$$\text{for } \theta \in [0,1], \text{ if } \theta \le \max_{(\underline{r},\hat{\underline{r}})\in\mathcal{N}} \max_{y\in\mathcal{Y}} \frac{\mathbb{W}_M(y|\underline{r})}{\mathbb{W}_M(y|\hat{\underline{r}})} \le \frac{1}{\theta}, \qquad (1)$$

where $\mathcal{N}$ is the set of all (ordered) pairs of DBs that differ in a single record, and $\mathbb{W}_M(y|\underline{r})$ is the probability of M putting

This work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, NSF Grants CCF-1524312, NSF ECCS-1646449, NIH Grant 1U01CA198941-01 and USARO under Contract W911NF-15-1-0279.

this limit, and hence the minimum expected L_1-fidelity of a $\theta-$DP DSM in the limit of large DBs. Our solution brings to light connections between DP and *Ehrhart theory* [2].

Ehrhart theory concerns integer-point enumeration of polytopes. The counts of the number of integer points in the $t-$th dilation of a polytope - the *Ehrhart polynomial* of the polytope - and the associated generating function - the *Ehrhart series*

978-1-5386-4780-6/18/$31.00©2018 IEEE

of the polytope - are fundamental constructs in Ehrhart theory. As we describe below, these constructs will play a central role in characterizing the limit we seek.

Our crucial first step of visualizing the LP through a graph paves the way to developing these connections with discrete geometry. In particular, we relate the objective and constraints of the LP with the distance distribution of vertices in this graph. In the limit of large DBs, the distance distribution of this graph is given by the Ehrhart polynomial of a suitably defined convex polytope. Our solution has two parts - upper and lower bound. To characterize an upper bound on the limit we seek, we identify feasible solutions to the sequence of LP's, whose objective values, in the limit is given by a simple functional of the Ehrhart series of the above mentioned convex polytope. Sec. III-A provides a descriptive derivation of the upper bound and Sec. IV, the mathematical steps. We appeal to weak duality theorem for the lower bound. Note that every feasible solution to the dual of the above LP evaluates to a lower bound on the minimum expected fidelity. We therefore consider the sequence of dual LPs and identify a sequence of feasible solutions for the same. We prove that these feasible solutions evaluate to, in the limit, the same functional as obtained in the upper bound. This enable us conclude that the Ehrhart series of the above mentioned convex integral polytope yields the minimum expected L_1-fidelity of a $\theta-$DP DSM, thereby establishing a connection between objects of fundamental interest of the two disciplines/areas.

While DP [1] has been a subject of intense research (See [3] and references therein, [4], [5]), much of this is aimed at studying variants of the geometric/Laplacian mechanism, leaving the question of their optimality open. Hardt and Talwar [6] considered 'continuous extensions' of the (min-max) problem and developed novel lower bounding techniques based on geometric arguments. [6] and [7] are based on a clever use of the Markov inequality. Geng and Viswanath [8], [9] focus on noise-adding mechanisms and proved optimality of 'staircase mechanisms' for a general class of convex utility functions by appealing to functional analytic techniques. More recently, [10] developed lower bounds based on non-existence of certain fingerprinting codes. All these techniques have been developed for the minimax setting and as we discuss in Rem. 2, do not yield a lower bound for the problem studied herein.

II. Problem Statement

DB as input and outputs a DB. Since permutations are irrelevant, a DB is equivalently represented through its histogram. We therefore concern ourselves with designing a histogram sanitization mechanism (HSM). For a DB $\underline{r} \in \mathcal{R}^n$ and a record $a_k \in \mathcal{R}$, we let $\mathrm{h}(\underline{r})_k = \sum_{i=1}^{n} \mathbb{1}_{\{r_i = a_k\}}$ denote number of subjects with record a_k, and $\mathrm{h}(\underline{r}) := (\mathrm{h}(\underline{r})_1, \cdots, \mathrm{h}(\underline{r})_K)$ denote the histogram corresponding to DB $\underline{r} \in \mathcal{R}^n$. Let $\mathcal{H}^n := \{(h_1, \cdots, h_K) \in \mathbb{Z}^K : h_i \ge 0, \sum_{k=1}^{K} h_k = n\}$ denote the collection of histograms. When K is set to a particular value, we let $\mathcal{H}_K^n$ denote $\mathcal{H}^n$. In this article, we measure fidelity between a pair of histograms through its L_1-distance. We employ DP to quantify vulnerability to privacy breaches. A pair $\underline{r}, \hat{\underline{r}} \in \mathcal{R}^n$ of DBs are *neighboring* if $\underline{r}$ and $\hat{\underline{r}}$ differ in exactly one entry, or equivalently $|\mathrm{h}(\underline{r}) - \mathrm{h}(\hat{\underline{r}})|_1 = 2$.

Definition 1: A pair $\underline{h}, \hat{\underline{h}} \in \mathcal{H}^n$ is neighboring if $|\underline{h} - \hat{\underline{h}}|_1 = 2$. A HSM $M : \mathcal{H}^n \Rightarrow \mathcal{H}^n$ is $\theta-$DP $(0 < \theta < 1)$ if for every pair $\underline{h}, \hat{\underline{h}} \in \mathcal{H}^n$ of neighboring histograms and every histogram $\underline{g} \in \mathcal{H}^n$, we have $\theta\ \mathbb{W}_M(\underline{g}|\underline{h}) \le \mathbb{W}_M(\underline{g}|\hat{\underline{h}}) \le \theta^{-1}\ \mathbb{W}_M(\underline{g}|\underline{h})$.

We formulate the problem of characterizing the minimum *expected* fidelity of a $\theta-$DP HSM. Towards that end, we model a pmf on the space of DBs. For a record $a_k \in \mathcal{R}$, let $p_k > 0$ denote the probability that a subject's record is a_k. Moreover, the n records that make up the DB are IID with pmf $\underline{p} := (p_1, \cdots, p_K)$. The probability that the histogram of the randomly chosen DB $\underline{R} \in \mathcal{R}^n$ is

$$P\left(\mathrm{h}(\underline{R}) = \underline{h}\right) = \sum_{\underline{r}\in\mathcal{R}^n : \mathrm{h}(\underline{r}) = \underline{h}} P(\underline{R} = \underline{r}) = \sum_{\underline{r}\in\mathcal{R}^n : \mathrm{h}(\underline{r}) = \underline{h}} \underline{p}^{\mathrm{h}(\underline{r})} = \binom{n}{\underline{h}} \underline{p}^{\underline{h}}, \quad (2)$$

where here and henceforth, we let $\underline{p}^{\underline{h}} : = \prod_{k=1}^{K} p_k^{h_k}$. (2) follows from the fact that the number of DBs whose histogram is $\underline{h} \in \mathcal{H}^n$ is the multinomial coefficient $\binom{n}{\underline{h}} := \binom{n}{h_1 \cdots h_K}$. In passing, we note that the multinomial pmf (2) with a generic pmf $\underline{p}$ on the set $\mathcal{R}$, is indeed the most generic pmf on the space of histograms. Throughout, we make no assumption on $\underline{p}$ resulting in a generic study. We now formulate our problem.

Given a privacy budget $\theta \in (0, 1)$, our goal is to characterize $D_K^*(\theta) := \lim_{n\to\infty} D_*^n(\theta)$, where,

$$D_*^n(\theta) := \min_{\mathbb{W}(\cdot|\cdot)} \sum_{\underline{h}\in\mathcal{H}^n} \sum_{\underline{g}\in\mathcal{H}^n} \binom{n}{\underline{h}} \underline{p}^{\underline{h}} \mathbb{W}(\underline{g}|\underline{h}) |\underline{h} - \underline{g}|_1, \text{ subject } (3)$$

$$\sum_{\underline{g}\in\mathcal{H}^n} \mathbb{W}(\underline{g}|\underline{h}) \overset{(4a)}{=} 1\ \forall\ \underline{h} \in \mathcal{H}^n, \quad \mathbb{W}(\underline{g}|\underline{h}) - \theta\ \mathbb{W}(\underline{g}|\hat{\underline{h}}) \overset{(4b)}{\ge} 0 \quad (4)$$