Preserving Privacy and Fidelity via Ehrhart Theory

Arun Padakandla
Purdue University

P. R. Kumar
Texas A & M University

Wojciech Szpankowski
Purdue University

***Abstract*—We consider the problem of designing a database sanitization mechanism (DSM) that minimizes, in the expected sense, the $\mathbb{L}_1-$distortion between the histograms of original and sanitized databases, while being $\theta-$differentially private (DP). The expected $\mathbb{L}_1-$distortion of a corresponding optimal $\theta-$DP DSM provides for an important utility-privacy trade-off. This problem reduces to a prohibitively complex linear program (LP). Using tools from Ehrhart theory, analytic combinatorics and LP theory, we solve this problem and thereby provide a simple closed form computable expression characterizing this trade-off.**

I. Introduction

Fine-grained and high-dimensional data gathered from subjects regarding their preferences, characteristics is being increasingly mined for socio, economic and scientific benefits. Correlations, discovered via such querying, among attributes previously thought unrelated, could lead to significant scientific breakthroughs and/or economic benefits. Consequently, databases (DBs) are subject to unrestrained querying, and are also being traded among corporations/agencies to facilitate informed policy making. Trading of DBs containing private information, amongst *untrusted* agencies, and their *unrestrained* querying results in catastrophic loss of subject privacy. It is imperative that we adopt an architecture for DB handling, trading and querying that simultaneously (i) permits unrestrained querying, (ii) provides useful/accurate responses, while (iii) being secure against privacy breaches. An information-theoretic study of the privacy-utility trade-off of such an architecture is the overarching goal of our work.

Motivated by these considerations, we focus on an architecture wherein the original DB is sanitized via a provably secure mechanism. The original DB is destroyed and further querying/trading is performed of the sanitized DB (Fig. 1). How do we sanitize a DB and secondly, what is the performance/utility of this architecture? The need to quantify vulnerability of a DB sanitizing mechanism (DSM), and more importantly provide provable upper bounds on the same, leads us to differential privacy (DP) [1]. DP models a DSM, and more generally a query-response mechanism, as a randomized algorithm and quantifies the vulnerability of the latter via its sensitivity to individual records. Specifically, a mechanism M is $\theta-$DP

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

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 out $y \in \mathcal{Y}$ as the response when input with DB $\underline{r}$. The reader will note that larger values of θ correspond to less vulnerable mechanisms, and the increased robustness is achieved at the cost of reduced accuracy of the query response. Key properties of DP - composition and post-processing - justify our choice of a DP DSM. In particular, the latter states that querying a DB sanitized via $\theta-$DP DSM is, irrespective of the query and the querying mechanism, at least as robust as a $\theta-$DP mechanism. In other words, sanitizing a DB via a DP mechanism provides for an impermeable firewall against privacy breaches.

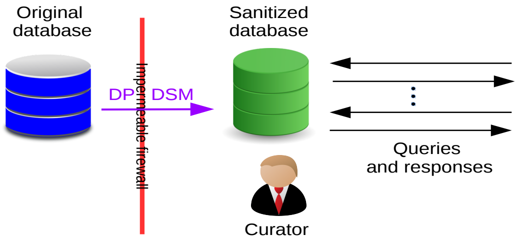

Fig. 1.

The utility/performance of this architecture (Fig. 1) is governed by accuracy of the responses obtained by querying the sanitized DB. Our focus being statistical queries, we consider a fidelity measure on the space of histograms and quantify (loss of) utility/performance by a fidelity between the histograms of the original and sanitized DBs. This leads us to our problem, posed in a (conventional) information-theoretic framework. Given a privacy budget, in terms of the DP parameter $\theta \in (0,1)$, what is the minimum *expected* fidelity of an optimal $\theta-$DP DSM, in the *asymptotic* regime of *large* DBs. Section II contains a mathematical formulation.

Identifying an optimal mechanism reduces to a prohibitively complex optimization problem (Rem. 1) with an exponential number of linear constraints. Seeking to identify structure of the optimal mechanism, we consider the $\mathbb{L}_1$ or TV divergence measure, in which case the objective function is linear, thereby resulting in a linear program (LP). We are thus posed with the task of identifying the limit of solutions to a sequence of LPs, each of which, is subject to exponentially many constraints. One of our main contributions is a precise characterization of this limit, and hence the minimum expected $\mathbb{L}_1-$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*

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.

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 $\mathbb{L}_1-$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

We let $M : \mathcal{A} \Rightarrow \mathcal{B}$ denote a mechanism (randomized algorithm) with input set $\mathcal{A}$ and output set $\mathcal{B}$. When input $a \in \mathcal{A}$, the mechanism M produces output $b \in \mathcal{B}$ with probability $\mathbb{W}_M(b|a)$.[1] $\mathbb{W}_M : \mathcal{A} \Rightarrow \mathcal{B}$ also refers to mechanism $M : \mathcal{A} \Rightarrow \mathcal{B}$. Consider a DB with n *subjects*. Each subject is identified with a *record* which stores his/her data. We let $\mathcal{R} = \{a_1, \cdots, a_K\}$ denote the set of possible records. K can be arbitrary, but will remain fixed throughout our study. We let $\underline{r} := (r_1, \cdots, r_n) \in \mathcal{R}^n$ denote a generic DB with n records.

Our focus in this article is to design a DB sanitization mechanism (DSM), i.e., a randomized algorithm that takes a 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 $\mathcal{H}^n := \{(h_1, \cdots, h_K) \in \mathbb{Z}^K : h_i \geq 0, \sum_{k=1}^K h_k = n\}$ 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 $\mathbb{L}_1-$distance. We employ DP to quantify vulnerability to privacy breaches. A pair $\underline{r}, \underline{\hat{r}} \in \mathcal{R}^n$ of DBs are *neighboring* if $\underline{r}$ and $\underline{\hat{r}}$ differ in exactly one entry, or equivalently $|\mathrm{h}(\underline{r}) - \mathrm{h}(\underline{\hat{r}})|_1 = 2$.

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

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,

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

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

$$\forall (\underline{h}, \underline{\hat{h}}) \in \mathcal{H}^n \times \mathcal{H}^n \text{ satisfying } |\underline{h} - \underline{\hat{h}}|_1 = 2 \text{ and } \forall \ \underline{g} \in \mathcal{H}^n; \quad (5)$$

$$\text{and } \mathbb{W}(\underline{g}|\underline{h}) \geq 0 \text{ for all } (\underline{g}, \underline{h}) \in \mathcal{H}^n \times \mathcal{H}^n. \quad (6)$$

Remark 1: It can be verified that the optimization problem (3) has $(n+1)^{2K}$ decision variables and $\mathcal{O}(K^2|\mathcal{H}^n|^2) = \mathcal{O}(K^2(n+1)^{2(K-1)})$ constraints of the form (4)(b). Owing to its complexity, we are unaware of its solution even for $K = 2$.

Remark 2: One may replace the $|\cdot|_1$ distance with a generic fidelity measure $\mathcal{F}(\cdot, \cdot)$ on $\mathcal{H}^n$ and recover problem formulations studied in [4], [11], among others. In particular Ghosh, Roughgarden and Sundararajan [11] study the $K = 2$ case for a fairly generic fidelity measure, and prove structural properties of an optimal mechanism. While these hold for

[1] In the CS literature, such as in [1], $\mathbb{W}(b|a)$ is denoted $M(b)_a$.

each n, they do not pin down an optimal mechanism, leaving $D^*_2(\theta)$ unknown. On the one hand, [6] studies a min-max problem setting. Secondly, their continuous extension results in a larger constraint set, lending the lower bounds developed therein invalid for the original discrete problem setting.

III. PRECISE CHARACTERIZATION OF $D^*_K(\theta)$

One of our main contributions is a simple computable expression for $D^*_K(\theta)$. Thm. 1 provides one such expression - a *hyper-geometric series*. The latter encapsulates the entire information from a power series. In Sec. III-A, we identify this power series and elaborate on how and why it is related to $D^*_K(\theta)$.

Theorem 1: (i) When $K = 2$, the limit $D^*_2(\theta) = \lim_{n\to\infty} D^n_*(\theta) = \frac{4\theta}{1-\theta^2}$. (ii) In general, the minimum expected $\mathbb{L}_1-$fidelity of a $\theta-$DP HSM is given by

$$D^*_K(\theta) = 2\theta\left\{\tfrac{K-1}{1-\theta} + \tfrac{S'_{K-1}(\theta)}{S_{K-1}(\theta)}\right\}, \text{ where}$$
$$S'_{K-1}(\theta) := \tfrac{dS_{K-1}(\theta)}{d\theta},\ S_{K-1}(\theta) = \sum_{j=0}^{K-1}\theta^j\left[\begin{pmatrix}K-1\\ j\end{pmatrix}\right]^2 \quad (7)$$

with $S'_{K-1}(\theta) := \frac{d}{d\theta}S_{K-1}(\theta)$. An optimal HSM is obtained as a truncation of a geometric mechanism $\mathbb{W}^*(\underline{g}|\underline{h}) = \frac{1}{\mathscr{E}_{\mathcal{P},f}(\theta)}\theta^{\frac{|\underline{g}-\underline{h}|_1}{2}}$, where $\mathscr{E}_{\mathcal{P},f}(\theta)$ is the normalizing constant related to the Ehrhart series shown in (8).

Below, we express $D^*_K(\theta)$ in terms of another important construct in analysis - the *Legendre polynomial*. We note that $S_{K-1}(\theta) = (1-\theta)^{K-1}L_{K-1}(\frac{1+\theta}{1-\theta})$ [12, Prob. 85], where $L_n(x) := \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n$ is the Legendre polynomial of degree n. This leads to the following.

Corollary 1: We have

$$D^*_K(\theta) = K\left\{\frac{1+4\theta-\theta^2}{1-\theta} + \frac{L_K(y)}{L_{K-1}(y)}\right\}, \text{ where } y = \frac{1+\theta}{1-\theta}.$$

Proof: Substitute $S_{K-1}(\theta) = (1-\theta)^{K-1}L_{K-1}(y)$ and use the defn of $L_n(x) := \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n$ [12, Prob. 86]. ■

Characterization for $D^*_K(\theta)$ provided in (7) hints at deeper connections between Ehrhart theory and DP. We prove that

$$D^*_K(\theta) = \tfrac{2\theta}{\mathscr{E}_{\mathcal{P},f}(\theta)}\tfrac{d\mathscr{E}_{\mathcal{P},f}(\theta)}{d\theta}, \text{ where}$$
$$\mathscr{E}_{\mathcal{P},f}(\theta) \overset{(b)}{=} (1-\theta)\mathrm{Ehr}_{\mathcal{P}}(\theta) \overset{(c)}{=} 1 + \textstyle\sum_{\mathrm{d}=1}^{\infty}\mathrm{N_d}\theta^{\mathrm{d}}, \quad (8)$$

N_d is the number of *integer* points on the *face* of the convex polytope

$$\mathcal{P}_d = \{(x_1,\cdots,x_K)\in\mathbb{R}^K : \sum_{k=1}^{K}x_k = 0, \sum_{k=1}^{K}|x_k| \leq 2d\}. \quad (9)$$

and $\mathrm{Ehr}_{\mathcal{P}}(\theta)$ is the Ehrhart series of this polytope. (7) is a culmination of leveraging tools from discrete geometry, combinatorics (Ehrhart theory) and analytic methods. In the following, we introduce these tools. We refer the reader to [2] for a beautiful exposition of Ehrhart theory.

A. Excursions into Ehrhart theory and Linear programming

$D^*_K(\theta)$ is the limit of solutions to a sequence of LPs (3). We begin with the privacy-constraint (PC) graph which greatly aids in visualizing the same and naturally leads us into Ehrhart theory. Consider a graph $G = (V, E)$ with vertex set $V = \mathcal{H}^n$ and an edge set $E = \left\{(\underline{h}, \hat{\underline{h}}) \in \mathcal{H}^n \times \mathcal{H}^n : |\underline{h} - \hat{\underline{h}}|_1 = 2\right\}$. For every vertex $\underline{h} \in V$, visualize the sub-collection $(\mathbb{W}(\underline{g}|\underline{h}) : \underline{g} \in \mathcal{H}^n)$ of decision variables as a function of V, i.e., as values lying on V, corresponding to $\underline{h} \in V$ (See Fig. 2). The values $(\mathbb{W}(\underline{g}|\underline{h}) : \underline{g} \in \mathcal{H}^n)$ and $(\mathbb{W}(\underline{g}|\hat{\underline{h}}) : \underline{g} \in \mathcal{H}^n)$ corresponding to two neighboring vertices $\underline{h}, \hat{\underline{h}}$ have to be within θ and $\frac{1}{\theta}$ of each other everywhere, i.e., at every $\underline{g}$ (See Fig. 2). In addition, the values corresponding to any node must be non-negative and sum to 1. The PC graph also provides a visualization of the objective function. $|\underline{g} - \underline{h}|_1$ is exactly twice $d_G(\underline{g}, \underline{h})$, where the latter denotes the length of a shortest path from $\underline{g} \in V$ to $\underline{h} \in V$. (Proof in [13, Appendix 7]). Two useful consequences follow. Firstly, the values corresponding to a node, say $\underline{h}$, that are equidistant from $\underline{h}$ are multiplied by identical coefficients in the objective function. Formally, $\binom{n}{\underline{h}}|\tilde{\underline{g}} - \underline{h}|_1 = \binom{n}{\underline{h}}|\underline{g} - \underline{h}|_1$ iff $d_G(\tilde{\underline{g}}, \underline{h}) = d_G(\underline{g}, \underline{h})$. Secondly, coefficients associated with the values increase with their distance from $\underline{h}$. Formally, if $d_G(\tilde{\underline{g}}, \underline{h}) > d_G(\underline{g}, \underline{h})$, then $\binom{n}{\underline{h}}|\tilde{\underline{g}} - \underline{h}|_1 > \binom{n}{\underline{h}}|\underline{g} - \underline{h}|_1$. These observations lets us restate our objective function (3) as

$$D^n_{\mathcal{H}}(\mathbb{W}) \overset{(a)}{=} \sum_{\underline{h}\in\mathcal{H}^n}\sum_{d=1}^{n}\sum_{\underline{g}\in\mathcal{H}^n:|\underline{g}-\underline{h}|_1=2d}\binom{n}{\underline{h}}\underline{p}^{\underline{h}}\mathbb{W}(\underline{g}|\underline{h})2d \quad (10)$$

$$= \sum_{\underline{h}\in\mathcal{H}^n}\binom{n}{\underline{h}}\underline{p}^{\underline{h}}\sum_{d=1}^{n}2d\sum_{\underline{g}\in\mathcal{H}^n:d_G(\underline{g},\underline{h})=d}\mathbb{W}(\underline{g}|\underline{h}). \quad (11)$$

In arriving at (10)(a), we used the fact that for any $\underline{g}, \underline{h} \in \mathcal{H}^n$, we have $|\underline{g} - \underline{h}|_1$ is an even integer and at most $2n$. Proof is in [13, Appendix 7]. Consider a HSM $M : \mathcal{H}^n \Rightarrow \mathcal{H}^n$ for which $\mathbb{W}(\underline{g}|\underline{h}) = f(\underline{h}, |\underline{g} - \underline{h}|_1)$ is a function only of the distance between the vertices. In the sequel, we will prove this sub-collection contains a mechanism that is optimal in the limit $n \to \infty$. For such a HSM, (11) reduces to

$$D^n_{\mathcal{H}}(\mathbb{W}) = \sum_{\underline{h}\in\mathcal{H}^n}\binom{n}{\underline{h}}\underline{p}^{\underline{h}}\sum_{d=1}^{n}2dN_d(\underline{h})f(\underline{h}, 2d),$$
$$\text{where } N_d(\underline{h}) = |\left\{\underline{g} \in \mathcal{H}^n : d_G(\underline{g}, \underline{h}) = d\right\}| \quad (12)$$

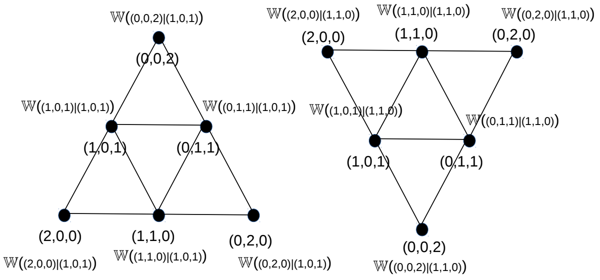

Fig. 2. The PC graphs for $K = 3, N = 2$ are depicted. $(\mathbb{W}(\underline{g}|(1,0,1)) : \underline{g} \in \mathcal{H}^2_3)$, $(\mathbb{W}(\underline{g}|(1,1,0)) : \underline{g} \in \mathcal{H}^2_3)$ are associated with the nodes of the graph on the left and right respectively. Since $(1,1,0)$ and $(1,0,1)$ are neighbors, at every node, the two values have to be within θ and $\frac{1}{\theta}$ of each other.

is the number of vertices at graph distance d from $\underline{h}$.

To evaluate the RHS of $D^n_{\mathcal{H}}(\mathbb{W})$ above, we will need to characterize the sum $\sum_{d=1}^{n} dN_d(\underline{h})f(\underline{h},d)$. Let us consider the sequence $N_1(\underline{h}), N_2(\underline{h}), \cdots, N_n(\underline{h})$ which may be regarded as the distance distribution of the vertex $\underline{h} \in V = \mathcal{H}^n$. Consider Fig. 3 and two sequences $(N_d(\underline{h}) : d = 1, 2, \cdots)$ and $(N_d(\underline{\tilde{h}}) : d = 1, 2, \cdots)$ for any pair $\underline{h}, \underline{\tilde{h}} \in V$ within the dotted circle. These sequences agree on the initial few terms, henceforth referred to as the *head*, and disagree in a few subsequent terms due to the presence of the boundary. As the boundary recedes (i.e. $n \to \infty$), the first term of disagreement recedes, and the head elongates. Alternatively stated, the heads of the sequences $(N_d(\underline{h}) : d = 1, 2, \cdots)$ for $\underline{h}$ within the dotted circle become invariant with $\underline{h}$. Formally, there exists a distance $r \in \mathbb{N}$ such that, for every $\underline{h}$ in the dotted circle, $N_d(\underline{h}) \to N_d$ for all $d = 1, 2, \cdots, r-1$. Moreover $r \to \infty$ as the boundary recedes, i.e., $n \to \infty$. We characterize N_d by considering $\underline{c} := n\underline{p}$. Observe that $N_d(\underline{c}) = |\{\underline{g} \in \mathcal{H}^n : d_G(\underline{g}, \underline{c}) = d\}| =$

$$\begin{aligned} &= |\{\underline{z} \in \mathbb{Z}^K : \underline{c} + \underline{z} \in \mathcal{H}^n, |\underline{z}|_1 = 2d\}| \\ &= |\{\underline{z} \in \mathbb{Z}^K\colon z_i \geq -np_i, z_1 + \cdots + z_K = 0, |\underline{z}|_1 = 2d\}|. \end{aligned}$$

As $n \to \infty$, the lower bound on z_i vanishes, and we have

$$N_d(\underline{c}) \to N_d := |\{\underline{z} \in \mathbb{Z}^k : \textstyle\sum_{k=1}^{K} z_k = 0, |\underline{z}|_1 = 2d\}|. \quad (13)$$

N_d is the number of *integer* points on the *face* of the *integral convex polytope* (9).

Indeed, if $L_{\mathcal{P}}(d) := |\mathbb{Z}^K \cap \mathcal{P}_d|$, then $N_d = L_{\mathcal{P}}(d) - L_{\mathcal{P}}(d-1)$. Notice that $L_{\mathcal{P}}(d)$ is the number of integral points in the $d-$th dilation of the integral convex polytope $\mathcal{P} := \mathcal{P}_1$. $L_{\mathcal{P}}(d)$ and its generating function plays a central role in this article. Ehrhart theory concerns the enumeration of integer points in a integral convex polytope and the objects associated with these counts. We present the foundational results in Ehrhart theory that we will have opportunity to use in our study.

A convex $l-$polytope is a convex polytope of dimension l. A convex $l-$polytope whose vertices have integral co-ordinates is an integral convex $l-$polytope. $L_{\mathcal{P}}(d)$ is the number of integral points in the $d-$th dilation of the integral convex $l-$polytope. Our pursuit of $L_{\mathcal{P}}(d)$ and the associated objects is aided by the following fundamental fact due to Eugène Ehrhart. Ehrhart's theorem states that if $\mathcal{P}$ is an integral convex $l-$polytope, then $L_{\mathcal{P}}(d)$ is a polynomial in d of degree l. We refer to $L_{\mathcal{P}}(d)$ as *Ehrhart's polynomial*. We will identify N_d, and hence $L_{\mathcal{P}}(d)$, precisely in our proof. As evidenced by (8), we will have opportunity to study the generating function of the counts $L_{\mathcal{P}}(d) : d \in \mathbb{N}$. We refer to the formal power series $\mathrm{Ehr}_{\mathcal{P}}(\mathrm{z}) = 1 + \sum_{\mathrm{d}=1}^{\infty} \mathrm{L}_{\mathcal{P}}(\mathrm{d})\mathrm{z}^{\mathrm{d}}$ as the *Ehrhart series* of $\mathcal{P}$, and let $\mathscr{E}_{\mathcal{P},f}(z) := (1-z)\mathrm{Ehr}(\mathrm{z})$. Since $N_d = L_{\mathcal{P}}(d) - L_{\mathcal{P}}(d-1)$, (8)(b) and (c) can be verified. We now partially explain, through a heuristic albeit imprecise limiting argument, the occurrence of $\mathscr{E}_{\mathcal{P},f}(\theta)$ in (8).

Suppose one were to consider the popular Laplace/geometric mechanism $\mathscr{G} : \mathcal{H}^n \Rightarrow \mathcal{H}^n$ and characterize its fidelity. In that case, $\mathbb{W}_{\mathscr{G}}(\underline{g}|\underline{h}) \propto \theta^{\frac{|\underline{g}-\underline{h}|_1}{2}}$, and hence

$$\mathbb{W}_{\mathscr{G}}(\underline{g}|\underline{h}) = \frac{\theta^{d_G(\underline{g},\underline{h})}}{E_{\underline{h}}(\theta)}, \text{ where } E_{\underline{h}}(\theta) = 1 + \sum_{d=1}^{n} N_d(\underline{h})\theta^d \quad (14)$$

is a normalizing constant chosen to ensure $\sum_{\underline{g} \in \mathcal{H}^n} \mathbb{W}_{\mathscr{G}}(\underline{g}|\underline{h}) = 1$. It will be apparent that $\mathbb{W}_{\mathscr{G}}(\cdot|\underline{h})$ is $\theta-$DP only if $E_{\underline{h}}(\theta)$ is invariant with $\underline{h}$. For any (finite) $n \in \mathbb{N}$, this is not true, leading to obstacles in defining a feasible $\theta-$DP HSM analog to the geometric mechanism. Reserving these elements to the proof, we put forth a heuristic limiting argument. As $n \to \infty$, we noted that $N_d(\underline{h}) \to N_d$ and becomes invariant with $\underline{h}$, and hence it is plausible that $E_{\underline{h}}(\theta) \to \mathscr{E}_{\mathcal{P},f}(\theta)$ and $\mathbb{W}_{\mathscr{G}}(\underline{g}|\underline{h}) \to (\mathscr{E}_{\mathcal{P},f}(\theta))^{-1}\theta^{d_G(\underline{g},\underline{h})}$. We substitute this in the RHS of (10), to obtain

$$\lim_{n\to\infty} D^n_{\mathcal{H}}(\mathbb{W}_{\mathscr{G}}) = \lim_{n\to\infty} \sum_{\underline{h} \in \mathcal{H}^n} \binom{n}{\underline{h}} \underline{p}^{\underline{h}} \sum_{d \geq 1} 2d \sum_{\substack{\underline{g} \in \mathcal{H}^n: \\ d_G(\underline{g},\underline{h}) = d}} \frac{\theta^{d_G(\underline{g},\underline{h})}}{\mathscr{E}_{\mathcal{P},f}(\theta)} =$$

$$\lim_{n\to\infty} \sum_{\underline{h} \in \mathcal{H}^n} \binom{n}{\underline{h}} \underline{p}^{\underline{h}} \frac{\sum_{d\geq 1} 2dN_d\theta^d}{\mathscr{E}_{\mathcal{P},f}(\theta)} = \lim_{n\to\infty} \frac{2\theta}{\mathscr{E}_{\mathcal{P},f}(\theta)} \frac{d\mathscr{E}_{\mathcal{P},f}(\theta)}{d\theta}, \quad (15)$$

and the latter quantity is invariant with n, enabling us conclude $\lim_{n\to\infty} D^n_{\mathcal{H}}(\mathbb{W}_{\mathscr{G}}) = \frac{2\theta}{\mathscr{E}_{\mathcal{P},f}(\theta)} \frac{d\mathscr{E}_{\mathcal{P},f}(\theta)}{d\theta}$. In arriving at (15), we used the fact that $\frac{d\mathscr{E}_{\mathcal{P},f}(\theta)}{d\theta} = \sum_{d\geq 1} dN_d\theta^{d-1}$. These informal arguments provide a heuristic explanation for (8) and leaves certain interesting and non-trivial elements, that are addressed in Sec. IV.

IV. Proof of Thm 1

We identify a sequence of (i) upper bounds $D^u_n(\theta) \geq D^n_*(\theta) : n \in \mathbb{N}$ and (ii) lower bounds $D^l_n(\theta) \leq D^n_*(\theta) : n \in \mathbb{N}$, such that $\lim_{n\to\infty} D^u_n(\theta) = \lim_{n\to\infty} D^l_n(\theta)$, and characterize the corresponding limit to obtain $D^*_K(\theta)$. For the upper bounds, we identify a sequence $\mathbb{W}^n : \mathcal{H}^n \Rightarrow \mathcal{H}^n : n \in \mathbb{N}$ of $\theta-$DP HSMs and let $D^u_n(\theta) := D^n_{\mathcal{H}}(\mathbb{W}^n)$. For the lower bound, we identify feasible solutions to the dual of the LP (3) and evaluate its objective value. In the interest of brevity, we only derive the upper bound in this conference article. The lower bound is derived in [13].

We propose $\mathbb{W}^n : \mathcal{H}^n \Rightarrow \mathcal{H}^n$ be a cascade of mechanisms/channels $\mathbb{U}^n : \mathcal{H}^n \Rightarrow \mathcal{H}^n_{\text{ext}}$ and $\mathbb{V}^n : \mathcal{H}^n_{\text{ext}} \Rightarrow \mathcal{H}^n$. $\mathbb{U}^n$ is a geometric mechanism and outputs 'histograms' from an 'enlarged set of histograms'. $\mathbb{V}^n$ takes as input only the output of $\mathbb{U}^n$, and remaps $\mathcal{H}^n_{\text{ext}}$ to $\mathcal{H}^n$. More importantly, it shapes the joint distribution to minimize the expected fidelity. Since a geometric mechanism is, in general, optimal in most DP settings, and $\mathbb{V}^n$ is carefully shaped, we obtain a reasonably good sequence $\mathbb{W}^n$ of mechanisms that is, in the limit, optimal.

Our derivation of the upper bound involves two parts. In the first part, we specify mechanisms $\mathbb{U}^n$, $\mathbb{V}^n$ and relate $D_{\mathcal{H}}(\mathbb{W}^n) (= D^u_n(\theta))$ to $D(\mathbb{U}^n)$ - the fidelity of $\mathbb{U}$. In the second part, we characterize $\mathbb{D}(\mathbb{U}^n)$.

First part : We take a clue from (14). The normalizing terms $E_{\underline{h}}(\theta)$, $E_{\underline{\tilde{h}}}(\theta)$ differ because the tails of the sequences $N_d(\underline{h}) : d \geq 1$ and $N_d(\underline{\tilde{h}}) : d \geq 1$ differ. The latter is due to the presence of the boundary of $\mathcal{H}^n$ (or the PC graph). We enlarge $\mathcal{H}^n$ to eliminate the boundary. This we do by getting rid of the non-negativity constraint in the definition of $\mathcal{H}^n$. The enlarged 'set of histograms' is therefore $\mathcal{H}^n_{\text{ext}} :=\{(h_1, \cdots, h_K) \in \mathbb{Z}^K : \sum_{k=1}^K h_k = n\}$. $\mathcal{H}^n_{\text{ext}}$ is isomorphic to $\{\underline{z} \in \mathbb{Z}^K : \sum_{k=1}^K z_k = 0\}$ and N_d defined in (13) is the number of 'extended histograms' at an $\mathbb{L}_1$ distance of $2d$ from *any* element in $\mathcal{H}^n_{\text{ext}}$. N_d being invariant with $\underline{h}$, we define a $\theta-$DP mechanism $\mathbb{U}^n : \mathcal{H}^n \Rightarrow \mathcal{H}^n_{\text{ext}}$ analogous to the geometric mechanism in (14) as $\mathbb{U}^n(\underline{g}|\underline{h}) := (\mathscr{E}_{\mathcal{P},f}(\theta))^{-1} \theta^{\frac{|\underline{g}-\underline{h}|_1}{2}}$, where $\mathcal{P}$ corresponds to the convex polytope defined in (9). It can be verified that $\mathbb{U}^n$ is $\theta-$DP. By the post-processing theorem of DP, so long as $\mathbb{V}^n : \mathcal{H}^n_{\text{ext}} \Rightarrow \mathcal{H}^n$ takes only the output of $\mathbb{U}^n$, as input, the cascade mechanism $\mathbb{W}^n$ is $\theta-$DP.

The choice of $\mathbb{V}^n$ is based on the fact that the DBs whose histograms differ widely from the mean histogram $n\underline{p}$ contribute an exponentially (in n) small amount to the expected value. $\mathbb{V}^n$ maps the histogram outside the $\mathbb{L}_1-$ball of radius $Rn^{\frac{2}{3}}$ centered at $n\underline{p}$ to the histogram $n\underline{p}$. The histograms within radius $Rn^{\frac{2}{3}}$ of $n\underline{p}$ remain unchanged. Formally, let $\mathbb{V}^n(\underline{g}|\underline{h}) = 1$ if $\underline{g} = \underline{h}, |\underline{h} - n\underline{p}|_1 \leq Rn^{\frac{2}{3}}$, $\mathbb{V}^n(\underline{g}|\underline{h}) = 1$ if $\underline{g} = n\underline{p}, |\underline{h} - n\underline{p}|_1 > Rn^{\frac{2}{3}}$, and $\mathbb{V}^n(\underline{g}|\underline{h}) = 0$ otherwise. For completeness, we also note $\mathbb{W}^n(\underline{g}|\underline{h}) = \sum_{\underline{b} \in \mathcal{H}^n_{\text{ext}}} \mathbb{V}^n(\underline{g}|\underline{b})\mathbb{U}^n(\underline{b}|\underline{h})$.

Does $\mathbb{V}^n$ output a histogram in $\mathcal{H}^n$? The output of $\mathbb{V}^n$ is contained within a $\mathbb{L}_1-$ball of radius $\alpha_n = Rn^{\frac{2}{3}}$ centered at $n\underline{p} \in \mathcal{H}^n$. The boundary of $\mathcal{H}^n$ is at a $\mathbb{L}_1-$distance of at least $\beta_n = \min_{k=1,\cdots,K} np_k$ from $n\underline{p} \in \mathcal{H}^n$. Since $p_k > 0$ for all $k \in \{1, \cdots, K\}$, as $n \to \infty$, $\alpha_n \leq \beta_n$, and the range of $\mathbb{V}^n$ is contained within $\mathcal{H}^n$. The output of mechanism $\mathbb{V}^n$ is indeed a histogram. A formal proof is provided in [13, Appendix 10].

We now prove that $\lim_{n\to\infty} D(\mathbb{W}^n) \leq \lim_{n\to\infty} D(\mathbb{U}^n)$. We describe the main line of our argument. A formal proof is provided in [13, Appendix 11]. Let $D(\mathbb{W}^n, \underline{h}) = \sum_{\underline{g}\in\mathcal{H}^n} \mathbb{W}^n(\underline{g}|\underline{h})|\underline{g} - \underline{h}|_1$, $D(\mathbb{U}^n, \underline{h}) = \sum_{\underline{g}\in\mathcal{H}^n_{\text{ext}}} \mathbb{U}^n(\underline{g}|\underline{h})|\underline{g} - \underline{h}|_1$ denote (unweighted) contributions of $\underline{h}$ to $D^n_{\mathcal{H}}(\mathbb{W}^n)$ and $D(\mathbb{U}^n)$ respectively. Let $B(\frac{1}{2})$ and $B(1)$ be the $\mathbb{L}_1-$balls centered at $n\underline{p}$ of radii $\frac{R}{2}n^{\frac{2}{3}}$ and $Rn^{\frac{2}{3}}$ respectively. Let $B^c(1) := \mathcal{H}^n_{\text{ext}} \setminus B(1)$. For each $\underline{h} \in B(\frac{1}{2})$, the mechanism $\mathbb{V}^n$ has the effect of decreasing $\underline{h}$'s contribution. In other words, for any $\underline{h} \in B(\frac{1}{2})$, $D(\mathbb{W}^n, \underline{h}) \leq D(\mathbb{U}^n, \underline{h})$. This is because (i) $\mathbb{V}^n$ transfers mass placed on $\tilde{g} \in B^c(1)$ - an element farther from $n\underline{p}$ - to $n\underline{p}$, and (ii) $\mathbb{V}^n$ does not alter the mass placed on elements $\underline{g} \in B(1)$ (other than $n\underline{p}$). This is made precise in the sequence of steps [13, (21) - (23)] and can be followed ignoring the earlier steps therein. What about for $\underline{h} \in B^c(\frac{1}{2})$? The weights $\binom{n}{\underline{h}}\underline{p}^{\underline{h}}$ associated with these elements, when summed up, contribute an exponentially small amount. Formally, $\sum_{\underline{h}\in B^c(\frac{1}{2})} \binom{n}{\underline{h}}\underline{p}^{\underline{h}} \leq \exp\{-n\alpha\}$ for some $\alpha > 0$. Since $|\underline{g} - \underline{h}|_1 \leq 2n$ whenever $\underline{h}, \underline{g}\ in \mathcal{H}^n$, we have $D(\mathbb{W}^n, \underline{h}) \leq 2n \exp\{-\alpha n\}$ and hence $\sum_{\underline{h}\in B^c(\frac{1}{2})} \binom{n}{\underline{h}}\underline{p}^{\underline{h}} D(\mathbb{W}^n, \underline{h}) \to 0$ as $n \to 0$.

Second Part : From the definition of $\mathbb{U}^n(\cdot|\cdot)$ and (8), we note $2\theta \frac{d}{d\theta}\mathscr{E}_{\mathcal{P},f}(\theta) = \sum_{d=1}^\infty 2dN_d\theta^d$. With these, we have

$$D(\mathbb{U}^n) = \sum_{\underline{h}\in\mathcal{H}^n} \sum_{\underline{g}\in\mathcal{H}^n_{\text{ext}}} \binom{n}{\underline{h}} \underline{p}^{\underline{h}} \mathbb{U}^n(\underline{g}|\underline{h})|\underline{g} - \underline{h}|_1$$

$$= \sum_{\underline{h}\in\mathcal{H}^n} \binom{n}{\underline{h}} \underline{p}^{\underline{h}} \sum_{\underline{g}\in\mathcal{H}^n_{\text{ext}}} \frac{1}{\mathscr{E}_{\mathcal{P},f}(\theta)} \theta^{\frac{|\underline{g}-\underline{h}|_1}{2}} |\underline{g} - \underline{h}|_1$$

$$= \sum_{\underline{h}\in\mathcal{H}^n} \binom{n}{\underline{h}} \underline{p}^{\underline{h}} \frac{1}{\mathscr{E}_{\mathcal{P},f}(\theta)} 2\theta \frac{d\mathscr{E}_{\mathcal{P},f}(\theta)}{d\theta} = \frac{2\theta}{\mathscr{E}_{\mathcal{P},f}(\theta)} \frac{d\mathscr{E}_{\mathcal{P},f}(\theta)}{d\theta}, \tag{16}$$

where the first equality in (16) follows from $\sum_{\underline{g}\in\mathcal{H}^n_{\text{ext}}} \theta^{\frac{|\underline{g}-\underline{h}|_1}{2}} |\underline{g} - \underline{h}|_1 = \sum_{d=1}^\infty 2dN_d\theta^d$ (rearrange sum as in (11)). In the sequel, we let $\mathscr{D}_K(\theta)$ denote the RHS of (16). From (16), (8) it suffices to characterize either the Ehrhart series $\text{Ehr}_{\mathcal{P}}(\theta)$ or $\mathscr{E}_{\mathcal{P},f}(\theta)$, of $\mathcal{P} = \mathcal{P}_1$, where $\mathcal{P}_d$ is the polytope characterized in (9). In the interest of brevity, we refer to [14], wherein it is stated $\mathscr{E}_{\mathcal{P},f}(\theta) \overset{(a)}{=} \frac{S_{K-1}(\theta)}{(1-\theta)^{K-1}}$. The proof [15, Thm 3.1, Eqn (3.7)] is via contour integration. We shall provide in a subsequent version of this article, an alternate proof via a simple counting principle, Ehrhart theory and analytic methods. Substituting this in (16), verify that $D(\mathbb{U}^n) = \mathscr{D}_K(\theta) = 2\theta \left\{ \frac{K-1}{1-\theta} + \frac{S'_{K-1}(\theta)}{S_{K-1}(\theta)} \right\}$. This concludes our second step.

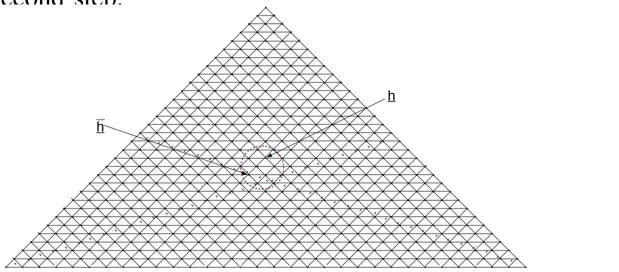

Fig. 3. Consider distance distribution of nodes within the dotted circle.

References

[1] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Proc. of TCC 2006, pp. 265–284.

[2] M. Beck and S. Robins, Computing the Continuous Discretely : Integer-Point Enumeration in Polyhedra, 2nd ed. Springer, 2015.

[3] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Comp. Sc., 2014.

[4] W. Wang, L. Ying, and J. Zhang, "On the relation between identifiability, differential privacy, and mutual-information privacy," IEEE Transactions on Information Theory, vol. 62, no. 9, pp. 5018–5029, Sept 2016.

[5] K. Kalantari, L. Sankar, and A. D. Sarwate, "Optimal differential privacy mechanisms under hamming distortion for structured source classes," in 2016 IEEE ISIT, July 2016, pp. 2069–2073.

[6] M. Hardt and K. Talwar, "On the geometry of differential privacy," in Proc. of 2010 STOC, ser. STOC '10. ACM, pp. 705–714.

[7] A. De, "Lower bounds in differential privacy," in Proc. of the 9th Intl. Conf. on Theory of Cryptography. Springer-Verlag, 2012, pp. 321–338.

[8] Q. Geng and P. Viswanath, "The optimal noise-adding mechanism in differential privacy," IEEE Trans. on Info. Th., vol. 62, no. 2, 2016.

[9] ——, "Optimal noise adding mechanisms for approximate differential privacy," IEEE Trans. on Info. Th., vol. 62, no. 2, pp. 952–969, 2016.

[10] M. Bun, J. Ullman, and S. Vadhan, "Fingerprinting codes and the price of approximate differential privacy," in Proc. of 2014 STOC. ACM.

[11] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," SIAM Jrnl on Computing, vol. 41, no. 6, pp. 1673–1693.

[12] G. Pólya and G. Szeg o, Problems and Theorems in Analysis. Springer, 1976, vol. 2.

[13] A. Padakandla, P. R. Kumar, and W. Szpankowski, "Precise asymptotics of Utility-Privacy tradeoff via Ehrhart Theory," available at https://www.cs.purdue.edu/homes/arunpr/preprints/UtilityPrivacy.pdf .

[14] "OEIS Sequence a103881, http://oeis.org/a103881."

[15] J. H. Conway and N. J. A. Sloane, "Low–dimensional lattices. vii. coordination sequences," Proc. of the Royal Soc. of London A: Math., Physical and Engg. Sciences, vol. 453, no. 1966, pp. 2369–2389, 1997.