (Zhang et al., 2021).

We contribute to both lines of work by introducing low-error, private variants for measuring non-linear dependencies.

7. Discussion and Future Work

In practice, publishing MIC values privately requires a broad consideration of the analysis process. If multiple statistics are published, then each must be allocated some fraction of an overall privacy budget. Moreover, choosing which statistics to publish based on the data may leak information, and thus the selection process itself should be differentially private. To optimize accuracy, we recommend that the parameters B and c be optimized for the size n of the dataset of interest and the desired privacy level ϵ, though linear interpolation across precomputed values can be used for speed. The privacy semantics of the results depend on what each data point represents. For example, we obtain user-level privacy if distinct individuals contribute each point and event-level privacy if points represent different events.

There are several promising directions to improve and apply this work. We have not yet considered how to use the local sensitivity (Nissim et al., 2007) of a given dataset to improve the accuracy of a DP MIC statistic. Also, it may be possible to remove the requirement for MICr that the range of the data be known in advance by estimating it in a differentially private way. Moreover, we note that the noisy master grid of MICr-Geom is already differentially private, and so more-accurate methods to estimate MIC from this object seem possible, potentially by exploiting the regular pattern of bias we observed experimentally (where bias decreases with non-private MICe).

Acknowledgments

This work was supported by the Office of Naval Research. The authors would like to thank Joan Feigenbaum for advising and supporting the early stages of this work, Argyris Oikonomou and Grigoris Velegkas for their contributions in an initial analysis of the sensitivity of MIC, and Yakir Reshef for helpful feedback on the consistency of MICe.

References

Albanese, D., Riccadonna, S., Donati, C., and Franceschi, P. A practical tool for maximal information coefficient analysis. *Gigascience*, 2018.

Bun, M., Kamath, G., Steinke, T., and Wu, Z. S. Private hypothesis selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Cao, D., Chen, Y., Chen, J., Zhang, H., and Yuan, Z. An improved algorithm for the maximal information coefficient and its application. *Royal Society open science*, 2021.

Chatterjee, S. A new coefficient of correlation. *Journal of the American Statistical Association*, 2021.

Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., and Wang, T. Privacy at scale: Local differential privacy in practice. In *Proceedings of the International Conference on Management of Data*, 2018.

Dwork, C. Differential privacy and the US census. In *Proceedings of the Symposium on Principles of Database Systems (PODS)*, 2019.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference (TCC)*, 2006.

Ghosh, A., Roughgarden, T., and Sundararajan, M. Universally utility-maximizing privacy mechanisms. *SIAM J. Comput.*, 2012.

Gillenwater, J., Joseph, M., and Kulesza, A. Differentially private quantiles. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., and Craig, D. W. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 2008.

Kamath, G., Singhal, V., and Ullman, J. Private mean estimation of heavy-tailed distributions. In *Proceedings of the Conference on Learning Theory (COLT)*, 2020.

Kinney, J. B. and Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *Proceedings of the National Academy of Sciences (PNAS)*, 2014.

Lazarsfeld, J. and Johnson, A. Consistency of the maximal information coefficient estimator. *CoRR*, abs/2107.03836, 2021.

Lopez-Paz, D., Hennig, P., and Schölkopf, B. The randomized dependence coefficient. *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the ACM Symposium on Theory of computing (STOC)*, 2007.

Prospectus, B. Baseball prospectus statistics reports (2009), Accessed March 2020. URL www.baseballprospectus.com/sortable/.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. Detecting novel associations in large data sets. *Science*, 2011.

Reshef, D. N., Reshef, Y. A., Mitzenmacher, M., and Sabeti, P. C. Cleaning up the record on the maximal information coefficient and equitability. *Proceedings of the National Academy of Sciences (PNAS)*, 2014.

Reshef, D. N., Reshef, Y. A., Sabeti, P. C., and Mitzenmacher, M. An empirical study of the maximal and total information coefficients and leading measures of dependence. *The Annals of Applied Statistics*, 2018.

Reshef, Y. A., Reshef, D. N., Finucane, H. K., Sabeti, P. C., and Mitzenmacher, M. Measuring dependence powerfully and equitably. *The Journal of Machine Learning Research (JMLR)*, 2016.

Reshef, Y. A., Reshef, D. N., Sabeti, P. C., and Mitzenmacher, M. Equitability, interval estimation, and statistical power. *Statistical Science*, 2020.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, 1998.

Wang, R., Li, Y. F., Wang, X., Tang, H., and Zhou, X. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2009.

Zhang, W., Kamath, G., and Cummings, R. Paprika: Private online false discovery rate control. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Appendices

A. MICe and MICe-Lap Details

This section provides more details on the MICe and MICe-Lap statistics and develops the proofs of Theorem 3.1 (MICe sensitivity) and Theorem 3.2 (MICe-Lap privacy).

A.1. Computing MICe

We first give more details on computing MICe (which also provides the foundation for computing MICr).

For any $c > 0$ and $B := B(n)$, MICe can be computed efficiently using the OPTIMIZEAXIS routine of Reshef et al. (2011): for a fixed dataset D of size n and a fixed (wlog) $k \leq \ell$, OPTIMIZEAXIS takes as input a fixed column partition Γ_x of size ℓ, a fixed master row partition Γ_y of size $\hat{k}$, and simultaneously outputs the size-k subgrid $P_k \subseteq \Gamma_y$ (where $k \leq \hat{k}$) that maximizes $I^\star(D|_{(P_k,\Gamma_x)})$ for all $k \leq \min\{\ell, B(n)/\ell\}$ in $O((\hat{k})^2 k\ell)$ time.

We refer the reader to Appendix 3 of Reshef et al. (2011) for exact implementation details of OPTIMIZEAXIS, which uses a dynamic-programming-based approach to perform the maximization. Using this subroutine, the runtime of computing MICe from Definition 2.3 can be stated as follows:

Theorem A.1 (Reshef et al. (2016))**.** *For any dataset D of size n, $B := B(n)$, and $c > 0$,* MICe(D, B, c) *can be computed using* OPTIMIZEAXIS *in $O(c^2 B^4)$ time.*

Proof. As stated in Definition 2.3, for a fixed ℓ, for every $k \leq \ell$, the (k, ℓ)'th entry of the equicharacteristic matrix uses a master row mass equipartition of size $\hat{k} = c \cdot \ell$, for some fixed $c > 0$. Then for each $\ell = 2, \ldots, B/2$, the OPTIMIZEAXIS routine runs in $O(c^2\ell^2 \cdot k\ell) = O(c^2\ell^2 B)$ time, since $k\ell \leq B$. In sum, running this routine for each $\ell = 2, \ldots B/2$ (to compute all equicharacteristic matrix entries where $\ell \geq k$) takes $O(c^2 B^4)$ time. By symmetry, computing all equicharacteristic matrix entries where $k > \ell$ also takes $O(c^2 B^4)$ time, and thus the theorem statement follows. □

A.2. MICe Sensitivity Upper Bound

We now develop the proof of Theorem 3.1, which bounds the ℓ_1-sensitivity of the MICe statistic.

First, recall from Section 2 that for a fixed $k \leq \ell$ and for a fixed D of n points, $\mathcal{E}(D, c, k, [\ell])$ is the set of all grids G with k rows and ℓ columns where the corresponding distribution $D|_G$ has mass-equipartitioned columns. Equivalently, for $G \in \mathcal{E}(D, c, k, [\ell])$, the count matrix $\mathbf{A}_{D,G}$ is of size $k \times \ell$, and its column sums are all equal[2]. For $k > \ell$, the set $\mathcal{E}(D, c, [k], \ell)$ is analogoulsy defined as the set of all $k \times \ell$ grids G where $D|_G$ has equipartitioned rows.

Recall also from Section 2 that for a dataset D of n points, a maximum grid size parameter $B := B(n)$, and a constant $c > 0$, we define

$$\text{MICe}(D, B, c) \ = \ \max_{k\cdot\ell\leq B} \left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell} , \tag{1}$$

where $\mathbf{M}_D^{\mathcal{E}(D,c)}$ is the equicharacteristic matrix of D.

Then for any (k, ℓ) pair, the (k, ℓ) entry of $\mathbf{M}_D^{\mathcal{E}(D,c)}$ is computed by

$$\left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell} = \begin{cases} \displaystyle\max_{G\in\mathcal{E}(D,c,k,[\ell])} \frac{I(D|_G)}{\log_2 k} & \text{if } k \leq \ell \\ \displaystyle\max_{G\in\mathcal{E}(D,c,[k],\ell)} \frac{I(D|_G)}{\log_2 \ell} & \text{if } k > \ell \end{cases} . \tag{2}$$

Fix $k \leq \ell$ (which we assume without loss of generality, since the analogous arguments when $k > \ell$ follow by symmetry).

[2] In the case that n/ℓ is not whole, let the first $\ell - 1$ columns of $\mathbf{A}_{D,G}$ contain $\lceil(n/\ell)\rceil$ points each, and let ℓ'th column contain the remainder of the points. Additionally, we assume that the coordinates of every point in D are unique, since we can perturb the position of each point by a small amount without affecting the output of the MICe statistic.

Because the set $\mathcal{E}(D, c, k, [\ell])$ depends on D, it follows that $\mathcal{E}(D, c, k, [\ell]) \neq \mathcal{E}(D', c, k, [\ell])$ for datasets $D \neq D'$. This means that in (2), the values $(\mathbf{M}_D^{\mathcal{E}(D,c)})_{k,\ell}$ and $(\mathbf{M}_{D'}^{\mathcal{E}(D',c)})_{k,\ell}$ are defined as maximizations over nonequal constraint sets.

To derive an upper bound on the sensitivity of MICe$(\cdot, B, c)$, it is helpful to define these two maximizations using a *common* constraint set — one that depends on n, but not on D or D'. To this end, we will proceed by developing a new, equivalent formulation for entries in $\mathbf{M}_D^{\mathcal{E}(D,c)}$, where the constraint set depends only on n.

A.2.1. Alternate Formulation of Equicharacteristic Matrix

For a $k \times \ell$ grid G and a fixed dataset D of size n, recall that $D|_G$ is a joint probability distribution over the discrete space $\{1, \ldots, k\} \times \{1, \ldots, \ell\}$. So the distribution can be represented by the count matrix $\mathbf{P}_{D,G} \in \mathbb{R}^{k \times \ell}$, where

$$\mathbf{P}_{D,G} \;=\; \frac{1}{n} \cdot \mathbf{A}_{D,G} \;.$$

Since by definition $\sum_{i,j} (\mathbf{A}_{D,G})_{i,j} = n$, we must have $\sum_{i,j} (\mathbf{P}_{D,G})_{i,j} = 1$. For readability, when D and G are fixed and clear from context, we will drop the subscripts and write $\mathbf{A}$ and $\mathbf{P}$.

The row and column marginal distributions of $D|_G$ can also be represented in vector form. First, letting $(\mathbf{a}_{D,G})_{i,*}$ and $(\mathbf{a}_{D,G})_{*,j}$ denote the i'th row vector and j'th column of $\mathbf{A}_{D,G}$ respectively, we define $\mathbf{r}_{D,G} \in \mathbb{Z}^k$ and $\mathbf{c}_{D,G} \in \mathbb{Z}^\ell$ as

$$\begin{aligned} \mathbf{r}_{D,G} &= (\|(\mathbf{a}_{D,G})_{1,*}\|_1, \ldots, \|(\mathbf{a}_{D,G})_{k,*}\|_1) \\ \mathbf{c}_{D,G} &= (\|(\mathbf{a}_{D,G})_{*,1}\|_1, \ldots, \|(\mathbf{a}_{D,G})_{*,\ell}\|_1) \;, \end{aligned}$$

where $\|\mathbf{r}_{D,G}\|_1 = \|\mathbf{c}_{D,G}\|_1 = n$. Then the row and column marginal distributions of $D|_G$ are given by $(\mathbf{r}_{D,G}/n)$ and $(\mathbf{c}_{D,G}/n)$ respectively.

Now observe that by definition, the mutual information $I(D|_G)$ depends only on the joint distribution $D|_G$, and not on the individual coordinates of points in D. Additionally, because we can represent $D|_G$ using the matrix $\mathbf{P} = \frac{1}{n} \cdot \mathbf{A}$, the joint distribution $D|_G$ depends only on the entries of $\mathbf{A}$, and not on the absolute positions of the column and row dividers of G.

It follows trivially that for a fixed dataset D of size n and two different $k \times \ell$ grids G and G', if $\mathbf{A}_{D,G} = \mathbf{A}_{D,G'}$, then $\mathbf{r}_{D,G} = \mathbf{r'}_{D,G}$, $\mathbf{c}_{D,G} = \mathbf{c'}_{D,G}$, and $\mathbf{P}_{D,G} = \mathbf{P}_{D,G'}$. Moreover, for a fixed D, if $\mathbf{r}_{D,G} = \mathbf{r}_{D,G'}$ and $\mathbf{c}_{D,G} = \mathbf{c}_{D,G'}$ then $\mathbf{A}_{D,G} = \mathbf{A}_{D,G'}$ and $\mathbf{P}_{D,G} = \mathbf{P}_{D,G'}$ (i.e., if $D|_G$ and $D|_{G'}$ have the same marginal row and column distributions, then $D|_G = D|_{G'}$).

Together, these two observations imply that for a fixed D and any two grids G and G', the two distributions $D|_G$ and $D|_{G'}$ are equal if and only if $\mathbf{r}_{D,G} = \mathbf{r'}_{D,G}$ and $\mathbf{c}_{D,G} = \mathbf{c'}_{D,G}$. We summarize these observations and equivalences with the following lemma.

Lemma A.2. *Fix a dataset D of n points, and fix any k and ℓ. For any grids G and G' of size $k \times \ell$, the following statements are all equivalent:*

$$D|_G \;=\; D|_{G'} \tag{i}$$

$$\mathbf{A}_{D,G} \;=\; \mathbf{A}_{D,G'} \tag{ii}$$

$$\mathbf{P}_{D,G} \;=\; \mathbf{P}_{D,G'} \tag{iii}$$

$$\mathbf{r}_{D,G} = \mathbf{r}_{D,G'} \;\; \textit{and} \;\; \mathbf{c}_{D,G} = \mathbf{c}_{D,G'} \;. \tag{iv}$$

Proof. Statements (i) and (ii) are equivalent by the definition of a count matrix $\mathbf{A}$, and (iii) is equivalent to (ii) by the definition of the $\mathbf{P}$. Statement (ii) also implies (iv) by the definition of the vectors $\mathbf{r}$ and $\mathbf{c}$.

To complete the proof, we show that (iv) implies (ii). For this, notice that for a fixed D, the entries of $\mathbf{A}_{D,G}$ can be computed given the vectors $\mathbf{r}_{D,G}$ and $\mathbf{c}_{D,G}$. For example, the entry $(\mathbf{A}_{D,G})_{1,1}$ is the number of points in D whose y-coordinate has descending rank at most $\mathbf{r}_{D,G}(1)$, and whose x-coordinate has ascending rank at most $\mathbf{c}_{D,G}(1)$; the entry $(\mathbf{A}_{D,G})_{1,2}$ is the number of points in D whose y coordinate has descending rank at least $\mathbf{r}_{D,G}(1) + 1$ and at most $\mathbf{r}_{D,G}(1) + 1 + \mathbf{r}_{D,G}(2)$, and whose x coordinate has ascending rank at most $\mathbf{c}_{D,G}(1)$, etc.

Thus for a fixed D, the entries of $\mathbf{A}_{D,G}$ can be computed as a function of D and the vectors $\mathbf{r}_{D,G}$ and $\mathbf{c}_{D,G}$, but without a direct dependence on the positions of the column and row dividers of G. It then follows that for a fixed D and any two grids G and G' of the same size, if $\mathbf{r}_{D,G} = \mathbf{r}_{D,G'}$ and $\mathbf{c}_{D,G} = \mathbf{c}_{D,G'}$, then $\mathbf{A}_{D,G} = \mathbf{A}_{D,G'}$, which completes the proof. □

Observe that for a fixed n and k and ℓ, for every dataset D of size n, and for every $G \in \mathcal{E}(D, c, k, [\ell])$, the vectors $\mathbf{c}_{D,G} \in \mathbb{Z}^\ell$ are all identical. This follows from the fact that for any D of size n, the count matrices $\mathbf{A}_{D,G}$ for any $G \in \mathcal{E}(D, c, k, [\ell])$ have row sums that depend only on n and ℓ, and not on D. Let $\mathbf{c}^{n,[\ell]} \in \mathbb{Z}^\ell$ denote this unique vector.

Additionally, observe that there are only finitely many vectors $\mathbf{r} \in \mathbb{Z}^k$ satisfying $\|\mathbf{r}\|_1 = n$. Let $\mathcal{R}(n, k)$ denote the set of all such vectors where $T(n, k) = |\mathcal{R}(n, k)|$, and fix some enumeration $\mathcal{R}(n, k) = \{\mathbf{r}^1, \mathbf{r}^2, \ldots, \mathbf{r}^{T(n,k)}\}$.

So for a fixed dataset D of size n,

$$\forall\, G \in \mathcal{E}(D, c, k, [\ell]) \quad \left(\exists \text{ unique } i \in [T(n, k)] \text{ such that } \mathbf{r}_{D,G} = \mathbf{r}^i\right) \ , \tag{3}$$

which follows directly from the definition of $\mathbf{r}_{D,G}$. So for any $i \in [T(n, k)]$, for any two $k \times \ell$ grids G and G' such that $\mathbf{r}_{D,G} = \mathbf{r}_{D,G'} = \mathbf{r}^i$, it follows by Lemma A.2 that $D|_G = D|_{G'}$ (since we also have $\mathbf{c}_{D,G} = \mathbf{c}_{D,G'} = \mathbf{c}^{n,[\ell]}$), and thus $I(D|_G) = I(D|_{G'})$.

In the other direction, it is easy to see that

$$\forall\, i \in [T(n, k)] \quad \left(\exists\, G \in \mathcal{E}(D, c, k, [\ell]) \text{ such that } \mathbf{r}_{D,G} = \mathbf{r}^i\right) \ , \tag{4}$$

since for any i, such a grid G can be constructed by sorting the points of D by their y coordinate and drawing a row divider between certain numbers of consecutive points as specified by $\mathbf{r}^i$.

For a fixed dataset D of size n, and for any $\mathbf{r}^i \in \mathcal{R}(n, k)$, let $\mathbf{P}\left(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]}\right) \in \mathbb{R}^{k\times\ell}$ be the probability matrix derived from the vectors $\mathbf{r}^i$ and $\mathbf{c}^{n,[\ell]}$ (i.e., as described in the proof of Lemma A.2). Then by (4), for every $i \in [T(n, k)]$, there exists some $G \in \mathcal{E}(D, k, [\ell])$ such that $\mathbf{P}(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]}) = \mathbf{P}_{D,G}$ and thus $I(\mathbf{P}(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]})) = I(D|_G)$.

Finally, by combining the preceding arguments, we state the following reformulation of entries $(\mathbf{M}_D^{\mathcal{E}(D,c)})_{k,\ell}$.

Lemma A.3. *For any fixed dataset D of size n, and for any $k \leq \ell$:*

$$\left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell} \;=\; \max_{G\in\mathcal{E}(D,c,k,[\ell])} \frac{I(D|_G)}{\log_2 k} \;=\; \max_{i\in T(n,k)} \frac{I\left(\mathbf{P}\left(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]}\right)\right)}{\log_2 k} \ . \tag{5}$$

Crucially, for any dataset D of size n and a fixed $k \leq \ell$, the constraint set in the new formulation of Lemma A.3 depends only on n and k, and *not* on D. This yields the following useful observation, which we state as a Corollary:

Corollary A.4. *For any n, for any $2 \leq k \leq \ell$, and for any two datasets D, D', both of size n:*

$$\left| \left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D'}^{\mathcal{E}(D',c)}\right)_{k,\ell} \right| \;\leq\; \max_{i\in T(n,k)} \left| I\left(\mathbf{P}(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right) \,-\, I\left(\mathbf{P}(D', \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right) \right| \ . \tag{6}$$

Proof. By Lemma A.3, we have

$$\begin{aligned}
\left| \left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D'}^{\mathcal{E}(D',c)}\right)_{k,\ell} \right| &= \left| \max_{i\in T(n,k)} \frac{I\left(\mathbf{P}(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right)}{\log_2 k} \,-\, \max_{i\in T(n,k)} \frac{I\left(\mathbf{P}(D', \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right)}{\log_2 k} \right| \\
&\leq \left| \max_{i\in T(n,k)} I\left(\mathbf{P}(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right) \,-\, \max_{i\in T(n,k)} I\left(\mathbf{P}(D', \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right) \right| ,
\end{aligned}$$

where the inequality follows from factoring out the common $\log_2 k$ term and noting that $\log_2 k \geq \log_2 2 = 1$ by assumption.

Given that the constraint sets in each maximization term are equal, we then have

$$\begin{aligned}
&\left| \max_{i\in T(n,k)} I\left(\mathbf{P}(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right) - \max_{i\in T(n,k)} I\left(\mathbf{P}(D', \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right) \right| \\
&\quad\leq \max_{i\in T(n,k)} \left| I\left(\mathbf{P}(D, \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right) \,-\, I\left(\mathbf{P}(D', \mathbf{r}^i, \mathbf{c}^{n,[\ell]})\right) \right|
\end{aligned}$$

which completes the proof. $\square$

A.2.2. Sensitivity of MICe via Sensitivity of Mutual Information

Using the new formulation from Lemma A.3 we will derive an upper bound on the sensitivity of MICe by deriving an upper bound on the sensitivity of mutual information with respect to a fixed $i \in T(n,k)$.

For this, note that for any pair of neighboring datasets $D \sim D'$ of size n, and for any $B := B(n)$ and $c > 0$:

$$\begin{aligned} |\ \mathrm{MICe}(D,B,c) - \mathrm{MICe}(D',B,c)\ | \ &= \left| \max_{k,\ell\,:\,k\cdot\ell\leq B} \left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell} - \max_{k,\ell\,:\,k\cdot\ell\leq B} \left(\mathbf{M}_{D'}^{\mathcal{E}(D',c)}\right)_{k,\ell} \right| \\ &\leq \max_{k,\ell\,:\,k\cdot\ell\leq B} \left| \left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D'}^{\mathcal{E}(D',c)}\right)_{k,\ell} \right| . \end{aligned} \tag{7}$$

Now by Corollary A.4, we know (assuming wlog that $k \leq \ell$) that

$$\left| \left(\mathbf{M}_D^{\mathcal{E}(D,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D'}^{\mathcal{E}(D',c)}\right)_{k,\ell} \right| \ \leq \max_{i\in T(n,k)} \left| I\left(\mathbf{P}(D,\mathbf{r}^i,\mathbf{c}^{n,[\ell]})\right) \ - \ I\left(\mathbf{P}(D',\mathbf{r}^i,\mathbf{c}^{n,[\ell]})\right) \right| . \tag{8}$$

We further claim the following bound:

Lemma A.5. *For any $n \geq 6$ and any $2 \leq k \leq \ell$, and any neighboring datasets $D \sim D'$ both of size n:*

$$\max_{i\in T(n,k)} \left| I\left(\mathbf{P}(D,\mathbf{r}^i,\mathbf{c}^{n,[\ell]})\right) \ - \ I\left(\mathbf{P}(D',\mathbf{r}^i,\mathbf{c}^{n,[\ell]})\right) \right| \ \leq \ k\ell \cdot \left(\frac{2\log_2 n}{n} + \frac{4.8}{n}\right) .$$

Let us grant Lemma A.5 as true for now. The proof of Theorem 3.1 then follows easily:

Proof (of Theorem 3.1). Applying Lemma A.5 to (8) and substituting into (7) gives us

$$|\ \mathrm{MICe}(D,B,c) - \mathrm{MICe}(D',B,c)\ | \ \leq \max_{k,\ell\,:\,k\cdot\ell\leq B} \ k\ell \cdot \left(\frac{2\log_2 n}{n} + \frac{4.8}{n}\right) , \tag{9}$$

which holds for any $D \sim D'$ of size n. Observing that $k \cdot \ell \leq B$ completes the proof. □

It remains to prove Lemma A.5, which we proceed to do in the following subsection.

A.2.3. Proof of Lemma A.5

Fix any $n \geq 6$ and any $2 \leq k \leq \ell$, and consider any fixed pair of neighboring datasets $D \sim D'$, both of size n.

Fix any $\mathbf{r}^i \in \mathcal{R}(n,k)$, and let

$$\begin{aligned} \mathbf{P} \ &:= \ \mathbf{P}(D,\mathbf{r}^i,\mathbf{c}^{n,[\ell]}) \quad \text{with entries } p(i,j) \text{ and column sums } p(*,j) \\ \mathbf{P}' \ &:= \ \mathbf{P}(D',\mathbf{r}^i,\mathbf{c}^{n,[\ell]}) \quad \text{with entries } p'(i,j) \text{ and column sums } p'(*,j) . \end{aligned}$$

Let $\mathbf{A}$ and $\mathbf{A}'$ denote the corresponding count matrices for $\mathbf{P}$ and $\mathbf{P}'$, respectively. Because $\mathbf{P}$ and $\mathbf{P}'$ are non-negative matrices representing two-dimensional, discrete joint distributions, we define information-theoretic properties of $\mathbf{P}$ and $\mathbf{P}'$ in the natural way. Specifically, let

$$\begin{aligned} I(\mathbf{P}) \ &= \ H_x(\mathbf{P}) - H_{x|y}(\mathbf{P}) \\ H_x(\mathbf{P}) \ &= \ \sum_{j\in[\ell]} p(*,j)\log_2\left(\frac{1}{p(*,j)}\right) \\ H_{x|y}(\mathbf{P}) \ &= \ \sum_{i\in[k]}\sum_{j\in[\ell]} p(i,j)\log_2\left(\frac{p(*,j)}{p(i,j)}\right) . \end{aligned}$$

The analogous definitions for $\mathbf{P}'$ are defined in terms of $p'(i,j)$ and $p'(*,j)$.

Our goal is to give an upper bound on $|I(\mathbf{P}) - I(\mathbf{P}')|$, which we can write as

$$\begin{aligned} |I(\mathbf{P}) - I(\mathbf{P}')| \; &= \; \left| \; H_x(\mathbf{P}) - H_{x|y}(\mathbf{P}) \; - \; \left(H_x(\mathbf{P}') - H_{x|y}(\mathbf{P}')\right) \; \right| \\ &= \; \left| \; H_x(\mathbf{P}) - H_x(\mathbf{P}') \; + \; H_{x|y}(\mathbf{P}') - H_{x|y}(\mathbf{P}) \; \right| \; . \end{aligned}$$

Because $\mathbf{P}$ and $\mathbf{P}'$ are both generated using $\mathbf{c}^{n,[\ell]}$, it follows by definition that $p(*,j) = p'(*,j)$ for all $j \in [\ell]$, and thus $H_x(\mathbf{P}) = H_x(\mathbf{P}')$. Then

$$\begin{aligned} |I(\mathbf{P}) - I(\mathbf{P}')| \; &= \; \left| \; H_{x|y}(\mathbf{P}') - H_{x|y}(\mathbf{P}) \; \right| \\ &= \; \left| \sum_{i\in[k]} \sum_{j\in[\ell]} p'(i,j) \log_2 \left(\frac{p'(*,j)}{p'(i,j)}\right) \; - \; \sum_{i\in[k]} \sum_{j\in[\ell]} p(i,j) \log_2 \left(\frac{p(*,j)}{p(i,j)}\right) \right| \\ &\leq \; \sum_{i\in[k]} \sum_{j\in[\ell]} \left| p'(i,j) \log_2 \left(\frac{p'(*,j)}{p'(i,j)}\right) \; - \; p(i,j) \log_2 \left(\frac{p(*,j)}{p(i,j)}\right) \right| \\ &= \; \sum_{i\in[k]} \sum_{j\in[\ell]} \; | \; p'(i,j) \log_2 p'(*,j) - p'(i,j) \log_2 p'(i,j) - p(i,j) \log_2 p(*,j) + p(i,j) \log_2 p(i,j) \; | \\ &= \; \sum_{i\in[k]} \sum_{j\in[\ell]} \; | \; (p'(i,j) - p(i,j)) \log_2 p(*,j) \; + \; p(i,j) \log_2 p(i,j) - p'(i,j) \log_2 p'(i,j) \; | \; . \end{aligned} \tag{10}$$

Here, the third line is due to the triangle inequality, and the final line is due to $p(*,j) = p'(*,j)$ for all j.

For any $(i,j) \in [k] \times [\ell]$, let

$$\alpha(i,j) \; = \; | \; (p'(i,j) - p(i,j)) \log_2 p(*,j) \; + \; p(i,j) \log_2 p(i,j) - p'(i,j) \log_2 p'(i,j) \; | \; .$$

The following lemma gives a uniform upper bound on $\alpha(i,j)$ for any (i,j):

Lemma A.6. *For any $n \geq 6$, any $2 \leq k \leq \ell$, and any $\mathbf{P}$ and $\mathbf{P}'$ as described above,*

$$\alpha(i,j) \; \leq \; \frac{2\log_2 n}{n} + \frac{4.8}{n}$$

for any $(i,j) \in [k] \times [\ell]$.

Granting this lemma true for now and applying it to expression (10) then implies

$$| \; I(\mathbf{P}) - I(\mathbf{P}') \; | \; \leq \; \sum_{i\in[k]} \sum_{j\in[\ell]} \frac{2\log_2 n}{n} + \frac{4.8}{n} \; \leq \; k\ell \cdot \left(\frac{2\log_2 n}{n} + \frac{4.8}{n}\right) \; .$$

Because we have considered an arbitrary vector $\mathbf{r}^i \in \mathcal{R}(n,k)$, the statement of Lemma A.5 follows.

Thus it only remains to prove Lemma A.6.

Proof (of Lemma A.6). For a fixed grid $G \in \mathcal{E}(D,c,k,[\ell])$, let ϕ_x and ϕ'_x denote the x-coordinate point mapping functions wrt G for D and D' respectively (i.e., for $d \in D$, the function $\phi_x(d)$ returns the column partition index of d wrt $D|_G$). Then observe that for any $(i,j) \in [k] \times [\ell]$, we have

$$| \; p(i,j) - p'(i,j) \; | \leq 2/n \; .$$

This is because for every column $j \in [\ell]$, there is at most one point $d \in D$ where $\phi_x(d) = j$ but $\phi'_x(d) \in \{j-1, j+1\}$ (equivalently, at most one point $d' \in D'$ where $\phi'_x(d') \in \{j-1, j+1\}$ but $\phi_x(d) = j$), The same argument holds for every row $i \in [k]$. Thus $\mathbf{A}$ and $\mathbf{A}'$ differ by at most 2 at any entry (i,j), and so $|p(i,j) - p'(i,j)| \leq 2/n$ for all (i,j).

Using this fact, we will derive an upper bound on $\alpha(i,j)$ for any fixed (i,j) via three cases: (1) when $p(i,j) = p'(i,j)$, (2) when $p(i,j) > p'(i,j)$, and (3) when $p(i,j) < p'(i,j)$. For ease of readability, and when (i,j) is fixed and clear from context, we we will write $p := p(i,j)$ and $p' := p'(i,j)$. Thus for fixed (i,j) we seek to bound

$$\alpha \; = \; | \; (p' - p) \log_2 p(*,j) + p \log_2 p - p' \log_2 p' \; | \; .$$

We have the following cases:

1. *Case when $p = p'$:*

   We trivially have that $\alpha = 0$.

2. *Case when $p > p'$:*

   (a) *Case when $p - p' = 1/n$:*

   Note that $p - p' = 1/n$ implies $p(*, j) \geq p \geq 1/n$. If $p = 1/n$ (and $p' = 0$), then

$$\begin{aligned} \alpha \;=\; \left| -\frac{1}{n}\log_2 p(*, j) + \frac{1}{n}\log_2 \frac{1}{n} \right| \;&=\; \left| \frac{1}{n}\log_2 \frac{1/n}{p(*, j)} \right| \\ &=\; \left| \frac{1}{n}\log_2 \frac{p(*, j)}{1/n} \right| \;. \end{aligned}$$

   Since $1/n \leq p(*, j) \leq 1$, it follows that $\alpha \leq (\log_2 n)/n$.

   On the other hand, say $p \geq 2/n$. Then

$$\begin{aligned} \alpha \;&=\; \left| -\frac{1}{n}\log_2 p(*, j) + p\log_2 p - (p - 1/n)\log_2(p - 1/n) \right| \\ &=\; \left| p\log_2 \frac{p}{p - 1/n} + \frac{1}{n}\log_2 \frac{p - 1/n}{p(*, j)} \right| \\ &\leq\; \left| p\log_2 \frac{p}{p - 1/n} \right| + \left| \frac{1}{n}\log_2 \frac{p - 1/n}{p(*, j)} \right| \;, \end{aligned} \tag{11}$$

   where the last line is due to the triangle inequality. Observe that the left hand term in (11) is decreasing in p, and since $p \geq 2/n$ by assumption:

$$\left| p\log_2 \frac{p}{p - 1/n} \right| \;\leq\; \left| \frac{2}{n}\log_2 \frac{2/n}{2/n - 1/n} \right| \;=\; \frac{2}{n} \tag{12}$$

   For the right hand term in (11), note that $p - 1/n = p' \leq p(*, j)$ (since the joint mass in any cell (i, j) is at most the marginal mass of that cell's column), so this term is also decreasing in p. Again since $p \geq 2/n$ by assumption, we have

$$\left| \frac{1}{n}\log_2 \frac{p - 1/n}{p(*, j)} \right| \;\leq\; \left| \frac{1}{n}\log_2 \frac{2/n - 1/n}{p(*, j)} \right| \;=\; \left| \frac{1}{n}\log_2 \left(n \cdot p(*, j)\right) \right| \;\leq\; \frac{\log_2 n}{n} \;, \tag{13}$$

   where the final inequality holds because $2/n \leq p(*, j) \leq 1$.

   Substituting (12) and (13) back into (11) gives $\alpha \leq (2/n) + (\log_2 n)/n$ when $p \geq 2/n$. Thus this bound holds for Case (2a) in general.

   (b) *Case when $p - p' = 2/n$:*

   The condition $p - p' = 2/n$ implies $p(*, j) \geq p \geq 2/n$. First consider when $p = 2/n$ and $p' = 0$. Then

$$\begin{aligned} \alpha \;=\; \left| -\frac{2}{n}\log_2 p(*, j) + \frac{2}{n}\log_2 \frac{2}{n} \right| \;&=\; \left| \frac{2}{n}\log_2 \frac{p(*, j)}{2/n} \right| \\ &=\; \left| \frac{2}{n}\log_2 \frac{n}{2} \right| \;\leq\; \frac{\log_2 n}{n} \;, \end{aligned}$$

   where the penultimate inequality follows from the assumption that $2/n \leq p(*, j) \leq 1$.

   On the other hand, consider when $p \geq 3/n$. Then

$$\begin{aligned} \alpha \;&=\; \left| -\frac{2}{n}\log_2 p(*, j) + p\log_2 p - (p - 2/n)\log_2(p - 2/n) \right| \\ &=\; \left| p\log_2 \frac{p}{p - 2/n} + \frac{2}{n}\log_2 \frac{p - 2/n}{p(*, j)} \right| \\ &\leq\; \left| p\log_2 \frac{p}{p - 2/n} \right| + \left| \frac{2}{n}\log_2 \frac{p - 2/n}{p(*, j)} \right| \;. \end{aligned} \tag{14}$$

As in Case (2a), the left hand term in (14) is decreasing in $p \geq 3/n$, meaning

$$\left| p\log_2 \frac{p}{p-2/n} \right| \;\leq\; \left| \frac{3}{n}\log_2 \frac{3/n}{3/n-2/n} \right| \;=\; \frac{3}{n}\log_2 3 \;\leq\; \frac{4.8}{n}. \tag{15}$$

Since $p \leq p(*, j)$, the right hand term in (14) is also decreasing in $p \geq 3/n$. So

$$\left| \frac{2}{n}\log_2 \frac{p-2/n}{p(*,j)} \right| \;\leq\; \left| \frac{2}{n}\log_2 \frac{3/n-2/n}{p(*,j)} \right| \;\leq\; \left| \frac{2}{n}\log_2 (n\cdot p(*,j)) \right| \;\leq\; \frac{2\log_2 n}{n}, \tag{16}$$

where the finaly inequality holds since $p(*, j) \leq 1$.

Substituting (15) and (16) back into (14) gives $\alpha \leq (4.8/n) + (2\log_2 n)/n$, which holds in general for Case (2b).

3. *Case when* $p' > p$:

   (a) *Case when* $p' - p = 1/n$:

   This condition implies $0 \leq p \leq 1 - 1/n$. First consider when $p = 0$ and $p' = 1/n$. Then

$$\alpha \;=\; \left| \frac{1}{n}\log_2 p(*,j) - \frac{1}{n}\log_2 \frac{1}{n} \right| \;=\; \left| \frac{1}{n}\log_2 \frac{p(*,j)}{1/n} \right| \;\leq\; \frac{\log_2 n}{n}\ ,$$

   where the final inequality is due to $1/n = p' \leq p(*, j) \leq 1$.

   Now consider when $p \geq 1/n$. Then

$$\begin{aligned}\alpha \;&=\; \left| \frac{1}{n}\log_2 p(*,j) + p\log_2 p - (p+1/n)\log_2(p+1/n) \right| \\ &=\; \left| p\log_2 \frac{p}{p+1/n} + \frac{1}{n}\log_2 \frac{p(*,j)}{p+1/n} \right| \\ &\leq\; \left| p\log_2 \frac{p}{p+1/n} \right| + \left| \frac{1}{n}\log_2 \frac{p(*,j)}{p+1/n} \right|\ .\end{aligned} \tag{17}$$

   The left hand term of (17) is increasing with p, and by the assumption that $p \leq 1 - 1/n$,

$$\left| p\log_2 \frac{p}{p+1/n} \right| \;\leq\; \left| \left(\frac{n-1}{n}\right)\log_2 \left(\frac{n-1}{n}\right) \right| \;\leq\; \left| \log_2 \frac{n-1}{n} \right| \;=\; \left| \log_2 \frac{n}{n-1} \right|\ .$$

   Because $(n + 2)/n \geq n/(n - 1)$ for all $n \geq 2$, we have

$$\left| p\log_2 \frac{p}{p+1/n} \right| \;\leq\; \left| \log_2 \frac{n+2}{n} \right| \;=\; |\ \log_2(1+2/n)\ |\ .$$

   Note that because $e \leq 2^{(3/2)}$, then $1 + x \leq e^x \leq 2^{(3/2)x}$ for all x. This implies $\log_2(1 + x) \leq (3/2)x$ for all x, and so

$$\left| p\log_2 \frac{p}{p+1/n} \right| \;\leq\; |\ \log_2(1+2/n)\ | \;\leq\; \frac{3}{2}\cdot\frac{2}{n} \;=\; \frac{3}{n}\ . \tag{18}$$

   The right hand term of (17) is decreasing in p since $p + 1/n = p' \leq p(*, j)$. Since $p \geq 1/n$ by assumption,

$$\begin{aligned}\left| \frac{1}{n}\log_2 \frac{p(*,j)}{p+1/n} \right| \;&\leq\; \left| \frac{1}{n}\log_2 \frac{p(*,j)}{1/n+1/n} \right| \\ &=\; \left| \frac{1}{n}\log_2 ((n/2)\cdot p(*,j)) \right| \;\leq\; \frac{\log_2(n/2)}{n} \;\leq\; \frac{\log_2 n}{n}\ ,\end{aligned} \tag{19}$$

   where the penultimate inequality holds because $p(*, j) \leq 1$.

   Substituting (18) and (19) into (17) gives $\alpha \leq (3/n) + (\log_2 n)/n$, which holds in general for Case (3a).

(b) *Case when* $p - p' = 2/n$:

This condition implies $0 \leq p \leq 1 - 2/n$. When $p = 0$ and $p' = 2/n$, we have

$$\alpha \;=\; \left| \frac{2}{n}\log_2 p(*,j) - \frac{2}{n}\log_2\frac{2}{n} \right| \;=\; \left| \frac{2}{n}\log_2\frac{p(*,j)}{2/n} \right| \;\leq\; \left| \frac{2}{n}\log_2(n/2) \right| \;\leq\; \frac{2\log_2 n}{n} \;,$$

where the penultimate inequality is due to $2/n = p' \leq p(*,j) \leq 1$.

Now consider when $p \geq 1/n$. We have

$$\begin{aligned}\alpha \;&=\; \left| \frac{2}{n}\log_2 p(*,j) + p\log_2 p - (p+2/n)\log_2(p+2/n) \right| \\ &=\; \left| p\log_2\frac{p}{p+2/n} + \frac{2}{n}\log_2\frac{p(*,j)}{p+2/n} \right| \\ &\leq\; \left| p\log_2\frac{p}{p+2/n} \right| + \left| \frac{2}{n}\log_2\frac{p(*,j)}{p+2/n} \right| \;. \end{aligned} \tag{20}$$

Similar to Case (3a), the left hand term of (20) is increasing with p, and by the assumption that $p \leq 1 - 2/n$,

$$\left| p\log_2\frac{p}{p+2/n} \right| \;\leq\; \left| \left(\frac{n-2}{n}\right)\log_2\left(\frac{n-2}{n}\right) \right| \;\leq\; \left| \log_2\frac{n-2}{n} \right| \;=\; \left| \log_2\frac{n}{n-2} \right| \;.$$

Because $(n+3)/n \geq n/(n-2)$ for all $n \geq 6$, we have

$$\left| p\log_2\frac{p}{p+2/n} \right| \;\leq\; \left| \log_2\frac{n+3}{n} \right| \;=\; |\log_2(1+3/n)| \;.$$

Then applying the identity $\log_2(1+x) \leq (3/2)x$ yields

$$\left| p\log_2\frac{p}{p+2/n} \right| \;\leq\; |\log_2(1+3/n)| \;\leq\; \frac{3}{2}\cdot\frac{3}{n} \;=\; \frac{4.5}{n} \;. \tag{21}$$

Now the right hand term of (20) is decreasing in p since $p + 2/n = p' \leq p(*,j)$. So by the assumption $p \geq 1/n$,

$$\begin{aligned}\left| \frac{2}{n}\log_2\frac{p(*,j)}{p+2/n} \right| \;&\leq\; \left| \frac{2}{n}\log_2\frac{p(*,j)}{1/n+2/n} \right| \\ &\leq\; \left| \frac{2}{n}\log_2((n/3)\cdot p(*,j)) \right| \;\leq\; \frac{2\log_2(n/3)}{n} \;\leq\; \frac{2\log_2(n)}{n} \;, \end{aligned} \tag{22}$$

where the penultimate inequality is due to $p(*,j) \leq 1$.

Now substituting expressions (21) and (22) into (20) gives the bound $\alpha \leq (4.5/n) + (2\log_n)/n$, which holds in general for Case (3b).

To summarize the bounds on α for each of the cases:

- *Case when* $p = p'$: $\alpha = 0$.
- *Cases when* $p \neq p'$:

	$\lvert p - p'\rvert = 1/n$	$\lvert p - p'\rvert = 2/n$
$p > p'$	Case (2a): $\alpha \leq (\log_2 n)/n + 2/n$	Case (2b): $\alpha \leq (2\log_2 n)/n + 4.8/n$
$p' > p$	Case (3a): $\alpha \leq (\log_2 n)/n + 3/n$	Case (3b): $\alpha \leq (2\log_2 n)/n + 4.5/n$

It follows that for any (i,j),

$$\alpha \;\leq\; \frac{2\log_2 n}{n} + \frac{4.8}{n} \;,$$

which completes the proof.

□

A.3. MICe-Lap Privacy

Here we prove the privacy guarantee of the MICe-Lap mechanism from Theorem 3.2. This requires first stating the post-processing principle, which says that the result of performaing any additional computation on the output of an ϵ-DP mechanism is still private.

Theorem A.7 (DP post-processing, Dwork & Roth (2014))**.** *Let $A : \mathcal{X} \to \mathcal{Y}$ be an ϵ-DP mechanism. Then for any function $f : \mathcal{Y} \to \mathcal{Z}$, the composition $(f \circ A) : \mathcal{X} \to \mathcal{Z}$ is also ϵ-DP.*

The privacy of the MICe-Lap mechanism then follows directly:

Theorem 3.2. MICe-Lap$(\cdot, B, c, \epsilon)$ *is ϵ-DP.*

Proof. This follows by the ϵ-DP privacy of the Laplace mechanism, and by applying the post-processing property from Theorem A.7 with truncation to the $[0, 1]$ interval. □

B. MICr Details

In this section, we give more details on the MICr statistic and develop the proofs of Theorem 4.2 (MICr consistency) and Theorem 4.3 (MICr sensitivity). Some of the notation and definitions were introduced previously in Sections 2 and 4, but we will restate them here for completeness and readability.

B.1. Computing MICr

B.1.1. The set of grids $\mathcal{R}(L, c, k, \ell)$

Recall for an interval I that $R_{I,\ell}$ is a size-ℓ range-equipartition of I, and $\mathcal{P}(I, k, [c\ell])$ is the set of all size-k subpartitions of a size-$(c\ell)$ range-equipartition of I. Then for fixed range bounds $L = L_x \times L_y$ and any finite $c > 0$:

- for any $2 \leq k \leq \ell$, the set $\mathcal{R}(L, c, k, [\ell])$ contains all grids $G = (P, Q)$ where $Q = R_{L_x,\ell}$ and $P \in \mathcal{P}(L_y, k, [c\ell])$.
- for any $2 \leq \ell < k$, the set $\mathcal{R}(L, c, [k], \ell)$ contains all grids $G = (P, Q)$ where $Q \in \mathcal{P}(L_x, \ell, [ck])$ and $P = R_{L_y,k}$.

For any L and c, and any $k, \ell \geq 2$, we overload the defintion of $\mathcal{R}$ and write

$$\mathcal{R}(L, c, k, \ell) \;=\; \begin{cases} \mathcal{R}(L, c, k, [\ell]) & \text{if } k \leq \ell \\ \mathcal{R}(L, c, [k], \ell) & \text{if } k > \ell \end{cases} . \tag{23}$$

B.1.2. Computing MICr on finite datasets

Recall now that for a fixed L and c, and for any dataset D of size n with range restricted to L, let $\mathbf{M}_D^{\mathcal{R}(L,c)}$ denote the sample range-equicharacteristic matrix for that dataset. For any $k, \ell \geq 2$, its entries are given by

$$\left(\mathbf{M}_D^{\mathcal{R}(L,c)}\right)_{k,\ell} \;=\; \max_{G \in \mathcal{R}(L,c,k,\ell)} \frac{I(D|_G)}{\log_2 \min\{k, \ell\}} . \tag{24}$$

Then for a maximum grid parameter $B := B(n)$, we define

$$\text{MICr}(D, B, L, c) \;=\; \max_{k,\ell \,:\, k \cdot \ell \leq B} \left(\mathbf{M}_D^{\mathcal{R}(L,c)}\right)_{k,\ell} . \tag{25}$$

Note that because MICr and MICe differ only in the *definition* of the respective master grids, but not in their size, MICr can be computed in time asympotically equivalent to MICe again using the OptimizeAxis routine described in Appendix A.1. Stated formally (and analogous to Theorem A.1):

Theorem B.1. *For any dataset D of size n restricted to range L, $B := B(n)$, and $c > 0$,* MICr(D, L, B, c) *can be computed using* OptimizeAxis *in $O(c^2 B^4)$ time.*

B.1.3. COMPUTING MICr ON JOINT DISTRIBUTIONS

Because the set $\mathcal{R}(L, c, k, \ell)$ depends only on L and c, and not on the dataset D, we can define a range-equicharacteristic matrix for joint distributions analogously to (24).

Again for a fixed L and c, and for any joint distribution (X, Y) with range restricted to L, let $\mathbf{M}_{(X,Y)}^{\mathcal{R}(L,c)}$ denote the range-equicharacteristic matrix for that distribution. For any $k, \ell \geq 2$, its entries are given by

$$\left(\mathbf{M}_{(X,Y)}^{\mathcal{R}(L,c)}\right)_{k,\ell} = \max_{G \in \mathcal{R}(L,c,k,\ell)} \frac{I((X,Y)|_G)}{\log_2 \min\{k,\ell\}} . \tag{26}$$

B.1.4. THE UNCONSTRAINED MICR STATISTIC

For theoretical purposes, we also define the following *unconstrained* variant of the MICr statistic with equicharacteristic matrix $\mathbf{M}_D^{\mathcal{R}(L,\infty)}$. Compared to $\mathcal{R}(L, c, k, \ell)$, the set $\mathcal{R}(L, \infty, k, \ell)$ (wlog when $k \leq \ell$) contains grids G with *any* row partition of $[y_0, y_1]$ of size k (i.e., not just those restricted to subpartitions of some finite-size master range-equipartition). Thus $\mathcal{R}(L, \infty, k, \ell)$ is exactly the set of grids G obtained from $\mathcal{R}(L, c, k, \ell)$ when taking $c \to \infty$.

More formally, for an interval I, let $\mathcal{P}(I, k)$ denote the set of *all* partitions of I of size k. And recall that the grid $R_{I,\ell}$ is a range equipartion of I of size ℓ. Then for a fixed $L = L_x \times L_y$:

- for $2 \leq k \leq \ell$, the set $\mathcal{R}(L, \infty, k, [\ell])$ contains all grids $G = (P, Q)$ where $Q = R_{L_x,\ell}$ and $P \in \mathcal{P}(L_y, k)$.
- for any $2 \leq \ell < k$, the set $\mathcal{R}(L, \infty, [k], \ell)$ contains all grids $G = (P, Q)$ where $Q \in \mathcal{P}(L_x, \ell)$ and $P = R_{L_y,k}$.

The set $\mathcal{R}(L, \infty, k, \ell)$ is then defined analogously to $\mathcal{R}(L, c, k, \ell)$ from (23).

For any dataset D of size n restricted to the range $L = L_x \times L_y$, let $\mathbf{M}_D^{\mathcal{R}(L,\infty)}$ denote its *unconstrained* range-equicharacteristic matrix. For $k, \ell \geq 2$, it entries are given by

$$\left(\mathbf{M}_D^{\mathcal{R}(L,\infty)}\right)_{k,\ell} = \max_{G \in \mathcal{R}(L,\infty,k,\ell)} \frac{I(D|_G)}{\log_2 \min\{k,\ell\}} . \tag{27}$$

Then for a maximum grid size parameter $B = B(n)$, define

$$\text{MICr}(D, L, B, c) = \max_{k \cdot \ell \leq B} \left(\mathbf{M}_D^{\mathcal{R}(L,\infty)}\right)_{k,\ell} .$$

For a jointly-distributed pair of random variables $\Pi = (X, Y)$ with range restricted to L, we similarly define the matrix $\mathbf{M}_\Pi^{\mathcal{R}(L,\infty)}$ with entries

$$\left(\mathbf{M}_\Pi^{\mathcal{R}(L,\infty)}\right)_{k,\ell} = \max_{G \in \mathcal{R}(L,\infty,k,\ell)} \frac{I(\Pi|_G)}{\log_2 \min\{k,\ell\}} . \tag{28}$$

B.2. MICr Consistency

Using the notation introduced in the previous section, we now develop the proof of Theorem 4.2, which shows that the MICr$(\cdot, L, B, c)$ statistic is a consistent estimator of MIC*. However, our first step is to show that the *unconstrained* MICr$(\cdot, L, B, \infty)$ statistic is a consistent estimator of MIC*. Once this step is established, we leverage the relationship between entries of $\mathbf{M}^{\mathcal{R}(L,c)}$ and $\mathbf{M}^{\mathcal{R}(L,\infty)}$ to prove the consistency of MICr$(\cdot, L, B, c)$.

B.2.1. CONSISTENCY OF THE MICr$(\cdot, L, B, \infty)$ ESTIMATOR

We will prove the following Theorem:

Theorem B.2. *Let $\Pi = (X, Y)$ be a jointly-distributed pair of random variables with range bounded by L, and let D_n be a dataset of n points sampled i.i.d. from Π. For every $0 < \alpha < 0.5$, and for every $\omega(1) \leq B(n) = O(n^\alpha)$,*

$$\text{MICr}(D_n, L, B(n), \infty) \longrightarrow \text{MIC*}(\Pi)$$

in probability as $n \to \infty$.

The proof of this theorem uses the following two lemmas, which use arguments adapted from Reshef et al. (2016) and Lazarsfeld & Johnson (2021), respectively.

The first lemma shows that, in the distributional setting, the supremum of the range-equicharacteristic matrix $\mathbf{M}_{\Pi}^{\mathcal{R}(L,\infty)}$ and the standard characteristic matrix $\mathbf{M}_{\Pi}^{\mathcal{G}}$ are equal, the latter of which is equal to MIC*(Π) by definition.

Lemma B.3. *Let $\Pi = (X, Y)$ be a pair of jointly-distributed random variables with range bounded by $L = L_x \times L_y$. Then*

$$\sup\ \mathbf{M}_{\Pi}^{\mathcal{R}(L,\infty)} \ = \ \sup \mathbf{M}_{\Pi}^{\mathcal{G}} \ = \ \text{MIC*}(\Pi)\ .$$

Proof. The statement follows by an argument identical to the proof of Theorem 21 from (Reshef et al., 2016). In particular, the mutual information of $I(X, Y)$ is the supremum of $I(X|_Q, Y)$ over all finite partitions Q. By definition, the (k, ℓ)'th entry of $\mathbf{M}_{\pi}^{\mathcal{R}(L,\infty)}$ uses the column partition $R_{L_x,\ell}$, which is a range-equipartition of L_x of size ℓ, to partition X. Thus taking ℓ in the limit, $I(X|_{R_{L_x,\ell}}, Y) = I(X, Y)$. The remainder of the argument in the original proof carries over unchanged. □

The next lemma shows that, when n is sufficiently large, for a dataset D_n of n points sampled i.i.d. from Π, a subset of corresponding entries (k, ℓ) in $\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}$ and $\mathbf{M}_{\Pi}^{\mathcal{R}(L,\infty)}$ have similar values with high probability.

Lemma B.4. *Let $\Pi = (X, Y)$ be a pair of jointly-distributed random variables with range bounded by L, and let D_n be a dataset of n points sampled i.i.d. from Π. For all n, and for every $0 < \alpha < 0.5$, there exists a constant $u > 0$ such that*

$$\left| \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} - \left(\mathbf{M}_{\Pi}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} \right| \ = \ O\left(\frac{1}{n^u}\right)$$

simultaneously for every $k\ell \le B(n) = O(n^\alpha)$ with probability at least $1 - O(n^{-1.5})$.

Proof. Because the the (k, ℓ) entries of $\mathbf{M}_{\Pi}^{\mathcal{R}(L,\infty)}$ and $\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}$ are both maximizations over the same set of grids $\mathcal{R}(L, \infty, k, \ell)$, the proof follows identically to that of Theorem 5 of Lazarsfeld & Johnson (2021). □

The proof of Theorem B.2 then follows.

Proof (of Theorem B.2). The proof is identical to that of Theorem 6 from Reshef et al. (2016). The exact argument goes through so long as (1), $\sup \mathbf{M}_{\Pi}^{\mathcal{R}(L,\infty)} = \sup \mathbf{M}_{\Pi}^{\mathcal{G}}$, and (2), the (k, ℓ)'th entry of $\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}$ converges in probability to the (k, ℓ)'th entry of $\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}$ as $n \to \infty$ for all $k\ell \le B(n) = O(n^\alpha)$. These arguments follow from Lemmas B.3 and B.4, respectively. □

B.2.2. Consistency of the MICr$(\cdot, L, B, c)$ estimator

We now use the consistency of MICr$(\cdot, L, B, \infty)$ to prove the consistency of MICr$(\cdot, L, B, c)$ for finite c. This was stated informally in Theorem 4.2 and is presented in more detail here.

Theorem B.5. *Let $\Pi = (X, Y)$ be a joint distribution with range bounded by L, and let D_n be a dataset of n points sampled i.i.d. from Π. For every finite $c > 0$, $\alpha \in (0, 0.5)$, and $\omega(1) \le B(n) = O(n^\alpha)$, the statistic* MICr$(D_n, L, B, c) \to$ MIC*(Π) *in probability as $n \to \infty$.*

To prove this theorem, we use (in addition to Theorem B.2) the following two analogs to Lemmas B.3 and B.4.

Lemma B.6. *Let $\Pi = (X, Y)$ be a pair of jointly-distributed random variables with range bounded by $L = L_x \times L_y$. Then for finite $c > 0$:*

$$\sup\ \mathbf{M}_{\Pi}^{\mathcal{R}(L,c)} \ = \ \sup \mathbf{M}_{\Pi}^{\mathcal{G}} \ = \ \text{MIC*}(\Pi)\ .$$

Proof. The proof follows identically to that of Proposition 8 from Reshef et al. (2016), which gives an analogous argument for the MICe(Π, c) case. The key observation is that, wlog for $k \le \ell$, the (k, ℓ)'th entry of $\mathbf{M}_{\Pi}^{\mathcal{R}(L,c)}$ is a maximization over row partitions $P \in \mathcal{P}(L_y, k, [c\ell])$, which recall is the set of all size-k sub-partitions of a size-$(c\ell)$ master range-equipartition of L_y. So as $\ell \to \infty$, the set $\mathcal{P}(L_y, k, [c\ell])$ approaches $\mathcal{P}(L_y, k)$. The remainder of the original argument then follows identically. □

Lemma B.7. *Let $\Pi = (X, Y)$ be a pair of jointly-distributed random variables with range bounded by L, fix a finite $c > 0$, and let D_n be a dataset of n points sampled i.i.d. from Π. For all n, and for every $0 < \alpha < 0.5$, there exists a constant $u > 0$ such that*

$$\left| \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \left(\mathbf{M}_{\Pi}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| = O\left(\frac{1}{n^u}\right)$$

simultaneously for every $k\ell \leq B(n) = O(n^\alpha)$ with probability at least $1 - O(n^{-1.5})$.

Proof. Similar to Lemma B.4, because the (k, ℓ) entries of $\mathbf{M}_{\Pi}^{\mathcal{R}(L,c)}$ and $\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}$ are both maximizations over the same set of grids $\mathcal{R}(L, c, k, \ell)$, the proof follows identically to that of Theorem 5 of Lazarsfeld & Johnson (2021). $\square$

In addition to these two lemmas, we state the following additional inequality that relates corresponding entries of $\mathbf{M}_{D}^{\mathcal{R}(L,c)}$ and $\mathbf{M}_{D}^{\mathcal{R}(L,\infty)}$ for some dataset D and finite c.

Lemma B.8. *Fix any dataset D with range restricted to L and any finite $c > 0$. Then for all $k, \ell \geq 2$*

$$\left(\mathbf{M}_{D}^{\mathcal{R}(L,c)}\right)_{k,\ell} \leq \left(\mathbf{M}_{D}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} . \tag{29}$$

Proof. By definition, for any finite $c > 0$ and any $k, \ell \geq 2$ the set $\mathcal{R}(L, c, k, \ell)$ is a subset of $\mathcal{R}(L, \infty, k, \ell)$. For a fixed dataset D, the (k, ℓ) entries of $\mathbf{M}_{D}^{\mathcal{R}(L,c)}$ and $\mathbf{M}_{D}^{\mathcal{R}(L,\infty)}$ are maximizations of a common function over the sets $\mathcal{R}(L, c, k, \ell)$ and $\mathcal{R}(L, \infty, k, \ell)$, respectively. Increasing the size of the constraint set can never decrease the maximum function value subject to the constraints, so the statement of the lemma follows. $\square$

Using these lemmas, we are now ready to prove Theorem B.5, which we do via a more direct adaptation of the proof of Lemma H.4 from Reshef et al. (2016).

Proof (of Theorem B.5). For any $0 < \alpha < 0.5$, fix $B(n) = O(n^\alpha)$. Our goal is to show that for a dataset D_n of n points sampled i.i.d. from Π, that

$$\text{MICr}(D_n, L, B(n), c) \longrightarrow \text{MIC*}(\Pi)$$

in probability as $n \to \infty$.

Recall that by definition, MIC*$(\Pi) = \sup \mathbf{M}_{\Pi}^{\mathcal{G}}$, so this means that for every $\epsilon > 0$ and every $0 < p \leq 1$, we must show that there exists some N such that

$$\Pr\left(\left| \max_{k,\ell\,:\,k\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \sup \mathbf{M}_{\Pi}^{\mathcal{G}} \right| < \epsilon \right) > 1 - p \tag{30}$$

for all $n \geq N$.

Fix ϵ and p. By Lemma B.6, we know $\sup M_{\Pi}^{\mathcal{R}(L,c)} = \sup \mathbf{M}_{\Pi}^{\mathcal{G}}$. So by definition, there must exist an entry (k, ℓ) such that

$$\left| \left(\mathbf{M}_{\Pi}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \sup\ \mathbf{M}_{\Pi}^{\mathcal{G}} \right| < (\epsilon/2) . \tag{31}$$

Denote such an entry by $(k_\epsilon, \ell_\epsilon)$.

Now by Lemma B.7, there exists some N_ϵ such that, for all $n \geq N_\epsilon$:

$$\Pr\left(\left| \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k_\epsilon,\ell_\epsilon} - \left(\mathbf{M}_{\Pi}^{\mathcal{R}(L,c)}\right)_{k_\epsilon,\ell_\epsilon} \right| < (\epsilon/2) \right) > 1 - p . \tag{32}$$

So with probability greater than $1 - p$, for all $n \geq N_\epsilon$,

$$\begin{aligned} \left| \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k_\epsilon,\ell_\epsilon} - \sup\ \mathbf{M}_{\Pi}^{\mathcal{G}} \right| &\leq \left| \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k_\epsilon,\ell_\epsilon} - \left(\mathbf{M}_{\Pi}^{\mathcal{R}(L,c)}\right)_{k_\epsilon,\ell_\epsilon} \right| + \left| \left(\mathbf{M}_{\Pi}^{\mathcal{R}(L,c)}\right)_{k_\epsilon,\ell_\epsilon} - \sup\ \mathbf{M}_{\Pi}^{\mathcal{G}} \right| \\ &< (\epsilon/2) + (\epsilon/2) \\ &= \epsilon . \end{aligned} \tag{33}$$

Here, the first line is due to the triangle inequality, and the second line is due to the bounds from (31) and (32).

Now let N_α be some integer such that, $B(N_\alpha) \geq k_\epsilon \cdot \ell_\epsilon$. Since $B(n) = O(n^\alpha)$ is increasing in n, this implies that $k_\epsilon \cdot \ell_\epsilon \leq B(N_\alpha) \leq B(n)$ for all $n \geq N_\alpha$, which means

$$\left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k_\epsilon,\ell_\epsilon} \leq \max_{k,\ell:\, k\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} . \tag{34}$$

Then for all $n \geq \max\{N_\epsilon, N_\alpha\}$, it follows that

$$\max_{k,\ell:\, k\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \geq \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k_\epsilon,\ell_\epsilon} > \sup \mathbf{M}_\Pi^{\mathcal{G}} - \epsilon \tag{35}$$

with probability greater than $1-p$, where the first inequality holds for $n \geq N_\alpha$ by (34), and the second inequality holds for $n \geq N_\epsilon$ by (33). This implies one direction of our goal from (30).

For the other direction, consider any $k, \ell \geq 2$. By Lemma B.8, we have for any n and fixed D_n that

$$\left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \leq \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} .$$

This implies that

$$\max_{k,\ell:\, k\cdot\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \leq \max_{k,\ell:\, k\cdot\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} . \tag{36}$$

Now by Theorem B.2, we know that

$$\max_{k,\ell:\, k\cdot\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} \longrightarrow \sup \mathbf{M}_\Pi^{\mathcal{G}}$$

in probability as $n \to \infty$. So there must exist some N_∞ such that for all $n \geq N_\infty$

$$\begin{aligned} &\left| \max_{k,\ell:\, k\cdot\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} - \sup \mathbf{M}_\Pi^{\mathcal{G}} \right| < \epsilon \\ &\implies \max_{k,\ell:\, k\cdot\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} < \sup \mathbf{M}_\Pi^{\mathcal{G}} + \epsilon \end{aligned} \tag{37}$$

with probability greater than $1-p$.

Then combining (36) and (37) says that for all $n \geq N_\infty$, with probability greater than $1-p$:

$$\max_{k,\ell:\, k\cdot\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \leq \max_{k,\ell:\, k\cdot\ell \leq B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,\infty)}\right)_{k,\ell} < \sup \mathbf{M}_\Pi^{\mathcal{G}} + \epsilon . \tag{38}$$

This implies the second direction of our goal from (30).

Setting $N \geq \max\{N_\epsilon, N_\alpha, N_\infty\}$ ensures that for all $n \geq N$, expression (30) always holds, which completes the proof. □

B.3. MICr Sensitivity

In this section, we develop the proof of Theorem 4.3, which gives an upper bound on the sensitivity of the MICr statistic. The high-level strategy is to reduce the sensitivity of the statistic to the sensitivity of mutual information with respect to a fixed grid, similar to the arguments developed in Section A.2.2 for the bound on the MICe sensitivity.

For a fixed dataset D of n points, a set of range limits L, a constant $c > 0$, and a maximum grid parameter $B := B(n)$, recall that

$$\mathrm{MICr}(D, B, L, c) = \left(\mathbf{M}_D^{\mathcal{R}(L,c)}\right)_{k,\ell} , \tag{39}$$

where $\mathbf{M}_D^{\mathcal{R}(L,c)}$ is the range-equicharacteristic matrix for D.

Now for any pair of neighboring datasets $D \sim D'$ of size n and fixed $B := B(n)$, L, and c:

$$
\begin{aligned}
|\ \mathrm{MICr}(D, B, L, c) - \mathrm{MICr}(D', B, L, c)\ | \ &= \left| \max_{k,\ell\,:\,k\cdot\ell \le B} \left(\mathbf{M}_D^{\mathcal{R}(L,c)}\right)_{k,\ell} - \max_{k,\ell\,:\,k\cdot\ell \le B} \left(\mathbf{M}_{D'}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| \\
&\le \max_{k,\ell\,:\,k\cdot\ell \le B} \left| \left(\mathbf{M}_D^{\mathcal{R}(L,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D'}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| . \qquad (40)
\end{aligned}
$$

As before, we will assume wlog that $k \le \ell$, as the corresponding arguments for the $k > \ell$ case follow by symmetry. So by the definition from (23), for any fixed $2 \le k \le \ell$, observe that

$$
\begin{aligned}
\left| \left(\mathbf{M}_D^{\mathcal{R}(L,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D'}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| &= \left| \max_{G\in\mathcal{R}(L,c,k,[\ell])} \frac{I\left(D|_G\right)}{\log_2 k} - \max_{G\in\mathcal{R}(L,c,k,[\ell])} \frac{I\left(D'|_G\right)}{\log_2 k} \right| \\
&\le \left| \max_{G\in\mathcal{R}(L,c,k,[\ell])} I\left(D|_G\right) - \max_{G\in\mathcal{R}(L,c,k,[\ell])} I\left(D'|_G\right) \right| \\
&\le \max_{G\in\mathcal{R}(L,c,k,[\ell])} \left|\ I\left(D|_G\right) - I\left(D'|_G\right)\ \right| . \qquad (41)
\end{aligned}
$$

Here, the first inequality is due to factoring out the $\log_2 k$ term and noting that $\log_2 k \ge 1$ when $k \ge 2$, and the final inequality follows from the fact that the set $\mathcal{R}(L, c, k, [\ell])$ is the same in both maximization terms.

We then claim the following upper bound on (41) that holds for any $D \sim D'$ of size n:

Lemma B.9. *For any $n \ge 4$ and any set of range bounds L, consider any pair of neighboring datasets $D \sim D'$, both of size n, and fix a pair $2 \le k \le \ell$. Then for any grid $G \in \mathcal{R}(L, c, k, [\ell])$:*

$$
|\ I(D|_G) - I(D'|_G)\ | \ \le \ \frac{4\log_2 n}{n} + \frac{6}{n} .
$$

Observe that this lemma holds for any $G \in \mathcal{R}(L, c, k, [\ell])$, and thus it holds for the maximizing grid in (41). Granting the lemma true for now then implies the upper bound on the MICr statistic (i.e., the proof of Theorem 4.3):

Proof (of Theorem 4.3). Apply Lemma B.9 to (41). Subtituting back into (40) implies that for any $D \sim D'$ of size $n \ge 4$ and any B:

$$
|\ \mathrm{MICr}(D, B, L, c) - \mathrm{MICr}(D', B, L, c)\ | \ \le \max_{k,\ell\,:\,k\cdot\ell \le B} \frac{4\log_2 n}{n} + \frac{6}{n} \ = \ \frac{4\log_2 n}{n} + \frac{6}{n} .
$$

The final equality follows by observing that the term from Lemma B.9 does not depend on k or ℓ. □

It now remains to prove Lemma B.9.

B.3.1. Proof of Lemma B.9

The proof follows similarly to that of Lemma A.5. Fix any $n \ge 4$, any $2 \le k \le \ell$, and any set of range boundaries L, and consider any pair of neighboring datasets $D \sim D'$, both of size n. Now consider any grid $G \in \mathcal{R}(L, c, k, [\ell])$. Let

$$
\begin{aligned}
\mathbf{P} \ &:= \ \mathbf{P}_{D,G} \quad \text{with entries } p(i,j), \text{ row sums } p(i,*), \text{ and column sums } p(*,j) \\
\mathbf{P}' \ &:= \ \mathbf{P}_{D',G} \quad \text{with entries } p'(i,j), \text{ row sums } p'(i,*), \text{ and column sums } p'(*,j) ,
\end{aligned}
$$

and let $\mathbf{A}$ and $\mathbf{A}'$ denote the corresponding count matrices for $\mathbf{P}$ and $\mathbf{P}'$ respectively. Because $\mathbf{P}$ and $\mathbf{P}'$ are non-negative matrices representing two-dimensional, discrete joint distributions, we define the mutual information of $\mathbf{P}$ and $\mathbf{P}'$ in the natural way. Specifically, let

$$
I(\mathbf{P}) \ = \ \sum_{i\in[k]} \sum_{j\in[\ell]} p(i,j) \log_2 \left(\frac{p(i,j)}{p(i,*)p(*,j)} \right) .
$$

Then to upper bound $|I(\mathbf{P}) - I(\mathbf{P}')|$, we have

$$| \ I(\mathbf{P}) - I(\mathbf{P}') \ | \ \leq \ \sum_{i \in [k]} \sum_{j \in [\ell]} \left| \ p(i,j) \log_2 \left(\frac{p(i,j)}{p(i,*)p(*,j)} \right) - p'(i,j) \log_2 \left(\frac{p'(i,j)}{p'(i,*)p'(*,j)} \right) \ \right| \ , \tag{42}$$

which follows from the triangle inequality.

For any $(i,j) \in [k] \times [\ell]$, let

$$\gamma(i,j) \ = \ \left| \ p(i,j) \log_2 \left(\frac{p(i,j)}{p(i,*)p(*,j)} \right) - p'(i,j) \log_2 \left(\frac{p'(i,j)}{p'(i,*)p'(*,j)} \right) \ \right| \ .$$

Now because G is fixed with respect to both D and D', observe that there at most two entries (i,j) such that $p(i,j) \neq p'(i,j)$, and thus at most two pairs (i,j) such that $\gamma(i,j) > 0$. Specifically for any $D \sim D'$, there are only two scenarios regarding how the entries of $\mathbf{P}$ and $\mathbf{P}'$ differ:

1. *Scenario A*: there is at most 1 pair of points $d \in D$ and $d' \in D'$ with nonequal coordinates, but $\phi(d,G) = \phi(d',G)$.

   Then for all $(i,j) \in [k] \times [\ell]$, $p(i,j) = p'(i,j)$ and thus $\gamma(i,j) = 0$.

   By (42), this means
   $$|I(\mathbf{P}) - I(\mathbf{P}')| = 0 \ .$$

2. *Scenario B*: there is at most 1 pair of points $d \in D$ and $d' \in D'$ with nonequal coordinates, and $\phi(d,G) \neq \phi(d',G)$.

   Then there is exactly one $(i,j) \in [k] \times [\ell]$ such that $p'(i,j) = p(i,j) - 1/n$ and thus $\gamma(i,j) \geq 0$. For this unique (i,j), we denote $\gamma(i,j)$ by γ^-.

   There is also exactly one $(i,j) \in [k] \times [\ell]$ such that $p'(i,j) = p(i,j) + 1/n$ and thus $\gamma(i,j) \geq 0$. For this unique (i,j), we similarly denote $\gamma(i,j)$ by γ^+.

   By (42), this means that
   $$|I(\mathbf{P}) - I(\mathbf{P}')| \ \leq \ \gamma^+ + \gamma^- \ . \tag{43}$$

   Intuitively, γ^- corresponds to the entry (i,j) that loses a point in $\mathbf{A}'$ compared to the corresponding entry in $\mathbf{A}$. Similarly, γ^+ corresponds to the entry (i,j) that gains a point in $\mathbf{A}'$ compared to the correpsonding entry in $\mathbf{A}$.

   Because $|D| = |D'| = n$, both such entries must exist in this scenario.

Let us assume a $D \sim D'$ and G from Scenario B (otherwise the lemma follows trivially).

We claim the following bound on γ^-:

Claim B.10. *For any $n \geq 4$ and for $\mathbf{P}$ and $\mathbf{P}'$ as described above, let $(i,j) \in [k] \times [\ell]$ be the cell such that*

$$p'(i,j) = p(i,j) - \frac{1}{n} \ .$$

Then

$$\gamma^- \ := \ \gamma(i,j) \ \leq \ \frac{2 \log_2 n}{n} + \frac{4}{n} \ .$$

For γ^+, we have a similar claim:

Claim B.11. *For any $n \geq 4$ and for $\mathbf{P}$ and $\mathbf{P}'$ as described above, let $(i,j) \in [k] \times [\ell]$ be the cell such that*

$$p'(i,j) = p(i,j) + \frac{1}{n} \ .$$

Then

$$\gamma^+ \ := \ \gamma(i,j) \ \leq \ \frac{2 \log_2 n}{n} + \frac{2}{n} \ .$$

If we grant these claims true for now, then by expression (43) it follows that

$$| \; I(\mathbf{P}) - I(\mathbf{P}') \; | \;\leq\; \frac{2\log_2 n}{n} + \frac{4}{n} + \frac{2\log_2 n}{n} + \frac{2}{n} \;=\; \frac{4\log_2 n}{n} + \frac{6}{n} \;,$$

which is the statement of the lemma. So it only remains to prove the two claims.

Proof (of Claim B.10). For readability, let $p := p(i,j)$ and $p' := p'(i,j)$. Since we assume $p' = p - 1/n$, we have that

$$\begin{aligned} p(i,*) - 1/n \leq p'(i,*) \leq p(i,*) \\ p(*,j) - 1/n \leq p'(*,j) \leq p(*,j) \;\; . \end{aligned}$$

Moreover, we have

$$\gamma^- \;=\; \left| \, p\log_2 \frac{p}{p(i,*)p(*,j)} \;-\; (p-1/n)\log_2 \frac{p-1/n}{p'(i,*)p'(*,j)} \, \right| \;.$$

Observe that when $p = 1/n$ and $p' = 0$, then

$$\gamma^- \;=\; \left| \, \frac{1}{n}\log_2 \frac{1/n}{p(i,*)p(*,j)} \, \right| \;.$$

When $p = 1/n$, then $1/n \leq p(i,*), p(*,j) \leq 1$ (and thus $1/n^2 \leq p(i,*)\cdot p(*,j) \leq 1$), so

$$\left| \, \frac{1}{n}\log_2 \frac{1/n}{p(i,*)p(*,j)} \, \right| \;\leq\; \max\left\{ \frac{1}{n}\log_2 \frac{1/n}{1/n^2}, \; \frac{1}{n}\log_2 \frac{1}{1/n} \right\} \;=\; \frac{\log_2 n}{n}$$

On the other hand, consider when $p \geq 2/n$ and $1/n \leq p' \leq 1 - 1/n$. We then have

$$\begin{aligned} \gamma^- \;&=\; \left| \, p\log_2 \left(\frac{p}{p-1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \right) \;+\; \frac{1}{n}\log_2 \frac{p-1/n}{p'(i,*)p'(*,j)} \, \right| \\ &\leq\; \left| \, p\log_2 \left(\frac{p}{p-1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \right) \, \right| \;+\; \left| \, \frac{1}{n}\log_2 \frac{p-1/n}{p'(i,*)p'(*,j)} \, \right| \;, \end{aligned} \tag{44}$$

where we use the triangle inequality.

For the left hand term in (44), say that

$$\frac{p}{p-1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \;>\; 1.$$

Then because $p'(i,*)\cdot p'(*,j) \leq p(i,*)\cdot p(*,j)$,

$$\left| \, p\log_2 \left(\frac{p}{p-1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \right) \, \right| \;\leq\; \left| \, p\log_2 \left(\frac{p}{p-1/n} \right) \, \right| \;\leq\; \left| \, \frac{2}{n}\log_2 \left(\frac{2/n}{2/n-1/n} \right) \, \right| \;=\; \frac{2}{n} \;.$$

Here, the final inequality is because the function is decreasing in p, and $p \geq 2/n$ by assumption.

On the other hand, consider when

$$\frac{p}{p-1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \;<\; 1.$$

Then because $p \leq p(i,*), p(*,j) \leq 1$, and because $p'(i,*)\cdot p'(*,j) \geq (p(i,*) - 1/n)\cdot(p(*,j) - 1/n)$:

$$\begin{aligned} \left| \, p\log_2 \left(\frac{p}{p-1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \right) \, \right| \;&=\; \left| \, p\log_2 \left(\frac{p-1/n}{p} \cdot \frac{p(i,*)p(*,j)}{p'(i,*)p'(*,j)} \right) \, \right| \\ &\leq\; \left| \, p\log_2 \left(\frac{p(i,*)p(*,j)}{(p(i,*)-1/n)(p(*,j)-1/n)} \right) \, \right| \\ &\leq\; \left| \, p(i,*)\log_2 \left(\frac{p(i,*)}{p(i,*)-1/n} \right) \;+\; p(*,j)\log_2 \left(\frac{p(*,j)}{p(*,j)-1/n} \right) \, \right| \;. \end{aligned}$$

Here, the two terms are decreasing in $p(i,*)$ and $p(*,j)$, respectively. Since $p(i,*), p(*,j) \geq p \geq 2/n$ by assumption, we have

$$\left| p(i,*) \log_2 \left(\frac{p(i,*)}{p(i,*) - 1/n} \right) \; + \; p(*,j) \log_2 \left(\frac{p(*,j)}{p(*,j) - 1/n} \right) \right| \tag{45}$$

$$\leq \left| \frac{2}{n} \log_2 \left(\frac{2/n}{2/n - 1/n} \right) \; + \; \frac{2}{n} \log_2 \left(\frac{2/n}{2/n - 1/n} \right) \right| \; = \; \frac{4}{n} \; , \tag{46}$$

which bounds the left hand term of (44) in general.

For the right hand term of (44), if

$$\frac{1}{n} \cdot \frac{p - 1/n}{p'(i,*)p'(*,j)} \; < \; 1,$$

then

$$\begin{aligned} \left| \frac{1}{n} \log_2 \frac{p - 1/n}{p'(i,*)p'(*,j)} \right| \; &= \; \left| \frac{1}{n} \log_2 \frac{p'(i,*)p'(*,j)}{p - 1/n} \right| \\ &\leq \; \left| \frac{1}{n} \log_2 \frac{1}{2/n - 1/n} \right| \; = \; \frac{\log_2 n}{n} \; . \end{aligned}$$

Here, the penultimate inequality is due to $p'(i,*) \cdot p'(*,j) \leq 1$ and $p \geq 2/n$ by assumption.

On the other hand, consider when

$$\frac{1}{n} \cdot \frac{p - 1/n}{p'(i,*)p'(*,j)} \; > \; 1 \; .$$

Then for the right hand term of (44), we have

$$\left| \frac{1}{n} \log_2 \frac{p - 1/n}{p'(i,*)p'(*,j)} \right| \; \leq \; \left| \frac{1}{n} \log_2 \frac{p - 1/n}{(p(i,*) - 1/n)(p(*,j) - 1/n)} \right| \; .$$

Now because $p \leq \min\{p(i,*), p(*,j)\}$, then

$$\begin{aligned} \left| \frac{1}{n} \log_2 \frac{p - 1/n}{p'(i,*)p'(*,j)} \right| \; &\leq \; \max \left\{ \left| \frac{1}{n} \log_2 \frac{1}{p(i,*) - 1/n} \right| , \left| \frac{1}{n} \log_2 \frac{1}{p(*,j) - 1/n} \right| \right\} \\ &\leq \; \left| \frac{1}{n} \log_2 \frac{1}{p(i,*) - 1/n} \right| \; + \; \left| \frac{1}{n} \log_2 \frac{1}{p(*,j) - 1/n} \right| \; , \end{aligned}$$

where the inequality is due to $\max\{x, y\} \leq x + y$ for all $x, y \geq 0$. Because the terms in the sum are decreasing in $p(i,*)$ and $p(*,j)$ respectively, and because $p(i,*), p(*,j) \geq 2/n$ by assumption, we have

$$\left| \frac{1}{n} \log_2 \frac{1}{p(i,*) - 1/n} \right| \; + \; \left| \frac{1}{n} \log_2 \frac{1}{p(*,j) - 1/n} \right| \; \leq \; \frac{2 \log_2 n}{n} \; . \tag{47}$$

This bounds the right hand term of (44) in general.

Now substituting the bounds (46) and (47) back into (44), we have

$$\gamma^- \; \leq \; \frac{2 \log_2 n}{n} + \frac{4}{n} \; ,$$

which completes the proof. □

The proof of Claim B.11 follows similarly:

Proof (of Claim B.11). Again for readability let $p := p(i,j)$ and $p' := p'(i,j)$. Since we assume $p' = p + 1/n$, we have that

$$\begin{aligned} p(i,*) \leq p'(i,*) \leq p(i,*) + 1/n \leq 1 \\ p(*,j) \leq p'(*,j) \leq p(*,j) + 1/n \leq 1 \end{aligned}$$

Then

$$\gamma^- \;=\; \left| \, p\log_2 \frac{p}{p(i,*)p(*,j)} \;-\; (p+1/n)\log_2 \frac{p+1/n}{p'(i,*)p'(*,j)} \, \right| .$$

Now observe that when $p = 0$ and $p' = 1/n$, then

$$\gamma^+ \;=\; \left| \, \frac{1}{n}\log_2 \frac{1/n}{p'(i,*)p'(*,j)} \, \right| \;\le\; \frac{\log_2 n}{n}$$

where the inequality is due to $1/n^2 \le p'(i,*) \cdot p'(*,j) \le 1$ since $p'(i,*), p'(*,j) \ge p' = 1/n$.

Consider now when $p \ge 1/n$. Then

$$\begin{aligned}
\gamma^+ \;&=\; \left| \, p\log_2\left(\frac{p}{p+1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)}\right) \;+\; \frac{1}{n}\log_2 \frac{p+1/n}{p'(i,*)p'(*,j)} \, \right| \\
&\le\; \left| \, p\log_2\left(\frac{p}{p+1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)}\right) \, \right| \;+\; \left| \, \frac{1}{n}\log_2 \frac{p+1/n}{p'(i,*)p'(*,j)} \, \right| ,
\end{aligned} \tag{48}$$

which follows by the triangle inequality.

For the left hand term in (44), say that

$$\frac{p}{p+1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \;>\; 1 .$$

Then

$$\begin{aligned}
\left| \, p\log_2\left(\frac{p}{p+1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)}\right) \, \right| \;&\le\; \left| \, p\log_2\left(\frac{p}{p+1/n} \cdot \frac{(p(i,*)+1/n)(p(*,j)+1/n)}{p(i,*)p(*,j)}\right) \, \right| \\
&\le\; \left| \, p\log_2\left(\frac{p}{p+1/n} \cdot \frac{1}{((n-1)/n)^2}\right) \, \right| \\
&\le\; \left| \, \frac{n-1}{n}\log_2\left(\frac{(n-1)/n}{1} \cdot \frac{n^2}{(n-1)^2}\right) \, \right| \\
&\le\; \left| \, \log_2\left(\frac{n}{n-1}\right) \, \right| .
\end{aligned}$$

Here, the second inequality holds given that the expression is increasing in both $p(i,*)$ and $p(*,j)$, both of which are at most $(n-1)/n$. The third inequality follows because the expression is increasing with $p \le (n-1)/n$.

Now because $2^{(29/20)} \approx 2.73 \ge e$, it follows that for all numbers x, $x \le e^x \le 2^{(29/20)x}$, which implies $\log_2 x \le (29/20)x$. Because $n/(n-1) \le (n+40/29)/n$ for all $n \ge 4$, then applying this identity means

$$\log_2 \frac{n}{n-1} \;\le\; \log_2 \frac{n+40/29}{n} \;\le\; \frac{29}{20} \cdot \frac{40}{29n} \;=\; \frac{2}{n} \tag{49}$$

for all $n \ge 4$, which holds by assumption of the claim. So in this case, the left hand term (48) is at most $2/n$.

On the other hand, suppose

$$\frac{p}{p+1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \;<\; 1 .$$

Then

$$\begin{aligned}
\left| \, p\log_2\left(\frac{p}{p+1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)}\right) \, \right| \;&=\; \left| \, p\log_2\left(\frac{p+1/n}{p} \cdot \frac{p(i,*)p(*,j)}{p'(i,*)p'(*,j)}\right) \, \right| \\
&\le\; \left| \, p\log_2\left(\frac{p+1/n}{p}\right) \, \right| \\
&\le\; \left| \, \frac{n-1}{n}\log_2\left(\frac{1}{(n-1)/n}\right) \, \right| \;\le\; \left| \, \log_2 \frac{n}{n-1} \, \right| .
\end{aligned}$$

Here, the first inequality holds given that $p(i,*) \cdot p(*,j) \leq p'(i,*) \cdot p'(j,*)$. The second inequality follows because the expression is increasing in p, which can be at most $(n-1)/n$ by assumption.

Applying the bound from (49) then yields

$$\left| p \log_2 \left(\frac{p}{p+1/n} \cdot \frac{p'(i,*)p'(*,j)}{p(i,*)p(*,j)} \right) \right| \leq \frac{2}{n} , \tag{50}$$

which bounds the left hand term of (48) in general.

For the right hand term of (48), consider when

$$\frac{p+1/n}{p'(i,*)p'(*,j)} < 1 .$$

Then because $p'(i,*), p'(*,j) \leq 1$ and $p \geq 1/n$ by assumption, the right hand term of (48) becomes

$$\left| \frac{1}{n} \log_2 \left(\frac{p+1/n}{p'(i,*)p'(*,j)} \right) \right| = \left| \frac{1}{n} \log_2 \left(\frac{p'(i,*)p'(*,j)}{p+1/n} \right) \right| \leq \left| \frac{1}{n} \log_2 \left(\frac{1}{1/n+1/n} \right) \right| \leq \frac{\log_2 n}{n} .$$

On the other hand, consider when

$$\frac{p+1/n}{p'(i,*)p'(*,j)} > 1 .$$

Then because $p \leq 1 - 1/n$ and because $p'(i,*) \cdot p'(*,j) \geq 1/n^2$ by the assumption that $p'(i,*), p'(*,j) \geq p \geq 1/n$,

$$\left| \frac{1}{n} \log_2 \left(\frac{p+1/n}{p'(i,*)p'(*,j)} \right) \right| \leq \left| \frac{1}{n} \log_2 \left(\frac{1}{1/n^2} \right) \right| = \frac{2 \log_2 n}{n} , \tag{51}$$

which bounds the right hand term of (48) in general.

Combining the bounds (50) and (51) and substituting back into (48) then shows

$$\alpha^+ \leq \frac{2 \log_2 n}{n} + \frac{2}{n} ,$$

which completes the proof. □

C. MICr-Lap Details

In this section, we provide the proofs of Theorem 4.4 (MICr-Lap privacy) and Theorem C.1 (MICr-Lap consistency).

C.1. MICr-Lap Privacy

We begin with the privacy guarantee of MICr-Lap (restated for convenience), whose proof follows identically to that of Theorem 3.2, which gave the privacy guarantee for MICe-Lap:

Theorem 4.4. MICr-Lap$(\cdot, L, B, c, \epsilon)$ *is* ϵ*-DP.*

Proof. This follows by the privacy of the Laplace mechanism, and because truncating to the $[0,1]$ interval is a post-processing operation that, by Theorem A.7, preserves privacy. □

C.2. MICr-Lap Consistency

We now prove that MICr-Lap is a consistent estimator of MIC*.

Theorem C.1. *For every finite* $c > 0$, $\epsilon > 0$, $\alpha \in (0, 0.5)$, *and* $\omega(1) \leq B(n) = O(n^\alpha)$, MICr-Lap$(\cdot, L, B, c, \epsilon)$ *is a consistent estimator of* MIC*$(\cdot)$.

Proof. Recall from Section 4 that the standard deviation of the MICr-Lap mechanism is at most

$$(\sqrt{2}/\epsilon) \cdot ((4 \log_2 n)/n + 6/n) ,$$

which is decreasing with n for a fixed ϵ. Then coupled with the fact that (non-private) MICr is a consistent estimator of MIC* when $\omega(1) \leq B(n) = O(n^\alpha)$ for $\alpha \in (0, 0.5)$ (Theorem B.5), it follows that MICr-Lap is *also* a consistent estimator under the same parameter settings of B for any c and ϵ. □

D. MICr-Geom Details

In this section, we provide more details on the MICr-Geom mechanism and give the proofs of its privacy (Theorem 4.6) and consistency (Theorems 4.7 and D.4).

D.1. MICr-Geom Privacy

We start with the privacy guarantee of MICr-Geom from Theorem 4.6, which is restated as follows:

Theorem 4.6. MICr-Geom(D, L, B, c, ϵ) *is* ϵ*-DP.*

The proof uses the privacy of the TruncGeom mechanism from (Ghosh et al., 2012), as well as the composition property of differentially private mechanisms. We state these two tools here:

Theorem D.1 (Privacy of TruncGeom, (Ghosh et al., 2012))**.** *Assume that* f *is the value of some function of a dataset* $D \in \mathbb{R}^{2\times n}$ *with* ℓ_1 *sensitivity 1, and range* $\{0, \dots, n\}$*. Then for any* $\epsilon > 0$ *and* n*,* TruncGeom(ϵ, n, f) *is* ϵ*-DP.*

Theorem D.2 (General Composition of DP Mechanisms, (Dwork & Roth, 2014))**.** *For any mechanisms* $\mathcal{A}_1 : \mathcal{X}_1 \to \mathcal{Y}_1$ *and* $\mathcal{A}_2 : \mathcal{X}_2 \to \mathcal{Y}_2$*, if* $\mathcal{A}_1$ *and* $\mathcal{A}_2$ *are* ϵ_1 *and* ϵ_2*-DP, respectively, then the mechanism* $\mathcal{A} : (\mathcal{X}_1 \times \mathcal{X}_2) \to (\mathcal{Y}_1 \times \mathcal{Y}_2)$ *defined by* $\mathcal{A}(x_1, x_2) = (\mathcal{A}_1(x_1), \mathcal{A}_2(x_2))$ *for* $x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2$ *is* $(\epsilon_1 + \epsilon_2)$*-DP.*

The differential privacy of MICr-Geom follows from Theorems D.1, D.2, and from the post-processing principle from Theorem A.7:

Proof (of Theorem 4.6). For each $\ell \geq 2$, the counts of all entries of the noisy count matrix $\widehat{\mathbf{A}}$ are all independently $(\epsilon/2)$-DP via the privacy of the TruncGeom mechanism from Theorem D.1. Since at most two corresponding entries between $\mathbf{A}_{D,G}$ and $\mathbf{A}_{D',G}$ can have non-zero difference for any neighboring $D \sim D'$ and any fixed G, it follows from Theorem D.2 that the entire noisy count matrix $\widehat{\mathbf{A}}$ is ϵ-DP. The same argument holds independently for all $k \geq 2$. Then by the post-processing principle, every (k, ℓ) entry of $\widehat{\mathbf{M}}^{\mathcal{R}(L,c)}_{D,\epsilon}$ is independently ϵ-DP. Because the MICr-Geom(D, L, B, c, ϵ) statistic only returns one such entry, the final output of the mechanism is ϵ-DP. □

D.2. Computing MICr-Geom

Recall from the definition of MICr-Geom in Section 4 that for a fixed ℓ, we construct a $c\ell \times \ell$ matrix $\widehat{\mathbf{A}}_{D,\Gamma}$ where each entry is generated using the TruncGeom mechanism from Definition 4.5. Because the TruncGeom outputs range from 0 to n, for any n and ϵ, an array representing the exact PMF function of the TruncGeom distribution can be constructed in linear time, with subsequent samples done in constant time. Because the subsequent steps of computing MICr-Geom do not differ from computing MICr, the run time of the former is asymptotically equivalent to the latter (i.e., from Theorem B.1). Stated formally:

Theorem D.3. *For any dataset* D *of size* n *restricted to* L*,* $B := B(n)$*,* $c > 0$*, and* $\epsilon > 0$*,* MICr-Geom(D, L, B, c, ϵ) *can be computed using* OPTIMIZEAXIS *and a pre-computed* TruncGeom *distribution in* $O(c^2 B^4)$ *time.*

D.3. MICr-Geom Consistency

In this section, we develop the proof of Theorem D.4, which shows that the MICr-Geom statistic is a consistent estimator of MIC*. Formally, we prove the following theorem:

Theorem D.4. *For every finite* $c > 0$*,* $\epsilon > 0$*,* $\alpha \in (0, 0.5)$*, and* $\omega(1) \leq B(n) = O(n^\alpha)$*,* MICr-Geom$(\cdot, L, B, c, \epsilon)$ *is a consistent estimator of* MIC*$(\cdot)$.

The proof of the theorem leverages Theorem 4.7, which shows that the error introduced by the MICr-Geom mechanism in the (k, ℓ)'th entry of the range-equicharacteristic matrix vanishes as n grows. Again restated:

Theorem 4.7 (Added error of MICr-Geom)**.** *Fix any* $\alpha \in (0, 0.5)$*, finite* $c > 0$*,* $\epsilon > 0$*, and dataset* D *of size* n*. For sufficiently large* n*, there exists some* $a > 0$ *such that*

$$\left| \left(\widehat{\mathbf{M}}^{\mathcal{R}(L,c)}_{D,\epsilon}\right)_{k,\ell} - \left(\mathbf{M}^{\mathcal{R}(L,c)}_{D}\right)_{k,\ell} \right| = O((c/\epsilon)n^{-a})$$

for all $k\ell \leq B(n) = O(n^\alpha)$ *simultaneously with probability at least* $1 - O(n^{-2})$*.*

Granting Theorem 4.7 true for now, the proof of Theorem D.4 is straightforward:

Proof (of Theorem D.4). Fix any jointly-distributed pair of random variables $\Pi = (X, Y)$ with range bounded by L. For any $0 < \alpha < 0.5$, fix $B(n) = O(n^\alpha)$, and fix finite $c > 0$ and $\epsilon > 0$. Our goal is to show that

$$\text{MICr-Geom}(D_n, L, B(n), c, \epsilon) \longrightarrow \text{MIC*}(\Pi) = \sup \mathbf{M}_\Pi^{\mathcal{G}}$$

in probability as $n \to \infty$. This means that for every $\tau > 0$ and every $0 < p \le 1$, we must show that there exists some N such that

$$\Pr\left(\left| \max_{k,\ell\colon k\ell \le B(n)} \left(\widehat{\mathbf{M}}_{D_n,\epsilon}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \sup \mathbf{M}_\Pi^{\mathcal{G}} \right| < \tau \right) > 1 - p \tag{52}$$

for all $n \ge N$.

By the triangle inequality, observe that

$$\left| \max_{k,\ell\colon k\ell \le B(n)} \left(\widehat{\mathbf{M}}_{D_n,\epsilon}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \sup \mathbf{M}_\Pi^{\mathcal{G}} \right| \le \left| \max_{k,\ell\colon k\ell \le B(n)} \left(\widehat{\mathbf{M}}_{D_n,\epsilon}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \max_{k,\ell\colon k\ell \le B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| + \tag{53}$$

$$\left| \max_{k,\ell\colon k\ell \le B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \sup \mathbf{M}_\Pi^{\mathcal{G}} \right| . \tag{54}$$

By Theorem B.5 (consistency of the non-private MICr statistic), the term in (54) converges in probability to 0 as $n \to \infty$. Similary, by Theorem 4.7, the term in (53) converges in probability to 0 as $n \to \infty$. Thus for any τ and p, there exist integers N_1 and N_2 such that both terms are each bounded from above by $\tau/2$ with probability greater than $1 - p$ for all $n \ge \max\{N_1, N_2\}$, which completes the proof. □

It reamins to prove Theorem 4.7, which we develop in the next subsection.

D.4. Proof of Theorem 4.7

The proof of this theorem relies on deriving a bound on the change in mutual information between the noisy distribution $\widehat{\mathbf{P}}_{D,\Gamma}^{\epsilon}$ and the non-noisy $\mathbf{P}_{D,\Gamma}$, where Γ is the master range-equipartition grid for a fixed (wlog) $k \le \ell$.

First, observe that because

$$\left| \max_{k,\ell\colon k\ell \le B(n)} \left(\widehat{\mathbf{M}}_{D_n,\epsilon}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \max_{k,\ell\colon k\ell \le B(n)} \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| \le \max_{k,\ell\colon k\ell \le B(n)} \left| \left(\widehat{\mathbf{M}}_{D_n,\epsilon}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| \tag{55}$$

our goal will be to derive a sufficiently small uniform upper bound on

$$\left| \left(\widehat{\mathbf{M}}_{D_n,\epsilon}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| \tag{56}$$

that holds for any $k\ell \le B(n)$ with probability at least $1 - p$. Taking a union bound over all possible k, ℓ pairs (of which there at are most $B(n)^2$) means that the uniform bound on the absolute difference holds simultaneously for all $k\ell \le B(n)$ with probability at least $1 - B(n)^2 p$.

To derive an upper bound on (56), recall that for a fixed (wlog) $k \le \ell$, the grid $\Gamma := \Gamma_{c,\ell}$ is the master range-equipartition grid for $\mathcal{R}(L, c, k, \ell)$. Then

$$\left| \left(\widehat{\mathbf{M}}_{D_n,\epsilon}^{\mathcal{R}(L,c)}\right)_{k,\ell} - \left(\mathbf{M}_{D_n}^{\mathcal{R}(L,c)}\right)_{k,\ell} \right| = \left| \max_{G \in \mathcal{R}(L,c,k,\ell)} I^\star\left(\widehat{\mathbf{P}}_{D,\Gamma,G}^{\epsilon}\right) - \max_{G \in \mathcal{R}(L,c,k,\ell)} I^\star\left(\mathbf{P}_{D,\Gamma,G}\right) \right|$$

$$\le \max_{G \in \mathcal{R}(L,c,k,\ell)} \left| I\left(\widehat{\mathbf{P}}_{D,\Gamma,G}^{\epsilon}\right) - I\left(\mathbf{P}_{D,\Gamma,G}\right) \right| \tag{57}$$

where the last line is due to the triangle inequality and since $\log_2 k \ge 1$.

Now for suitable $B(n)$, and for a fixed $k \le \ell$ such that $k\ell \le B(n)$, we claim the following probabilistic bound on (57).

Lemma D.5. *Fix any finite $c > 0$, $\epsilon > 0$, $\alpha \in (0, 0.5)$, and $B(n) \le O(n^\alpha)$. For sufficiently large n and any $k\ell \le B(n)$, let Γ denote the master range-equipartition for $\mathcal{R}(L, c, k, \ell)$. Then there exists some $u > 0$ such that, for any dataset D of size n and $G \in \mathcal{R}(L, c, k, \ell)$,*

$$\left| I\left(\widehat{\mathbf{P}}_{D,\Gamma,G}^{\epsilon}\right) - I\left(\mathbf{P}_{D,\Gamma,G}\right) \right| \le O\left(\frac{c}{\epsilon n^u}\right)$$

with probability at least $1 - O(n^{-3})$.

If we again grant this lemma true for now, the proof of Theorem 4.7 follows:

Proof (of Theorem 4.7). By Lemma D.5, for any $k\ell \le B(n) = O(n^\alpha)$, expression (57) is bounded from above by $O(c/(\epsilon n^u))$ for some constant $u > 0$ (that depends on k, ℓ) with probability at least $1 - O(n^{-3})$. Let $a > 0$ be the minimum of all such constants over all $k\ell \le B(n)$. Then for all $k\ell \le B(n)$,

$$\left| \left(\widehat{\mathbf{M}}^{\mathcal{R}(L,c)}_{D_n,\epsilon}\right)_{k,\ell} - \left(\mathbf{M}^{\mathcal{R}(L,c)}_{D_n}\right)_{k,\ell} \right| \le O\left(\frac{c}{\epsilon n^a}\right)$$

with probability at least $1 - O(B(n)^2/n^3) \ge 1 - O(n^{-2})$. The error probability holds simultaneously over all $k\ell \le B(n)$ and follows by taking a union bound over all such entries, of which there at most $B(n)^2 \le O(n^{2\alpha}) \le O(n)$ since $\alpha < 0.5$. In particular, this means the bound holds for the maximizing entry of expression (55), which completes the proof. □

It remains to prove Lemma D.5, which requires the most technical machinery.

D.4.1. Proof of Lemma D.5

We will use a tool from (Reshef et al., 2016), which relates the statistical distance between two discrete distributions to their difference in mutual information. For two discrete distributions Π and Ψ over $[k] \times [\ell]$, let $D_{TV}(\Pi, \Psi)$ denote their statistical (total variation) distance, and recall that $D_{TV}(\Pi, \Psi) = \frac{1}{2} \cdot \|\Pi - \Psi\|_1$. We have the following lemma of Reshef et al.:

Lemma D.6 ((Reshef et al., 2016) Proposition 40, Appendix B)**.** *Let Π and Ψ be discrete distributions over $[k] \times [\ell]$ for $k, \ell \ge 2$. For any $0 < \delta \le 1/4$, if $D_{TV}(\Pi, \Psi) \le \delta$, then $|\ I(\Pi) - I(\Psi)\ | \le O\left(\delta \log_2((\min\{k, \ell\})/\delta)\right)$.*

For a fixed dataset D of size n, finite $c > 0$, a fixed k, ℓ, and a fixed $G \in \mathcal{R}(L, c, k, \ell)$ with master range-equipartition Γ, we will bound the difference in mutual information between $\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma,G}$ and $\mathbf{P}_{D,\Gamma G}$ by providing an upper bound on $D_{TV}(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma,G}, \mathbf{P}_{D,\Gamma,G})$. But given that every $G \in \mathcal{R}(L, c, k, \ell)$ is a subpartition of Γ, it follows by the triangle inequality that

$$D_{TV}(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma,G}, \mathbf{P}_{D,\Gamma,G}) \le D_{TV}(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma}, \mathbf{P}_{D,\Gamma}) , \tag{58}$$

and thus our goal will be to provide a probabilistic upper bound on $D_{TV}(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma}, \mathbf{P}_{D,\Gamma})$.

Assume wlog that $k \le \ell$, which means that the master range-equipartition grid Γ is of size $c\ell \times \ell$. Fix $\epsilon > 0$, and let $\widehat{p}(i,j)$ and $p(i,j)$ denote the (i,j) entries of $\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma}$ and $\mathbf{P}_{D,\Gamma}$. By definition of the MICr-Geom mechanism from Section 4.3, recall that for all $(i,j) \in [c\ell] \times [\ell]$, we define

$$\begin{aligned} \widehat{a}(i,j) &\sim \text{TruncGeom}(\epsilon/2, n, p(i,j) \cdot n) \\ \widehat{n} &= \textstyle\sum_{i,j} \widehat{a}(i,j) \\ \widehat{p}(i,j) &= \frac{\widehat{a}(i,j)}{\widehat{n}} . \end{aligned}$$

For all (i,j) let $\Delta_{i,j}$ be a random variable defined by

$$\Delta_{i,j} = \widehat{a}(i,j) - p(i,j) \cdot n , \tag{59}$$

which can be thought of as the change in number of datapoints in cell (i,j) after applying the TruncGeom mechanism. For all (i,j), the value $\Delta_{i,j}$ is distributed like a doubly-geometric random variable centered at 0 with parameter $e^{(-\epsilon/2)}$, but with truncation at $n - (p(i,j) \cdot n)$ and $-(p(i,j) \cdot n)$.

So we can rewrite each $\widehat{p}(i,j)$ as

$$\widehat{p}(i,j) = p(i,j) \cdot \frac{n}{\widehat{n}} + \frac{\Delta_{i,j}}{\widehat{n}} , \tag{60}$$

and moreover, we can write

$$\widehat{n} = n + \sum_{i,j} \Delta_{i,j} . \tag{61}$$

It follows that

$$\begin{aligned} D_{TV}(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma}, \mathbf{P}_{D,\Gamma}) &= \frac{1}{2}\sum_{i,j} |\ \widehat{p}(i,j) - p(i,j)\ | \\ &= \frac{1}{2}\sum_{i,j} \left|\ p(i,j)\cdot\frac{n}{\widehat{n}} + \frac{\Delta_{i,j}}{\widehat{n}}\ -\ p(i,j)\ \right| . \end{aligned}$$

By the triangle inequality, and increasing the $1/2$ term to 1, we have

$$\begin{aligned} D_{TV}(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma}, \mathbf{P}_{D,\Gamma}) &\le \sum_{i,j}\left|\ p(i,j)\left(\frac{n}{\widehat{n}}-1\right)\ \right| + \sum_{i,j}\left|\ \frac{\Delta_{i,j}}{\widehat{n}}\ \right| \\ &= \left|\ \frac{n}{\widehat{n}}-1\ \right|\cdot\sum_{i,j}p(i,j) + \sum_{i,j}\left|\ \frac{\Delta_{i,j}}{\widehat{n}}\ \right| \\ &= \left|\ \frac{n}{\widehat{n}}-1\ \right| + \sum_{i,j}\left|\ \frac{\Delta_{i,j}}{\widehat{n}}\ \right| . \end{aligned} \tag{62}$$

where the final equality holds because $\sum_{i,j} p(i,j) = 1$.

The following lemma gives a uniform upper bound on $|\Delta_{i,j}|$ that will translate into bounds on both terms of (62). Note that the lemma is stated and proven with respect to the fixed parameters (namely ϵ) already stated in this section.

Lemma D.7. *For any $(i,j) \in [c\ell]\times[\ell]$ and for any number $x > 0$,*

$$\Pr(|\Delta_{i,j}| > x) \le \exp(-(\epsilon/2)\cdot x) .$$

Proof. Recall $\Delta_{i,j}$ is a doubly-geometric random variable with parameter $e^{(-\epsilon/2)}$ centered at 0 and truncated at $n-(p(i,j)\cdot n)$ and $-(p(i,j)\cdot n)$. Let Δ denote a *non-truncated* doubly-geometric random variable with parameter $e^{(-\epsilon/2)}$ centered at 0. It is easy to see that for any number $x > 0$,

$$\Pr(|\Delta_{i,j}| > x) \le \Pr(|\Delta| > x) .$$

Then using the CDF of doubly-geometric random variables, we have

$$\Pr(|\Delta| > x) = 1 - \left(\frac{1-\exp(-\epsilon/2)}{1+\exp(-\epsilon/2)}\right)\cdot\left(\sum_{-x}^{x}\exp(-(\epsilon/2)\cdot|x|)\right)$$

The doubly-geometric distribution with parameter $e^{(-\epsilon/2)}$ is a discrete approximation of a continuous Laplace distribution with parameter $2/\epsilon$. It is known that for any $x > 0$

$$\Pr(|\mathrm{Lap}(2/\epsilon)| > x) = \exp(-(\epsilon/2)\cdot x) ,$$

and that

$$1 - \left(\frac{1-\exp(-\epsilon/2)}{1+\exp(-\epsilon/2)}\right)\cdot\left(\sum_{-x}^{x}\exp(-(\epsilon/2)\cdot|x|)\right) \le \exp(-(\epsilon/2)\cdot x) .$$

Thus the tail of every $\Delta_{i,j}$ r.v. with parameter $\exp(-(\epsilon/2))$ is dominated by the tail of the non-truncated Δ r.v. with the same parameter, which is dominated by the tail of a continuous Laplace r.v. with parameter $2/\epsilon$, completing the proof. □

Using this tail bound (which observe applies uniformly to all (i,j)), we have the following corollary.

Corollary D.8. *For every (i,j) and for any $d > 0$,*

$$\Pr(|\Delta_{i,j}| > d\ln n) \le \exp(-(\epsilon/2)d\ln n) = n^{-((\epsilon/2)d)} .$$

By a union bound over all $c\ell^2$ entries (i, j), it follows that for all $d > 0$,

$$\sum_{i,j} |\Delta_{i,j}| \;\leq\; c\ell^2 \cdot (d \ln n) \tag{63}$$

with probability all but $(c\ell^2)/(n^{-((\epsilon/2)d)})$. Because $\widehat{n} = n + \sum_{i,j} \Delta_{i,j}$ (from expression (61)), then with this same probability:

$$\widehat{n} \leq n + \sum_{i,j} |\Delta_{i,j}| \;\leq\; n + c\ell^2 d \ln n \tag{64}$$

$$\text{and} \quad \widehat{n} \geq n - \sum_{i,j} |\Delta_{i,j}| \;\geq\; n - c\ell^2 d \ln n \;. \tag{65}$$

Using these bounds on $\hat{n}$ and $\sum_{i,j} \Delta_{i,j}$, we state and prove the following two bounds on the terms from expression (62).

Lemma D.9. *For sufficiently large n, and assuming $B(n) = O(n^\alpha)$ for $\alpha \in (0, 0.5)$, there exists a constant $u_1 > 0$ such that*

$$\left| \frac{n}{\widehat{n}} - 1 \right| \;\leq\; O\left(\frac{c}{\epsilon n^{u_1}}\right)$$

with probability at least $1 - O(n^{-3})$.

Proof. First, suppose that $(n/\widehat{n}) \geq 1$. Then $|(n/\widehat{n}) - 1|$ is maximized when $\widehat{n}$ is small. Using the lower bound on $\widehat{n}$ from (65), it follows that

$$\left| \frac{n}{\widehat{n}} - 1 \right| \;=\; \left| \frac{n - \widehat{n}}{\widehat{n}} \right| \;\leq\; \frac{c\ell^2 d \ln n}{n - c\ell^2 d \ln n}$$

with probability all but $n^{-((\epsilon/2)d)}$. By the assumption that $\alpha < 0.5$, it follows that $\ell^2 \leq (B(n))^2 \leq O(n^{2\alpha}) = o(n)$. Setting $d := 8/\epsilon$, it follows that $c\ell^2 d \ln n \leq O((c/\epsilon) n^{1+t_1})$ for some constant $t > 0$, and that $n - c\ell^2 d \ln n = \Omega(n)$, which means that

$$\left| \frac{n}{\widehat{n}} - 1 \right| \;\leq\; \frac{c\ell^2 d \ln n}{n - c\ell^2 d \ln n} \;\leq\; O\left(\frac{c}{\epsilon n^t}\right)$$

with probability all but $(c\ell^2)/n^{-(\epsilon/2)d} = O(cn^{2\alpha}/n^{-4}) = O(n^{-3})$.

On the other hand, suppose $(n/\widehat{n}) < 1$. Then $|1 - (n/\widehat{n})|$ is maximized when $\widehat{n}$ is large, and using the upper bound on $\widehat{n}$ from (64), it follows that

$$\left| 1 - \frac{n}{\widehat{n}} \right| \;=\; \left| \frac{\widehat{n} - n}{\widehat{n}} \right| \;\leq\; \frac{c\ell^2 d \ln n}{n + c\ell^2 d \ln n}$$

with probability all but $n^{-((\epsilon/2)d)}$. By similar arguments as in the first case and again setting $d := 8/\epsilon$, it then follows that for some constant $v > 0$,

$$\left| 1 - \frac{n}{\widehat{n}} \right| \;\leq\; \frac{c\ell^2 d \ln n}{n + c\ell^2 d \ln n} \;\leq\; O\left(\frac{c}{\epsilon n^v}\right)$$

with probability all but n^{-3}.

Setting $u_1 := \min\{t, v\}$ then concludes the proof. □

Lemma D.9 gives a bound on the first term from expression 62. We now prove a bound on the second term.

Lemma D.10. *For sufficiently large n, and assuming $B(n) = O(n^\alpha)$ for $\alpha \in (0, 0.5)$, there exists a constant $u_2 > 0$ such that*

$$\sum_{i,j} \left| \frac{\Delta_{i,j}}{\widehat{n}} \right| \;\leq\; O\left(\frac{c}{\epsilon n^{u_2}}\right)$$

with probability all but n^{-3}.

Proof. As in the proof of Lemma D.9, set $d := 8/\epsilon$. Again using the fact that $\ell^2 \leq B(n)^2 \leq O(n^{2\alpha}) = o(n)$ by the assumption that $\alpha < 0.5$, observe from (65) that $\hat{n} > 0$ with probability all but $O(n^{-3})$ for sufficiently large n. So with this same probability we can write

$$\sum_{i,j} \left| \frac{\Delta_{i,j}}{\widehat{n}} \right| = \frac{\sum_{i,j} |\Delta_{i,j}|}{\hat{n}} \leq \frac{c\ell^2(8/\epsilon)\ln n}{n - c\ell^2(8/\epsilon)\ln n} .$$

Here, since $\ell^2 \ln n \leq n^{2\alpha} \ln n$, it follows that the denominator of the final expression is $\Omega(n)$, and the numerator is $O((c/\epsilon)n^{1+u_2})$ for some $u_2 > 0$. Thus with probability all but $O(n^{-3})$, it follows that

$$\sum_{i,j} \left| \frac{\Delta_{i,j}}{\widehat{n}} \right| \leq \frac{c\ell^2(8/\epsilon)\ln n}{n - c\ell^2(8/\epsilon)\ln n} \leq O\left(\frac{c}{\epsilon n^{u_2}}\right) .$$ □

Together, Lemmas D.9 and D.10 give the following probabilistic bound on the total variation distance between $\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma}$ and $\mathbf{P}_{D,\Gamma}$ from expression (62), and thus a proof of Lemma D.5.

Proof (of Lemma D.5). By expressions (58) and (62), and using Lemmas D.9 and D.10, it follows that for sufficiently large n, there exist constants $u_1, u_2 > 0$ such that for any $G \in \mathcal{R}(L, c, k, \ell)$,

$$D_{TV}(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma,G}, \mathbf{P}_{D,\Gamma,G}) \leq D_{TV}(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma}, \mathbf{P}_{D,\Gamma}) \leq O\left(\frac{c}{\epsilon n^{u_1}}\right) + O\left(\frac{c}{\epsilon n^{u_2}}\right) \leq O\left(\frac{c}{\epsilon n^{u_3}}\right)$$

with probability at least $1 - O(n^{-3})$ for sufficiently large n, where $u_3 := \min\{u_1, u_2\} > 0$.

Then by Lemma D.6, it follows that

$$\left| I\left(\widehat{\mathbf{P}}^{\epsilon}_{D,\Gamma,G}\right) - I\left(\mathbf{P}_{D,\Gamma,G}\right) \right| \leq O\left(\frac{c}{\epsilon n^{u_3}} \cdot \log_2(n^{(\alpha+u_3)})\right) \leq O\left(\frac{c}{\epsilon n^{u}}\right)$$

with probability at least $1 - O(n^{-3})$ for some $u_3 > u > 0$ when n is sufficiently large. □

E. Experimental Results Appendix

In this section, we provide more details on the methodologies used for our experimental results from Section 5.

E.1. Implementation Details

As mentioned in Section 2, the MICe statistic in Definition 2.3 (and the analogous definitions of MICr and its private variants) varies slightly from the definition of MICe from Reshef et al. (2016), specifically when defining the maximization space of individual entries in the equicharacteristic matrix[3]. In particular, for $k \leq \ell$, when $\ell > \sqrt{B}$, Reshef et al. (2016) set the master row partition used in the optimization as a mass equipartition of size $c(B/\ell)$. In contrast, Definition 2.3 sets the master row partition as a mass equipartition of size $c \cdot \ell$, even when $\ell > \sqrt{B}$ (note that the definitions are identical in the case when $\ell \leq \sqrt{B}$).

The advantage of using these smaller master partitions is computational: the MICe variant of Reshef et al. (2016) can be computed in $O(c^2B^{2.5})$ time (Appendix H.1, Reshef et al. (2016)), in contrast with the $O(c^2B^4)$ running time from Theorem A.1 using Definition 2.3. On the other hand, for $B := B(n) = O(n^\alpha)$ where $\alpha \in (0, 0.5)$, the consistency of this variant can only be shown to hold when the output of the statistic is the maximum (k, ℓ) entry of the equicharacteristic matrix where $k \cdot \ell \leq B$ *and* k and ℓ are individually at most $\sqrt{B}$. Note however that although consistency for this variant is only shown to hold when imposing this additional constraint on k and ℓ, the statistic is still nonetheless both defined and can be computed (i.e., as in the implementation of Albanese et al. (2018)) as a maximization over *all* equicharacteristic matrix entries where $k \cdot \ell \leq B$.

The MICr statistic from Definition 4.1 (and its private variants) can be adjusted in an analogous way to reduce the size of the master row range equipartitions when $\sqrt{B} < \ell \leq B/2$. Again, the benefit is faster computation at the expense of a slightly different consistency guarantee (analogous to that of the MICe variant described above).

[3]Reshef et al. (2016) refer to this as the "clumped" variant of MICe, but to avoid confusion we refer to this as just MICe.

Primarily for computational considerations, our experiments in this work use implementations of MICe and MICr (and their private variants) that follow the original definition of Reshef et al. (2016) (i.e., use smaller master partition sizes for $\ell > \sqrt{B}$). While these variants hold slightly different consistency properties than their counterparts defined in Sections 2, 3, and 4, in practice the outputs between pairs of corresponding variants are similar. Note that an end-user with a fixed computational budget could use either set of definitions and tune the B and c parameters accordingly to meet their budget.

Moreover, we remark that the ϵ-DP guarantees of the private MICe and MICr mechanisms introduced in this work apply to either regime of master grid size settings.

In our experiments, to compute MICe we used the MINEPY library implementation from Albanese et al. (2018), and to compute MICr and its private variants, we developed an implementation available at `https://github.com/jlazarsfeld/dp-mic`.

E.2. Synthetic Data Experiments

E.2.1. Functional Relationships Used to Generate Distributions

As described in Section 5, we evaluated our private mechanisms on distributions based on a set of 21 functional relationships originally defined by Reshef et al. (2011). We list these functions here:

- F1: $f(x) = 0.2 \cdot \sin(12x - 6) + 1.1(x - 1) + 1$
- F2: $f(x) = 0.15 \cdot \sin(11x \cdot \pi) + (x + 0.05)$
- F3: $f(x) = 0.1 \cdot \sin(48x) + 2(x - 0.05)$
- F4: $f(x) = 0.2 \cdot \sin(48x) + 2(x - 0.05)$
- F5: $f(x) = 0.4 \cdot \cos(7x \cdot \pi) + 0.5$
- F6: $f(x) = 0.4 \cdot \cos(14x \cdot \pi) + 0.5$
- F7: $f(x) = 10 \cdot (x - 0.6)^3 + 2 \cdot x^2 + (1.5 - 3x)$
- F8: $f(x) = 4 \cdot (10 \cdot (x - 0.6)^3 + 2 \cdot x^2 + (1.5 - 3x)) - 1.4$
- F9: if $x \leq 0.99$ then $f(x) = \frac{x}{99}$, else $f(x) = 99x - 98$
- F10: $f(x) = 2^x - 1$
- F11: $f(x) = 8^{(x-0.3)} - 1$
- F12: $f(x) = x$
- F13: $f(x) = 4 \cdot (x - 0.5)^2 + 0.1$
- F14: $f(x) = 0.4 \cdot \sin(9x \cdot \pi) + 0.5$
- F15: $f(x) = 0.4 \cdot \sin(8x \cdot \pi) + 0.5$
- F16: $f(x) = 0.4 \cdot \sin(16x \cdot \pi) + 0.5$
- F17: if $x < 0.491$ then $f(x) = 0.05$, else if $x > 0.509$ then $f(x) = 0.95$, else $f(x) = 50 \cdot (x - 0.5) + 0.5$
- F18: $f(x) = 0.4 \cdot \cos(5x \cdot \pi \cdot (1 + x)) + 0.5$
- F19: $f(x) = 0.4 \cdot \sin(6x \cdot \pi \cdot (1 + x)) + 0.5$
- F20: if $x \leq 0.0528$ then $f(x) = 18x$, else if $x \geq 0.1$ then $f(x) = -1 \cdot \frac{x}{9} + \frac{1}{9}$, else $f(x) = -18x + 1.9$
- F21: if $x \leq 0.0051$ then $f(x) = 190x$, else if $x \geq 0.01$ then $f(x) = -1 \cdot \frac{x}{99} + \frac{1}{99}$, else $f(x) = -198x + 1.99$

For each function, we generated 9 joint distributions by placing $k = 100$ "generating" points on the function graph at evenly spaced intervals. We considered a bivariate Gaussian distribution centered at each point with $\rho = 0$ and identical variances. The magnitude of the variance was set according to the desired R^2 value of the resulting noisy functional relationship using the method described in Reshef et al. (2018).

E.2.2. Approximately Computing MIC*

For a joint distribution $\Pi = (X, Y)$, we used the method described in Section 3.5 of Reshef et al. (2016) to approximately compute MIC*(Π). This involves maximizing the mutual information of $I(X|_Q, Y|_P)$ for increasingly large discrete partitions P of size k chosen from a dense master mass equipartition Γ, and where Q is a dense, fixed-sized master mass equipartition . For this, we set the size of Γ to 260 and the size of Q to 360, and we found the computation insensitive to increases in master grid sizes beyond this point.

E.2.3. Tuning Parameters for MICr-Lap and MICr-Geom

To determine optimal B and c parameter settings for MICr-Lap and MICr-Geom at a particular value of n and ϵ, we defined the following objective function WSUM using the set of 189 distributions in $\mathcal{Q}$.

First, we sorted $\mathcal{Q}$ in increasing order of MIC* value (using the method described in the previous subsection). For the i'th distribution Π_i in sorted order, we defined a weight w_i by taking the length of the interval between the midpoint of MIC*(Π_{i-1}) and MIC*(Π_i) and the midpoint of MIC*(Π_i) and MIC*(Π_{i+1}). For the distributions with smallest and largest MIC* values, we used 0 and 1 as left and right interval endpoints respectively. Then for a fixed mechanism, B, c, n, and ϵ, we took the weighted sum of the mechanism's average unsigned error on each distribution wrt MIC* using the w_i's as weights.

By minimizing this function wrt the B and c parameters for each mechanism and (n, ϵ, B, c) combination, our goal was to determine parameter settings that ensured equal levels of error across the entire spectrum of low-correlation (low MIC*) and high-correlation (high MIC*) distributions.

	MICr-Geom		MICr-Lap	
	$\epsilon = 1.0$	$\epsilon = 0.1$	$\epsilon = 1.0$	$\epsilon = 0.1$
$n = 25$	(2, 12)	(2, 6)	(5, 8)	(5, 6)
$n = 250$	(1, 40)	(2, 10)	(5, 40)	(5, 40)
$n = 500$	(1, 40)	(2, 20)	(5, 60)	(5, 80)
$n = 1000$	(1, 60)	(2, 40)	(5, 80)	(5, 100)
$n = 5000$	(1, 150)	(1, 40)	(5, 150)	(5, 125)
$n = 10000$	(1, 150)	(1, 80)	(5, 150)	(5, 150)

Table 2. Each entry is the optimal (c, B) settings for the corresponding mechanism and (n, ϵ) that was found by minimizing a weighted sum of the mechanism's absolute error (wrt MIC*) over all $\Pi \in \mathcal{Q}$.

Table 2 shows these optimized parameters for the MICr-Lap and MICr-Geom mechanisms. For both mechanisms and both ϵ values, the optimal B values are generally increasing with n, which aligns with the intuition that both mechanisms' outputs converge toward MIC* with larger n. Notice also that for ϵ=0.1 and small n, the optimal c value of MICr-Geom is 2. Although the error of MICr-Geom wrt to MICr decreases with c, we expect the error of MICr wrt MIC* to decrease with *larger* c and B. For small n and ϵ, this leads to an optimization tradeoff that is more pronounced.

E.2.4. Bias/Variance Analysis for $\epsilon = 0.1$

Figure 4 shows the boxplots of the bias and variance of each private mechanism over the set of all distributions in $\mathcal{Q}$ as n varies and $\epsilon = 0.1$. These are the analogous plots for Figure 2 in Section 5 for $\epsilon = 1$. Again notice that as n grows, the IQR of bias tends to decrease for the MICr-Lap and MICr-Geom mechanisms, but this is less pronounced for MICr-Geom compared to the $\epsilon = 1$ case. Here, the bias/variance tradeoff between MICr-Lap (lower bias, higher variance) and MICr-Geom (higher bias, lower variance) is much more apparent, especially at smllaller values of n.

E.3. Real Data Experiments

Table 3 shows the median bias and variance for each mechanism over all datasets in each collection for $\epsilon = 0.1$ (analogous to Table 1 from Section 5).

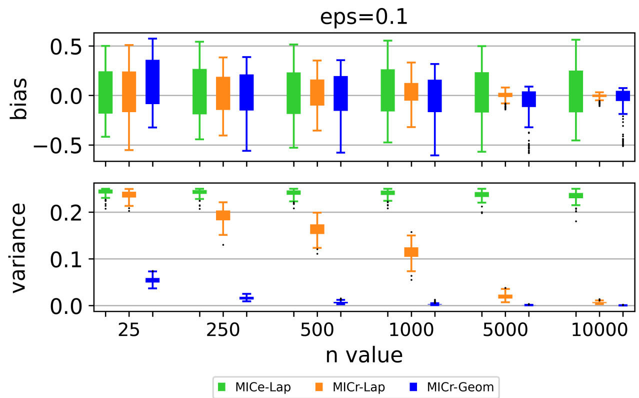

Figure 4. Boxplots of bias (top) and variance (bottom) of each private mechanism (over 50 iterations) over all distributions in $\mathcal{Q}$ for ϵ=0.1 and varying n.

	MICe-Lap	MICr-Lap	MICr-Geom
Spellman23	0.20 (0.24)	0.19 (0.24)	0.31 (0.05)
Baseball	0.36 (0.24)	0.25 (0.18)	0.33 (0.01)
Spellman4381	0.41 (0.24)	0.03 (0.01)	0.14 (8e-4)

Table 3. The median bias (average signed error wrt MICe over 100 runs) and median variance (over 100 iterations) of each private mechanism across all datasets of each collection for ϵ=0.1.

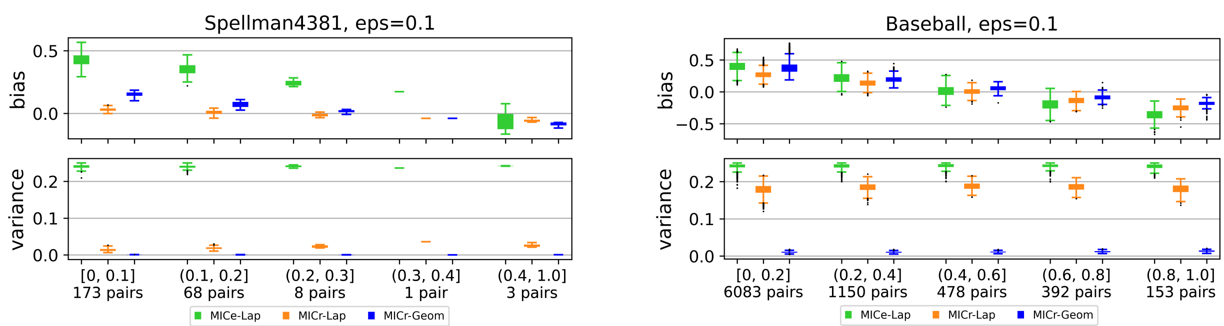

Figure 5. Bias and variance boxplots for each mechanism over datasets (pairs) in the Spellman4381 collection (left) and Baseball collection (right) binned by non-private MICe score for ϵ=0.1.

Figure 5 shows the bias and variance of each mechanism for the Spellman4381 and Baseball collections at $\epsilon = 0.1$ in the binned setting (analogous to Figure 3 in Section 5). Again notice the bias/variance tradeoff between the MICr-Lap and MICr-Geom mechanisms, which is especially apparent at this level of ϵ in the smaller ($n = 337$) Baseball datasets.