On an Information Theoretic Approach to Cardinality Estimation

Hung Q. Ngo
RelationalAI Inc., Berkeley, CA, USA

Abstract

This article is a companion to an invited talk at ICDT'2022 with the same title.

Cardinality estimation is among the most important problems in query optimization. It is well-documented that, when query plans go haywire, in most cases one can trace the root cause to the cardinality estimator being far off. In particular, traditional cardinality estimation based on selectivity estimation may sometimes under-estimate cardinalities by orders of magnitudes, because the independence or the uniformity assumptions do not typically hold.

This talk outlines an approach to cardinality estimation that is "model-free" from a statistical stand-point. Being model-free means the approach tries to avoid making any distributional assumptions. Our approach is information-theoretic, and generalizes recent results on worst-case output size bounds of queries, allowing the estimator to take into account histogram information from the input relations. The estimator turns out to be the objective of a maximization problem subject to concave constraints, over an exponential number of variables. We then explain how the estimator can be computed in polynomial time for some fragment of these constraints. Overall, the talk introduces a new direction to address the classic problem of cardinality estimation that is designed to circumvent some of the pitfalls of selectivity-based estimation. We will also explain connections to other fundamental problems in information theory and database theory regarding information inequalities.

The talk is based on (published and unpublished) joint works with Mahmoud Abo Khamis, Sungjin Im, Hossein Keshavarz, Phokion Kolaitis, Ben Moseley, XuanLong Nguyen, Kirk Pruhs, Dan Suciu, and Alireza Samadian Zakaria

2012 ACM Subject Classification Information systems → Query optimization; Information systems → Query planning; Information systems → Join algorithms

Keywords and phrases Cardinality Estimation, Information Theory, Polymatroid Bound, Worst-case Optimal Join

Digital Object Identifier 10.4230/LIPIcs.ICDT.2022.1

Category Invited Talk

Acknowledgements I would like to thank the technical program committee and the program chair Dan Olteanu of ICDT 2022 for inviting me to give a keynote talk at the conference

1 Introduction

Cardinality estimation [25] is a crucial component of the query optimization pipeline. The problem can be formulated as shown in Figure 1. We collect statistical profiles $s(\boldsymbol{D})$ of relations in the database $\boldsymbol{D}$. (Profiles are also called "system catalogs" [51], among other names.) For a given query Q, the problem is to come up with an accurate estimate $\hat{q}$ of $|Q(\boldsymbol{D})|$ from the profile $s(\boldsymbol{D})$, as quickly as possible. The main tradeoff dimensions are the *space complexity* of $s(\boldsymbol{D})$, the *speed* of computing $\hat{q}$, and the *accuracy* of the estimate.

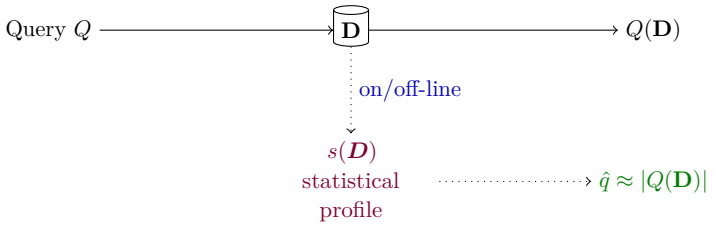

Figure 1 Cardinality Estimation.

Cardinality estimation is one of the most, if not the most, important component of the query optimization pipeline [38, 42, 40]. Cardinality estimates are the main parameters in the cost-estimators of logical query plans, physical query plan, parallel query processing, and in computing budgets for in-memory query processing [1]. Guy Lohman [42] expressed the issue succinctly:

> *The root of all evil*, the Achilles Heel of query optimization, is the estimation of the size of intermediate results, known as *cardinalities*. Everything in cost estimation depends upon how many rows will be processed, so the entire cost model is predicated upon the cardinality model. ... In my experience, the cost model may introduce errors of at most 30% for a given cardinality, but the cardinality model can quite *easily introduce errors of many orders of magnitude!*

Cardinality estimation is a very challenging problem. After all, the problem is fundamentally a lossy compression problem: different databases may have the same or similar (small) statistical profiles, how do we tell them apart? After more than half a century of theory and implementation of relational database systems, whose global market size is more than 60 billions USD in 2020[1], commercial database systems still routinely misestimate cardinalities by a factor of 1000 or more [38]. Two major reasons for the misestimation are: (1) relational RDBMSs employ estimators that make distributional assumptions about the data (such as uniformity) which may not hold in real workloads or in standard benchmarks, and (2) traditional estimators treat selection predicates independently, leading to error accumulation on large queries. Hence, estimation errors grow exponentially as the number of joins increases [30, 40]. The lack of whole-query constraints consideration also led to strange phenomena where "simply by swapping predicates or relations", the estimates can change drastically [38].

A natural question is, "how can database theory be (more) helpful in addressing these challenges?" There are a steady stream of works from the database systems community studying the cardinality estimation problem, year after year (see [43, 43, 11, 39, 26, 36, 19, 57, 25, 44] and references thereoff). Yet there are remarkably much fewer works from the database theory community on this fundamental problem. The Alice book [3] does not have any mention of cardinality estimation (theory or otherwise). The few database theory papers related to cardinality estimation were on aspects of an existing approach, and the vast majority of them were published in the 1990s [18, 52, 41, 12, 37, 13, 31, 24, 21, 23, 16]. There are very few papers, if any at all, that attempt to take a fresh look at cardinality estimation.

[1] https://www.gartner.com/en/documents/4001330/market-share-database-management-systems-worldwide-2020

This talk presents an cardinality estimation approach that is developed and implemented at RelationalAI Inc. (with collaborators). The approach is aimed to address both of the weaknesses mentioned above. The first weakness is the (strong) model-based assumptions which leading to the lack of robustness. The second weakness is the one-selection-at-a-time estimation strategy which leads to error accumulation, in part due to its inability to take into account statistical information available in the query but *outside* the scope of the selection predicate being considered.

Before describing the approach, we describe the cardinality estimation problem more formally. We shall start with the profile $s(\boldsymbol{D})$, and then describe different types of estimators $\hat{q}$, the two components shown in Figure 1.

There are several types of (statistical) profiles $s(\boldsymbol{D})$, which can be collected offline or online (after seeing the query Q). In the offline case, the simplest profiles involve integrity constraints such as functional dependencies, number of distinct values of a given attribute, and base-table cardinalities and number of disk pages. More complex profiles include synopses such as histograms, approximate histograms, frequency moments [29, 25] of degree sequences, moments and quantiles for numeric variables, trie or bigrams/trigrams for string variables, and so on [15]. In the online case, $s(\boldsymbol{D})$ can contain samples from the database, constructed based on the query Q. Sampling-based estimator is a deep topic both theoretical and practically [40, 19, 39, 41, 27]. More recently, machine learning based approaches are starting to show up in the literature [57, 36], but the jury is still out on whether they are practical enough to be in production.

This talk concentrates on the *offline* case, where $s(\boldsymbol{D})$ contains some forms of histogram information and integrity constraints. The setup is sufficiently simple for a fresh look at the problem, yet powerful enough to capture and even generalize standard settings in most RDBMSs. Simplicity also leads to practicality.

In what follows, we write $\boldsymbol{D}' \models s(\boldsymbol{D})$ to mean "database $\boldsymbol{D}'$ that has the same statistical profile $s(\boldsymbol{D})$", and $\boldsymbol{D}' \sim \mathcal{D}$ to mean "$\boldsymbol{D}'$ drawn from some database distribution $\mathcal{D}$". There are two types of cardinality estimators:

$$\hat{q} \approx \mathop{\mathbb{E}}_{\boldsymbol{D}' \sim \mathcal{D}} [|Q(\boldsymbol{D}')| \mid \boldsymbol{D}' \models s(\boldsymbol{D})] \qquad \text{average-case / model-based} \tag{1}$$

$$\hat{q} \approx \sup_{\boldsymbol{D}' \models s(\boldsymbol{D})} |Q(\boldsymbol{D}')| \qquad \text{worst-case / model-free} \tag{2}$$

The *average-case estimator* (1) relies on certain distributional assumptions about the data. The assumptions are modeled with a distribution $\mathcal{D}$ from which the data is drawn. The estimator aims to approximate the conditional expectation of $|Q(\boldsymbol{D}')|$ over all databases $\boldsymbol{D}'$ drawn from the distribution $\mathcal{D}$, conditioned on the fact that $\boldsymbol{D}'$ has the statistical profile $s(\boldsymbol{D})$. On the plus side, if $\mathcal{D}$ is a good model for the input data, then the estimator adapts well to the data, giving the query planner more accurate estimates, leading to better query plans. On the minus side, when the data does not conform on the assumptions baked into $\mathcal{D}$, one can end up with very bad query plans.

The *worst-case estimator* (2) approximates the worst-possible output size of query Q over all databases $\boldsymbol{D}'$ having the same statistical profile $s(\boldsymbol{D})$. (These are also called "pessimistic estimators" [11].) Two main advantages of worst-case estimators are that: (1) they are robust to outliers, heavy skews or corner cases do not affect the estimator and thus they help avoid query plans which explode in runtime under bad inputs, and (2) they can be used to guarantee memory budget during query evaluation. The main disadvantage is that, for a given dataset, the worst-case estimate may be far from the actual output size $|Q(\boldsymbol{D})|$, potentially leading to sub-optimal query plans. However, there are recent experimental research results showing that worst-case estimators can be quite effective on some benchmarks [11, 26].

Traditional approaches to cardinality estimation are all model-based. In the offline case (i.e. no sampling / learning), the main approach is stil the one proposed from the System-R days [54]. What we advocate for and present in this paper is model-free, designed to address the robustness concern of the traditional estimator.

The model-free estimation problem (2) is exactly the problem of estimating the worst-case output size of a join query. Studying this problem has led to a new class of join algorithms called *worst-case optimal join algorithms* [56, 47, 48]. For a certain class of statistical profile $s(\boldsymbol{D})$ (called "degree-constraints" in [6, 7]), the worst-case output size bound (2) is known to be deeply connected to important information theory questions [7, 5]). This is where our story begins: how do we solve the worst-case estimation problem (2) beyond the degree-constraints setup proposed in [7], taking into account, for example, histogram information?

The information-theoretic approach we present in this paper can be thought of as a "multiway cardinality estimator", to parallel the notion of "multiway-join operator" that worst-case optimal join algorithms are, as opposed to "binary-join cardinality estimator" which does not take into account constraints outside of the binary join being considered, which is also exactly why binary-join is not worst-case optimal!

The rest of this paper is organized as follows. Section 2 briefly presents the traditional cardinality estimation approach based on estimating the *selectivity*, and the notion of degree-constraints and polymatroid bounds. Section 3 presents an information-theoretic framework which begins with generalizing degree-constraints to the so-called "histogrammed frequency-moment constraints", and ends with an optimization problem for approximating (2) based on the entropy argument. Section 4 describes some of the main algorithmic and theoretical challenges we face while realizing the new approach. Section 5 concludes the paper.

2 Background

We use bold-face capital letters, such as $\boldsymbol{X}$, to denote tuples/sets of variables, capital letters such as X to denote a variable, bold-face lower-case letters, such as $\boldsymbol{x}$ to denote specific tuples in the domain $\mathsf{Dom}(\boldsymbol{X})$ of $\boldsymbol{X}$, and naturally non-bold-face letter x to denote a particular value in $\mathsf{Dom}(X)$.

2.1 Conjunctive Queries

We restrict our attention to estimating the output size of conjunctive queries, with a special emphasis on *full* conjunctive queries. We associate a *full conjunctive query* Q to a multi-hypergraph $\mathcal{H} := (\boldsymbol{V}, \mathcal{E})$, $\mathcal{E} \subseteq 2^{\boldsymbol{V}}$; the query is written as

$$Q(\boldsymbol{V}) \leftarrow \bigwedge_{\boldsymbol{F} \in \mathcal{E}} R_{\boldsymbol{F}}(\boldsymbol{F}), \tag{3}$$

with variables $\boldsymbol{V}$ and *atoms* $R_{\boldsymbol{F}}(\boldsymbol{F})$ for each $\boldsymbol{F} \in \mathcal{E}$. We also write $R_{\boldsymbol{F}}$ to avoid duplication. The atoms $R_{\boldsymbol{F}}$ represent either relational tables or built-in predicates whose columns are variables in $\boldsymbol{F}$.

▶ **Example 1.** *The following query corresponds to a triangle (hyper)graph, it is often called the "triangle query":*

$$Q_{\triangle}(A, B, C) \leftarrow E(A, B) \wedge E(A, C) \wedge E(B, C). \tag{4}$$

The relation E contains all edges of the graph that we want to count the number of triangles of.

Table 1 System-R-style selectivity estimation.

Predicate p	$s(p)$	note	default	assumption
$\neg p'$	$1 - s(p')$			
$p_1 \wedge p_2$	$s(p_1) \cdot s(p_2)$			independence
$A = c$	$1/d_A$	$d_A = \#$ of dist. vals	$\frac{1}{10}$	uniformity
$A > c$	$\frac{\max_A - c}{\max_A - \min_A}$	if known	$\frac{1}{3}$	uniformity
$c_1 < A < c_2$	$\frac{c_2 - c_1}{\max_A - \min_A}$	if known	$\frac{1}{4}$	uniformity
$R(A,B) \bowtie S(B,C)$	$\frac{1}{\max(d_B^R, d_B^S)}$	i.e. $\lvert R \bowtie S\rvert \approx \lvert R\rvert \cdot \lvert S\rvert \cdot s(\bowtie)$		uniformity
A IN L	$\min\{1/2, s(A = c)\lvert L\rvert\}$			
A IN Q	$\lvert Q\rvert/\lvert X\rvert$	X is cross-prod		

▶ **Example 2.** *Another example is the following query*

$$Q_+(A, B) \leftarrow R(A) \wedge S(B) \wedge A + B = 5 \tag{5}$$

Here, the predicate $A + B = 5$ is a built-in predicate, which is morally R_{AB} if we want to write it in the form (3)*.*

2.2 Traditional selectivity estimation

Most modern RDBMSs adopts a variant of System-R [53] cardinality estimation approach, which works as follows. Consider a (conjunctive) query which contains input relations and a collection of selection predicates. Without the predicates, the output size estimate is the product of input relation sizes. Each selection predicate reduces the estimate by a certain factor called the *selectivity* of the predicate. Table 1 summarizes how the selectivity of typical predicates are computed.

On the one hand, this approach has served us sufficiently well for the past 60 years or so, as evident by the commercial success. On the other hand, drawbacks of this approach are well-documented [38]. We list here a few key weaknesses:

- The approach is (probabilistic) model-based, but there does not seem to be a known theory for probabilistic guarantee regarding the quality of the estimate.
- The approach does not take into account all constraints at once, it is one predicate at a time, assuming independence. Hence, it is prone to under-estimation when the number of predicates involved is large.
- Is not entirely clear how to incorporate more known constraints into the estimator. The most obvious omission is in the fact that functional dependencies are *not* taken into account. In Example (2) above, for instance, our estimation approach will give a bound of $\min\{|R|, |S|\}$, while the traditional estimator approach likely gives $|R| \times |S|$.
- It is not clear how to estimate the output size of non-full conjunctive queries. (Of course, one can always take some trivial bound such as the cross-product of the distinct counts of the free variables.)
- The approach does not take into account histogram information when estimating join sizes. It does make use of the number of distinct values over the joined variable domain; however, in practice, histograms over the same attribute on different relations have mis-aligned boundaries.

2.3 Degree constraints

The notion of "degree constraints" was introduced in [7] to model a simple yet powerful form of statistical profiles $s(\boldsymbol{D})$. A *degree constraint* is a triple $(\boldsymbol{X}, \boldsymbol{Y}, N)$, where $\boldsymbol{X} \subsetneq \boldsymbol{Y} \subseteq \boldsymbol{V}$ and $N \in \mathbb{N}$. The relation $R_{\boldsymbol{F}}$ is said to *guard* the degree constraint $(\boldsymbol{X}, \boldsymbol{Y}, N)$ if $\boldsymbol{Y} \subseteq \boldsymbol{F}$ and

$$\max_{\boldsymbol{x}} |\pi_{\boldsymbol{Y}} \sigma_{\boldsymbol{X}=\boldsymbol{x}} R_{\boldsymbol{F}}| \leq N. \tag{6}$$

In plain language, the degree constraint states that: "in the relation $R_{\boldsymbol{F}}$, for every fixed binding $\boldsymbol{X} = \boldsymbol{x}$, there are at most N bindings $\boldsymbol{y}$ of $\boldsymbol{Y}$ for which $\boldsymbol{y}$ is in the projection of $R_{\boldsymbol{F}}$ onto the attributes $\boldsymbol{Y}$. Note that a given relation may guard multiple degree constraints.

Let DC denote a set of degree constraints. The input database $\boldsymbol{D}$ is said to *satisfy* DC if every constraint in DC has a guard, in which case we write $\boldsymbol{D} \models \mathsf{DC}$.

A *cardinality constraint* is an assertion of the form $|R_{\boldsymbol{F}}| \leq N$, for some $\boldsymbol{F} \in \mathcal{E}$; it is exactly the degree constraint $(\emptyset, \boldsymbol{F}, N)$ guarded by $R_{\boldsymbol{F}}$. A *functional dependency* $\boldsymbol{X} \to \boldsymbol{Y}$ is a degree constraint $(\boldsymbol{X}, \boldsymbol{Y}, 1)$. In particular, degree constraints strictly generalize both cardinality constraints and functional dependencies.

In the triangle query (4), suppose in addition to knowing that $|E| = N$ we also know that the out-degree of every vertex is bounded by D. Then this database (i.e. the graph G) satisfies the following degree constraints: $(\emptyset, \{u, v\}, N)$ for every pair $u, v \in [3]$. and $(\{1\}, \{1, 2\}, D)$ $(\{2\}, \{2, 3\}, D)$ and $(\{3\}, \{3, 1\}, D)$.

Our problem setting is general, where we are given a query of the form (3) and a set DC of degree constraints satisfied by the input database $\boldsymbol{D}$. The model-free cardinality estimation problem is to find a good upper bound of, or to determine exactly the quantity $\sup_{\boldsymbol{D} \models \mathsf{DC}} |Q(\boldsymbol{D})|$, the worst-case output size of the query given that the input satisfies the degree constraints. To describe the solution space, we need a detour to some classes of set functions.

2.4 Families of set functions

Let $n = |\boldsymbol{V}|$. A function $f : 2^{\boldsymbol{V}} \to \mathbb{R}_+$ is called a (non-negative) *set function* on $\boldsymbol{V}$. A set function f on $\boldsymbol{V}$ is *modular* if $f(\boldsymbol{S}) = \sum_{X \in \boldsymbol{S}} f(\{X\})$ for all $\boldsymbol{S} \subseteq \boldsymbol{V}$, is *monotone* if $f(\boldsymbol{X}) \leq f(\boldsymbol{Y})$ whenever $\boldsymbol{X} \subseteq \boldsymbol{Y}$, and is *sub-modular* if $f(\boldsymbol{X} \cup \boldsymbol{Y}) + f(\boldsymbol{X} \cap \boldsymbol{Y}) \leq f(\boldsymbol{X}) + f(\boldsymbol{Y})$ for all $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$. A function $h : 2^{\boldsymbol{V}} \to \mathbb{R}_+$ is said to be *entropic* if there is a joint distribution on $\boldsymbol{V}$ with entropy function H such that $h(\boldsymbol{S}) = H[\boldsymbol{S}]$ for all $\boldsymbol{S} \subseteq \boldsymbol{V}$.

Unless specified otherwise, we will only consider *non-negative* and *monotone* set functions f for which $f(\emptyset) = 0$; this assumption will be implicit in the entire paper. Furthermore, for $\boldsymbol{X} \subseteq \boldsymbol{Y}$, we will write $h(\boldsymbol{Y} \mid \boldsymbol{X}) := h(\boldsymbol{Y}) - h(\boldsymbol{X})$ for all our set functions h.

Let M_n and Γ_n denote the set of all (non-negative and monotone) modular and submodular set functions on $\boldsymbol{V}$, respectively. The set Γ_n is called the set of *polymatroidal functions*, or simply *polymatroids*. Let Γ_n^* denote the set of all entropic functions on n variables, and $\overline{\Gamma}_n^*$ denote its topological closure (in the Euclidean space, where we think of a polymatroid function $f : 2^{\boldsymbol{V}} \setminus \{\emptyset\} \to \mathbb{R}$ as a vector in $\mathbb{R}^{2^n - 1}$).

The notations $\Gamma_n, \Gamma_n^*, \overline{\Gamma}_n^*$ are standard in information theory. It is known [58] that Γ_n^* is a cone which is not topologically closed. And hence, when optimizing over this cone we take its topological closure $\overline{\Gamma}_n^*$, which is convex. It is easy to see that M_n and Γ_n are *polyhedral cones*. (Note that we can view them as either functions or vectors in $\mathbb{R}^{2^n - 1}$.)

There is another interesting class of set functions called *normal* functions [4, 6], defined as follows. For every $\boldsymbol{W} \subsetneq \boldsymbol{V}$, a *step function* $s_{\boldsymbol{W}} : 2^{\boldsymbol{V}} \to \mathbb{R}_+$ is defined by

$$s_{\boldsymbol{W}}(\boldsymbol{X}) = \begin{cases} 0 & \boldsymbol{X} \subseteq \boldsymbol{W} \\ 1 & \text{otherwise} \end{cases}$$

A function is normal if it is a non-negative linear combination of step functions. Let N_n denote the set of normal functions on $\boldsymbol{V}$.

As mentioned above, entropic functions satisfy non-negativity, monotonicity, and submodularity. Linear inequalities regarding entropic functions derived from these three properties are called *Shannon-type* inequalities. For a very long time, it was widely believed that Shannon-type inequalities form a complete set of linear inequalities satisfied by entropic functions, namely $\overline{\Gamma}_n^* = \Gamma_n$. This indeed holds for $n \leq 3$, for example. However, in 1998, in a breakthrough paper in information theory, Zhang and Yeung [59] presented a new inequality which cannot be inferred from Shannon-type inequalities. Their result proved that, $\overline{\Gamma}_n^* \subsetneq \Gamma_n$ for any $n \geq 4$.

The following inclusion chain can be found in a combination of [58, 4].

▶ **Theorem 3.** *The following chain of inclusion holds*

$$\mathsf{M}_n \subseteq \mathsf{N}_n \subseteq \Gamma_n^* \subseteq \overline{\Gamma}_n^* \subseteq \Gamma_n \tag{7}$$

When $n \geq 4$, all of the containments are strict.

2.5 Entropic and Polymatroid Bounds

In [7], we used families of set functions to describe answers to the worst-case cardinality estimation problem $\sup_{\boldsymbol{D} \models \mathsf{DC}} |Q(\boldsymbol{D})|$ This quantity is called the *worst-case output size* of the query, over databases satisfying the input degree constraints. Algorithms evaluating Q running in time $\tilde{O}(|\boldsymbol{D}| + \sup_{\boldsymbol{D} \models \mathsf{DC}} |Q(\boldsymbol{D})|)$ are called *worst-case optimal join algorithms* [47, 49, 56].

To obtain a bound in the general case, we employ the entropy argument, which is widely used in extremal combinatorics [34, 14, 50, 20], and in database theory [22, 6, 7]. The reader is referred to [47] for a brief historical account in relation to database theory.

Define the collection HDC of set functions satisfying the degree constraints DC:

$$\mathsf{HDC} := \{h \mid h : 2^{\boldsymbol{V}} \to \mathbb{R},\ h(\boldsymbol{Y}) - h(\boldsymbol{X}) \leq \log N, \qquad \forall (\boldsymbol{X}, \boldsymbol{Y}, N) \in \mathsf{DC}\}. \tag{8}$$

The entropy argument immediately gives the following result, first explicitly formulated in joint works with Abo Khamis and Suciu [6, 7]:

▶ **Theorem 4** (From [6, 7])**.** *Let Q be a conjunctive query and* DC *be a given set of degree constraints, then for any database $\boldsymbol{D}$ satisfying* DC*, we have*

$$\sup_{\boldsymbol{D} \models \mathsf{DC}} \log |Q(\boldsymbol{D})| = \max_{h \in \overline{\Gamma}_n^* \cap \mathsf{HDC}} h(\boldsymbol{V}) \qquad \textit{(entropic bound)} \tag{9}$$

$$\leq \max_{h \in \Gamma_n \cap \mathsf{HDC}} h(\boldsymbol{V}) \qquad \textit{(polymatroid bound)} \tag{10}$$

Furthermore, the entropic bound is asymptotically tight and the polymatroid bound is not.

3 An Information Theoretic Framework

This section first introduces a more powerful notion of constraints, enriching what can be stored in the profile $s(\boldsymbol{D})$. We explain how this type of constraints capture very well the kind of summary information that histograms store. Finally, we explain how to formulate a worst-case cardinality estimator subject to these constraints.

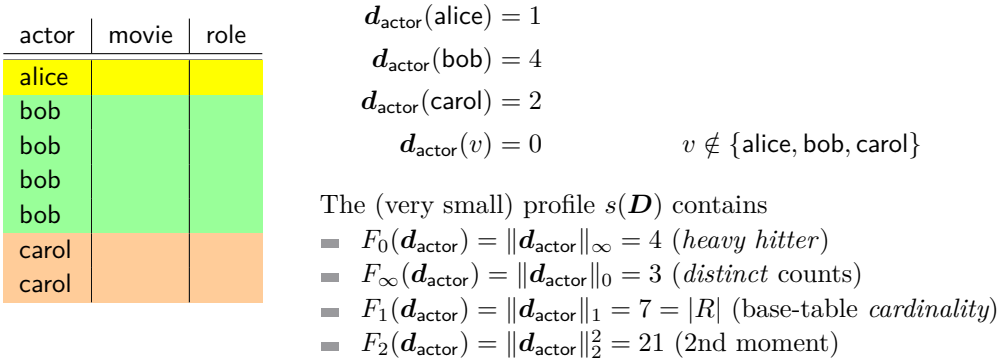

Figure 2 An example of degree-norm constraints.

3.1 Frequency-moment constraints

3.1.1 Motivation

To motivate the notion of frequency-moment constraints, let us consider an example shown in Figure 2. Let's say we have a table $R(\mathsf{actor}, \mathsf{movie}, \mathsf{role})$ and we would like to compute a small statistical profile (that goes into $s(\boldsymbol{D})$) of this table. We want to be able to capture as much of the joint distribution over three variables actor, movie, and role using as little space as possible.

One possible representation is to look at each of the marginal distributions over individual variables. On the variable actor, the marginal distribution is summarized with a *frequency vector* $\boldsymbol{d}_{\mathsf{actor}}$ which counts, for each actor, the number of times the actor occurs in the table. This marginal mass vector is typically too large to be part of the profile $s(\boldsymbol{D})$. Instead, the idea of frequency-moment constraints is to include in $s(\boldsymbol{D})$ some frequency-moment [8] of this vector: $F_\ell(\boldsymbol{d}_{\mathsf{actor}})$, for $\ell \in \{0, 1, 2, \infty\}$. The *frequency-moment* F_ℓ of a vector $\boldsymbol{v}$ is defined by

$$F_\ell(\boldsymbol{v}) := \begin{cases} \|\boldsymbol{v}\|_\ell & \ell \in \{0, 1, +\infty\} \\ \|\boldsymbol{v}\|_\ell^\ell & \ell \notin \{0, 1, +\infty\}. \end{cases} \tag{11}$$

Theoretically, we are certainly free to pick ℓ-moments for values of ℓ beyond $\{0, 1, 2, \infty\}$, but practically they are not very meaningful. As shown in Figure 2, the 0-, 1-, and ∞-moments capture commonly used statistics in RDBMSs: heavy hitters, distinct value counts, and base-table cardinalities.

3.1.2 Formal definition

Let R be a relation, and $\boldsymbol{X}, \boldsymbol{Y}$ be subsets of attributes of the relation. Define the *conditional frequency vector* $\boldsymbol{d}_{\boldsymbol{Y}|\boldsymbol{X}}$ to be

$$\boldsymbol{d}^R_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{x}) = |\pi_{\boldsymbol{Y}} \sigma_{\boldsymbol{X}=\boldsymbol{x}} R| \tag{12}$$

When R is clear from context, we drop the subscript R to reduce cluttering. Note that $\boldsymbol{X}$ can be empty, where $\boldsymbol{d}^R_{\boldsymbol{Y}|\emptyset}() = |\pi_{\boldsymbol{Y}}(R)|$ counts the number of distinct $\boldsymbol{Y}$-tuples in R.

▶ **Definition 5** (Frequency-moment constraint)**.** *A* frequency-moment constraint *(or just* FM-constraints *for short) is is a quintuple* $(\boldsymbol{X}, \boldsymbol{Y}, N, \ell, R)$*, where* $\boldsymbol{X} \subseteq \boldsymbol{Y}$ *are sets of variables,* $N \in \mathbb{R}_+$*, and* $\ell \in [0, +\infty]$ *is a nonnegative real number or infinity.* R *is an input relation. The constraint states that*

$$F_\ell(\boldsymbol{d}^R_{Y|X}) \le N \tag{13}$$

The values of ℓ that are most meaningful in practice are $\{0, 1, 2, +\infty\}$.

Note the following fact: for the same relation R, we have

$$F_0(\boldsymbol{d}_{\boldsymbol{Y}|\boldsymbol{X}}) = F_\infty(\boldsymbol{d}_{\boldsymbol{X}|\emptyset}) \qquad F_1(\boldsymbol{d}_{\boldsymbol{Y}|\boldsymbol{X}}) = F_\infty(\boldsymbol{d}_{\boldsymbol{X}\cup\boldsymbol{Y}|\emptyset}) \tag{14}$$

In particular, adding the ability to measure 1- or 0-moments does not move us beyond the degree constraints setting of ∞-frequency-moments. This fact will change when we use FM-constraints in the context of histograms, as presented in the next section.

In the obvious way, FM-constraints capture profile information typically used in RDBMSs, such as functional dependencies, base-table cardinalities, distinct value counts, and heavy hitters. It is a strict generalization of degree constraints.

3.1.3 Related recent works

Independent of our work, there are a couple of recent works which dealt with degree sequences and their norms.

Jayaraman, Ropell and Rudra [32] considered the join problem where the input database contains arity-2 relations, each of which has a degree vector some of whose norms are given as input to the join computation problem. They derived a worst-case optimal join algorithm under this input.

Deeds, Suciu, Balazinska, and Cai [17] considered the problem setting where entire degree vectors $\boldsymbol{d}^R_{\boldsymbol{Y}|X}$ are given, along with maximum tuple multiplicities. They derived novel bounds for the output size given this information.

The setup and results of both papers are orthogonal to what is presented in this talk.

3.2 Histograms

The frequency-moment constraints are powerful building blocks for summarizing the data in a database $\boldsymbol{D}$. To increase the granularity of the statistical profile $s(\boldsymbol{D})$, we partition the data and have frequency-moment constraints for each part of the data space. This is the idea behind histograms [2, 29]. Given a relation $R_{\boldsymbol{Y}}$ and a set $\boldsymbol{X} \subset \boldsymbol{Y}$ of its attributes, to build an $\boldsymbol{X}$*-histogram*, we partition the active domain of $\boldsymbol{X}$ into some k parts:

$$\mathsf{Dom}(\boldsymbol{X}) := \prod_{X\in\boldsymbol{X}} \mathsf{Dom}(X) = B_1 \cup B_2 \cup \cdots \cup B_k.$$

The B_is are called "buckets". Let $\mathcal{B} = \{B_1, B_2, \ldots, B_k\}$. For each bucket $B \in \mathcal{B}$, the B-conditional frequency vector $\boldsymbol{d}^R_{\boldsymbol{Y}|\boldsymbol{X}\in B}$ is defined by:

$$\boldsymbol{d}^R_{\boldsymbol{Y}|\boldsymbol{X}\in B}(\boldsymbol{x}) = |\pi_{\boldsymbol{Y}}\sigma_{\boldsymbol{X}=\boldsymbol{x}}R| \qquad \boldsymbol{x} \in B \tag{15}$$

As usual, we drop R and write $\boldsymbol{d}_{\boldsymbol{Y}|\boldsymbol{X}\in B}$ when R is clear from context.

If $|\boldsymbol{X}| = 1$, then the histogram is a 1D-histogram, otherwise it is a multidimensional histogram. In typical RDBMSs, there are about $k \approx 200$ buckets (e.g. MS SQL Server). Furthermore, the top 10 or so heavy hitters are each in a bucket by themselves. For 1D-histograms, the bucketization is done via *equi-depth* partitioning[29]. Multi-dimensional histograms are algorithically more complicated and require more space to store [46, 45].

▶ **Definition 6.** *A* histogrammed FM-constraint *(or HFM-constraint for short) is a tuple* $(\mathcal{B}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{c}, \ell, R)$*, where* $\boldsymbol{X} \subseteq \boldsymbol{Y}$ *are sets of variables,* $\mathcal{B}$ *is a partition of* $\mathsf{Dom}(\boldsymbol{X})$*,* $\boldsymbol{c} = (c_B)_{B\in\mathcal{B}}$ *is a vector of real numbers, and* $\ell \in [0, +\infty]$*. The constraint states that*

$$F_\ell(\boldsymbol{d}^R_{\boldsymbol{Y}|\boldsymbol{X}\in B}) \le c_B \qquad \forall B \in \mathcal{B} \tag{16}$$

Unlike in the histogram-free case, we can no longer use the ∞-moments to capture the 0- and 1-moments. Each moment is its own useful statistics.

3.3 A model-free cardinality estimator under HFM-constraints

This section explains how in an on-going work, in collaboration with Keshavarz and Nguyen [35], we were able to generalize Theorem 4 to the case when the input constraints are HFM-constraints. For the sake of clarity and to simplify the exposition, we will restrict the HFM-constraints to only 1D-histogram constraints.[2]

3.3.1 A highly simplified example

To illustrate the main ideas, we start with an example where the query is a simple join between two relations

$$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z).$$

We further assume that there are HFM-constraints on both R and S, where the bucketization on Y is *identical* on both R and S. In particular, suppose the input HFM-constraints are of the form:

- $(\mathcal{B}, Y, XY, \boldsymbol{r}, \infty, R)$, where $\boldsymbol{r} = (r_B)_{B\in\mathcal{B}}$
- $(\mathcal{B}, Y, XY, \boldsymbol{c}, 0, R)$, where $\boldsymbol{c} = (c_B)_{B\in\mathcal{B}}$
- $(\mathcal{B}, Y, YZ, \boldsymbol{s}, \infty, S)$, where $\boldsymbol{s} = (s_B)_{B\in\mathcal{B}}$

More concretely, the constraints state the following, for *every* $B \in \mathcal{B}$:

- given any $y \in B$, we have $|\pi_X \sigma_{Y=y} R| \le r_B$
- $|\sigma_{Y\in B} \pi_Y R| \le c_B$
- given any $y \in B$, we have $|\pi_X \sigma_{Y=y} S| \le s_B$

We now apply the entropy argument to upper-bound $|Q|$. The starting point is the traditional entropy argument. Fix a particular (but arbitrary) input, including relations R and S. Consider the uniform distribution on (X, Y, Z) chosen from the join $R(X, Y) \wedge S(Y, Z)$. (Note that we do not assume anything about input distribution; the uniformity considered here is only for mathematical reasoning purposes.) Next, we depart from the entropy argument used to prove the likes of Theorem 4. We add one more random variable to the joint distribution. Let $J \in \mathcal{B}$ be the categorical variable, where $J = B$ iff $Y \in B$. Define

$$p_B := \mathsf{Pr}[J = B] \qquad \boldsymbol{p} := (p_B)_{B\in\mathcal{B}} \tag{17}$$

Consider the joint distribution on (X, Y, Z, J). Let h be its entropy function. Then, $h \in \overline{\Gamma}^*_4$; and the following holds:

$$\log|Q| = h(XYZ) \qquad h(J \mid Y) = 0 \qquad h(J) = -\sum_i p_i \log p_i \qquad \|\boldsymbol{p}\|_1 = 1 \qquad \boldsymbol{p} \ge \boldsymbol{0} \tag{18}$$

[2] This is the typical case in all existing RDBMSs: by default multidimensional histograms are not built. Sometimes they are built on-demand, but they are not used in estimating the join sizes, only to estimate selectivities of filter conditions on the base-tables.

Furthermore, the following inequalities hold, for every $B \in \mathcal{B}$:

$$h(X \mid J = B) \le \log r_B \tag{19}$$
$$h(Y \mid J = B) \le \log c_B \tag{20}$$
$$h(Z \mid J = B) \le \log s_B. \tag{21}$$

We simplify the constraints above by aggregating them:

$$h(X \mid J) = \sum_B h(X \mid J = B) \cdot p_B \le \langle \log \boldsymbol{r}, \boldsymbol{p} \rangle \tag{22}$$
$$h(Y \mid J) = \sum_B h(Y \mid J = B) \cdot p_B \le \langle \log \boldsymbol{d}, \boldsymbol{p} \rangle \tag{23}$$
$$h(Z \mid J) = \sum_B h(Z \mid J = B) \cdot p_B \le \langle \log \boldsymbol{s}, \boldsymbol{p} \rangle \tag{24}$$

Overall, we have the following optimization problem, which is our worst-case cardinality estimator for the example input:

$$\begin{aligned}
\max \quad & h(XYZ) && (25)\\
\text{s.t.} \quad & h(Y \mid J) \le \langle \boldsymbol{p}, \lg \boldsymbol{c} \rangle && (26)\\
& h(X \mid J) \le \langle \boldsymbol{p}, \lg \boldsymbol{r} \rangle && (27)\\
& h(Z \mid J) \le \langle \boldsymbol{p}, \lg \boldsymbol{s} \rangle && (28)\\
& h \in \overline{\Gamma}_4^* && (29)\\
& \boldsymbol{p} \ge 0, && (30)\\
& h(J) = -\langle \boldsymbol{p}, \lg \boldsymbol{p} \rangle && (31)\\
& h(J \mid Y) = 0 && (32)\\
& \|\boldsymbol{p}\|_1 = 1. && (33)
\end{aligned}$$

For this problem to be solvable, we replace $\overline{\Gamma}_4^*$ with Γ_4, which contains all the Shannon-type inequalities on h.

To show that the above optimization problem makes sense and is non-trivial, we now show that the estimator matches a combinatorial bound. We do so by applying only the constraints from the above optimization problem (with $\overline{\Gamma}_4^*$ replaced by Γ_4) to derive a bound on $|Q|$:

$$\begin{aligned}
\lg |Q| &= h(XYZ) && (34)\\
(\text{since } h(JY] = h(Y]) &= h(XYZJ) = h(XYZ|J) + h(J) && (35)\\
(\text{since } h \in \Gamma_4) &\le h(X|J) + h(Y|J) + h(Z|J) + h(J) && (36)\\
&\le \sum_{B \in \mathcal{B}} (\lg r_B + \lg c_B + \lg s_B - \lg p_B)) \cdot p_B && (37)\\
&= \sum_{B \in \mathcal{B}} (\lg(r_B c_B s_B / p_B)) \cdot p_B && (38)\\
(\text{Jensen}) &\le \lg \left(\sum_B r_B c_B s_B \right). && (39)
\end{aligned}$$

We arrive at the bound

$$|Q| \le \sum_{B \in \mathcal{B}} r_B c_B s_B \tag{40}$$

which is *exactly* what we would expect from the input conditions; furthermore, it is easy to see that the bound is tight!

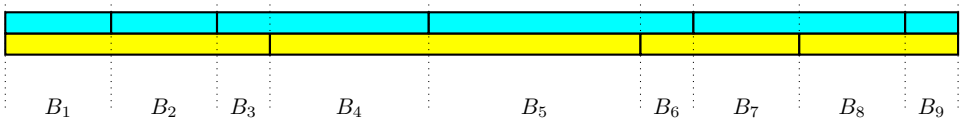

Figure 3 When boundaries of the blue-histograms and yellow histograms do not align.

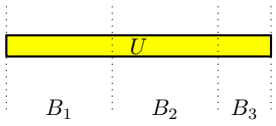

Figure 4 Refined partitioning of a yellow interval.

3.3.2 When histogram boundaries do not align

We now remove the unrealistic assumption that the histogram boundaries on R and S align perfectly. We assume the input HFM-constraints are

- $(\mathcal{B}', Y, XY, \boldsymbol{r}, \infty, R)$, where $\boldsymbol{r} = (r_{B'})_{B' \in \mathcal{B}'}$
- $(\mathcal{B}', Y, XY, \boldsymbol{c}, 0, R)$, where $\boldsymbol{c} = (c_{B'})_{B' \in \mathcal{B}'}$
- $(\mathcal{B}'', Y, YZ, \boldsymbol{s}, \infty, S)$, where $\boldsymbol{s} = (s_{B''})_{B'' \in \mathcal{B}''}$

The natural idea is to compute a refined partition $\mathcal{B}$ of $\mathsf{Dom}(Y)$, from intersecting the partitions $\mathcal{B}'$ and $\mathcal{B}''$, as shown in Figure 3. The issue is that we do not know the numbers r_B, c_B, and s_B; hence, we make them variables, and write down extra constraints to relate them to $r_{B'}$, $c_{B'}$, and $s_{B''}$.

The extra constraints depend on the norm-ℓ; for instance, suppose an interval $U \in \mathcal{B}'$ is refined into $B_1 \cup B_2 \cup B_3$ as shown in Figure 4, then we have two extra constraints:

$$c_{B_1} + c_{B_2} + c_{B_3} = c_U \tag{41}$$

$$\max\{r_{B_1}, r_{B_2}, r_{B_3}\} = r_U. \tag{42}$$

3.3.3 The general case

The above examples can be generalized as follows. Consider a collection HFM of HFM-constraints, for a query Q over variables $\boldsymbol{V}$. Recall that HFM-constraints are a strict super set of FM-constraints, which is a strict superset of DC-constraints, which is a strict superset of FD-constraints and cardinality constraints. An HFM-constraint $(\mathcal{B}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{c}, \ell, R)$ is an FM-constraint if $\mathcal{B} = \{\mathsf{Dom}(X)\}$, and an FM-constraint is a DC-constraint if $\ell = \infty$. An HFM-constraint is called *proper* if $|\mathcal{B}| > 1$ (i.e. it is a legitimate partition of $\mathsf{Dom}(\boldsymbol{X})$). Thanks to (14), if $\ell \in \{0, 1, \infty\}$, then we can assume that all non-proper HFM-constraints are DC-constraints. The size $|\boldsymbol{X}|$ is called the *dimensionality* of the HFM-constraint.

▶ **Definition 7.** *We call a collection $\mathcal{C}$ of HFM-constraints* simple *if the following conditions are met:*

- *Every constraint in* HFM *has* $\ell \in \{0, 1, +\infty\}$
- *Every proper HFM-constraint in* HFM *is one-dimensional (i.e.* $|\boldsymbol{X}| = 1$*)*

The idea to parallel Theorem 4 is to write down a set HC of all constraints, and formulate an optimization problem which mirrors the above examples and that of Theorem 4. We will assume that the input collection $\mathcal{C}$ of HFM-constraints are simple.

To start describing HC, we begin by assuming the "aligned boundary" case for all proper HFM-constraints. In particular, if there was at least one proper HFM-constraint $(\mathcal{B}, X, \boldsymbol{Y}, \boldsymbol{c}, \ell, R)$ on X, then *all* other constraints of the form $(\mathcal{B}', X, \boldsymbol{Y}', \boldsymbol{c}', \ell', R')$ on X, then *all* other constraints of the form must have $\mathcal{B}' = \mathcal{B}$.

For each such X, we add a new categorical variable J (to the joint distribution over $\boldsymbol{V}$), and a new vector $\boldsymbol{p} = (p_B)_{B\in\mathcal{B}}$, with the following constraints to HC:

$$h(J) = -\langle \boldsymbol{p}, \lg \boldsymbol{p}\rangle \qquad h(J \mid X) = 0 \qquad \|\boldsymbol{p}\|_1 = 1 \qquad \boldsymbol{p} \geq \boldsymbol{0} \tag{43}$$

Next, for each HFM-constraint $(\mathcal{B}, X, \boldsymbol{Y}, \boldsymbol{c}, \ell, R)$, we add the following constraints to HC:

$$\text{if } \ell = 0 \qquad h(X|J) \leq \langle \boldsymbol{p}, \lg \boldsymbol{d}\rangle \tag{44}$$

$$\text{if } \ell = 1 \qquad h(X\boldsymbol{Y}|J) \leq \langle \boldsymbol{p}, \lg \boldsymbol{d}\rangle \tag{45}$$

$$\text{if } \ell = \infty \qquad h(\boldsymbol{Y}|J) \leq \langle \boldsymbol{p}, \lg \boldsymbol{d}\rangle \tag{46}$$

The constraints for $\ell \in \{0, \infty\}$ were already explained in the example above (see (22), (23), (24)). The constraint for $\ell = 1$ also follows the same reasoning; the only difference is that the F_1-frequency moment counts the number of $(X, \boldsymbol{Y})$ tuples, and hence the bound is on $h(X\boldsymbol{Y}|J)$.

Now, when the boundaries of all the bucketizations $\mathcal{B}$ on the same variable X are *not* aligned, we do the following.

- Create the finest partition $\mathcal{B}$ of $\mathsf{Dom}(X)$ from taking the intersections of all the input bucketizations $\mathcal{B}'$ on $\mathsf{Dom}(X)$.
- For each input HFM-constraint $(\mathcal{B}', X, \boldsymbol{Y}, \boldsymbol{c}', \ell, R)$, create a new HFM-constraint $(\mathcal{B}, X, \boldsymbol{Y}, \boldsymbol{c}, \ell, R)$, where $\boldsymbol{c}$ are variables
- Add the following constraints on the quantities $\boldsymbol{c}$, depending on ℓ. Since $\mathcal{B}$ is a finer partition than $\mathcal{B}'$, every interval $U \in \mathcal{B}'$ is a union of some intervals $U = B_1 \cup \cdots \cup B_q$ in $\mathcal{B}$. We relate $c_{B_1}, \ldots, c_{B_q}$ to c'_U as follows.

$$\text{if } \ell \in \{0, 1\} \qquad c_{B_1} + \cdots + c_{B_k} \leq c'_U \tag{47}$$

$$\text{if } \ell = \infty \qquad c_{B_i} \leq c'_U \qquad \forall i \tag{48}$$

Let m denote the number of variables X for which there is a bucketization from $\mathcal{C}$ from, then the joint distribution we considered is on $n + m$ variables: $n = |\boldsymbol{V}|$, and there is one variable J for each such X. Thus, $h \in \overline{\Gamma}^*_{n+m}$. In addition, we have variables $\boldsymbol{p}$ for each X, and new variables $\boldsymbol{c}$ for each input HFM-constraint on which the bucketizations do not align. Together, the unknowns involve $(h, \boldsymbol{P}, \boldsymbol{C})$ where $\boldsymbol{P}$ collects all the unknowns $\boldsymbol{p}$, and $\boldsymbol{C}$ collects all the unknown $\boldsymbol{c}$. Let HC denote the list of all constraints. Then, we have the following.

▶ **Theorem 8** (From [35]). *Let Q be a conjunctive query and $\mathcal{C}$ be a given set of simple HFM-constraints, then for any database $\boldsymbol{D}$ satisfying $\mathcal{C}$, we have*

$$\sup_{\boldsymbol{D}\models\mathcal{C}} \log|Q(\boldsymbol{D})| \leq \max\{h(\boldsymbol{V}) \mid h \in \overline{\Gamma}^*_{n+m}, (h, \boldsymbol{P}, \boldsymbol{C}) \in \mathsf{HC}\} \quad \textit{(entropic bound)} \tag{49}$$

$$\leq \max\{h(\boldsymbol{V}) \mid h \in \Gamma_{n+m}, (h, \boldsymbol{P}, \boldsymbol{C}) \in \mathsf{HC}\} \quad \textit{(polymatroid bound)} \tag{50}$$

The (generalized) polymatroid bound (50) is our model-free estimator. It is possible to relax the simplicity assumption on $\mathcal{C}$, but the description of the abound will be much more involved. For the sake of clarify, we refrain from doing so here.

4 Research Questions

The approach we sketched in the previous section raises some very interesting and challenging research problems, with deep connections to information theory, linear programming, combinatorial optimization, and statistical estimation. This section outlines some research questions arising from our framework. We taxonomize the research questions in three broad categories:

- The first category involves questions surrounding the computability of the entropic bounds.
- The second category involves questions on how to compute the polymatroid bound efficiently.
- The third category aims to capture questions where probabilistic guarantees are taken into account, getting data-type specific information and the 2nd-moment ($\ell = 2$) constraints involved in the model.

4.1 Computability and information inequality

In order to compute the best (worst-case) cardinality estimate, we want to compute the entropic bound. For simplicity, let's start with the entropic bound (9) under degree constraints only. The constraints in HDC are linear constraints, and $\overline{\Gamma}_n^*$ is known to be a closed convex cone. Hence, the entropic bound is a *conic programming* problem [10] of the form:

$$\begin{aligned} \min \quad & \langle \boldsymbol{c}, \boldsymbol{h} \rangle && (51)\\ \text{s.t.} \quad & \boldsymbol{A}\boldsymbol{h} \le \boldsymbol{b} && (52)\\ & \boldsymbol{h} \in \overline{\Gamma}_n^*, && (53) \end{aligned}$$

where $c_{\boldsymbol{V}} = -1$ and $c_{\boldsymbol{X}} = 0$ for $\boldsymbol{X} \subset \boldsymbol{V}$. The inequalities in $\boldsymbol{A}\boldsymbol{h} \le \boldsymbol{b}$ come from the degree constraints: $h(\boldsymbol{Y}) - h(\boldsymbol{X}) \le N$. We write down a particular Lagrangian dual problem. To do so, associate dual variables $\boldsymbol{\delta}$ to the inequalities $\boldsymbol{A}\boldsymbol{h} \le \boldsymbol{b}$. The Lagrangian is

$$\mathcal{L}(\boldsymbol{\delta}) = \inf_{\boldsymbol{h} \in \overline{\Gamma}_n^*} \langle \boldsymbol{c}, \boldsymbol{h} \rangle + \langle \boldsymbol{A}\boldsymbol{h} - \boldsymbol{b}, \boldsymbol{\delta} \rangle = -\langle \boldsymbol{b}, \boldsymbol{\delta} \rangle + \inf_{\boldsymbol{h} \in \overline{\Gamma}_n^*} \langle \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta}, \boldsymbol{h} \rangle \tag{54}$$

Let $(\overline{\Gamma}_n^*)^*$ denote the *dual cone* of the cone $\overline{\Gamma}_n^*$.

- If $\boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta} \in \overline{\Gamma}_n^*$, then $\langle \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta}, \boldsymbol{h} \rangle \ge 0$ and thus $\inf_{\boldsymbol{h} \in \overline{\Gamma}_n^*} \langle \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta}, \boldsymbol{h} \rangle = 0$.
- If $\boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta} \notin \overline{\Gamma}_n^*$, then $\inf_{\boldsymbol{h} \in \overline{\Gamma}_n^*} \langle \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta}, \boldsymbol{h} \rangle = -\infty$.

Since the Lagrangian dual problem is to maximize $\mathcal{L}(\boldsymbol{\delta})$ subject to $\boldsymbol{\delta} \ge \boldsymbol{0}$, solving the above conic programming problem is essentially equivalent to solving the following dual:

$$\begin{aligned} \min \quad & \langle \boldsymbol{b}, \boldsymbol{\delta} \rangle && (55)\\ \text{s.t.} \quad & \boldsymbol{\delta} \ge \boldsymbol{0} && (56)\\ & \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta} \in (\overline{\Gamma}_n^*)^*. && (57) \end{aligned}$$

There is evidence that the dual problem (55) may not be decidable. Even checking for feasibility of a given solution seems hard. The reason is as follows. The statement $\boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta} \in (\overline{\Gamma}_n^*)^*$ is equivalent to

$$\langle \boldsymbol{c} + \boldsymbol{A}^\top \boldsymbol{\delta}, \boldsymbol{h} \rangle = \langle \boldsymbol{c}, \boldsymbol{h} \rangle + \langle \boldsymbol{A}\boldsymbol{h}, \boldsymbol{\delta} \rangle \ge 0 \qquad \forall h \in \overline{\Gamma}_n^*.$$

In particular, this is saying that the inequality

$$h(\boldsymbol{V}) \le \sum_{(X,Y) \in \mathsf{DC}} \delta_{Y|X} h(Y|X) \tag{58}$$

is a valid *information inequality*, i.e. an inequality that holds for all almost entropic functions. In joint work with Abo Khamis, Kolaitis, and Suciu [4], we studied this class of problems (of deciding the validity of information inequalities, and their generalization). While the

(un)decidability of these bounds are open, we were able to put them on the arithmetic hierarchy. Studying these inequalities are also closely related to (the decidability of) the problem of query containment under bag-semantic [4].

On the plus side, it was known [7] that the entropic bound under degree constraints is asymptotically tight! It is open, however, whether the generalization of the bound under HFM-constraints (49) is tight or not.

4.2 Computational complexity of the polymatroid bounds

Our next best hope is thus put on the polymatroid bound (10) and its histogrammed counterpart (50). These are optimization problems on an exponential number of variables and constraints.

Consider the simpler bound (10), under input DC constraints. On the negative news side, it is known [7] that the bound is not tight in general, namely there is a gap between the entropic bound (9) and the polymatroid bound on some input instances. In fact, one can construct a family of input instances for which the gap-ratio goes to infinity. Furthermore, as mentioned in [47], the exact computational complexity of computing the polymatroid bound (10) is open.

What gives us hope that the bound is computably tractable to begin with? After all, the linear program has an exponential number of variables and constraints. This is when some positive news emerge. We know that, under certain assumptions about the input degree constraints, we know that the polymatroid bound is not only computable in polynomial time, but also is tight (i.e. it is equal to the entropic bound:

- If DC contains only cardinality constraints, then we can show that [6] the polymatroid bound is exactly equal to the AGM bound [9]. One way to prove this is to use Lovasz "modularization" technique to show that one can replace polymatroids by *modular* polymatroids in the optimization problem while retaining the same objective value. The dual of the modular polymatroid optimization problem is exactly the AGM bound.
- There is one simple further relaxation we can make: if in addition to cardinality constraints, we have *simple* FDs, then the bound is also tight an computable in PTime. This fact was observed in [22] and generalized in [6] in terms of the lattice of FD closures. In particular, it was shown in [6] that if the FD-closure lattice is *distributive* [55], then the bound is tight and computable in PTime. (Simple FDs implies distributive FD-closure lattice.)
- Another case when the bound is tight and PTime-computable is when the set DC of input degree constraints is acyclic [47]. One can use this fact in another way, in order to obtain an upper-approximation of the bound: find a minimal subset of DC that is acyclic and use that as the approximation.
- Finally, recently in [28] we showed that if all degree constraints are *simple* then the bound is tight and PTime-computable. A degree constraint is simple if it is of the form (X, Y, N) with $|X| \leq 1$. In particular, all cardinality constraints are simple, all simple FDs are simple, and in addition we can also have proper degree constraints with $|X| = 1$. The fact that this bound is tight was shown in [4] where we showed that the polymatroids can be replaced by *normal* polymatroids in this case. Thanks to (7), this proves tightness of the bound. However, it still requires an exponential-sized description to describe the normal polymatroids N_n. In [28] we proved PTime-computability by characterizing the optimal solution using network flow analysis. We also proved that the bound is tight using a different strategy than what was used in [28]. Surprisingly, if the input DC are not (necessarily) simple, then computing the normal-polymatroid bound is NP-hard [28].

Next, consider the bound (50) under HFM-constraints. This is a concave maximization problem, which can be numerically solved [10]. An interesting research question is to find good upper-approximation to this bound that can be computed efficiently. Following the Jensen inequality strategy presented in Section 3.3.1, we can eliminate the extra variables J and $\boldsymbol{p}$; however, we are still studying how much we lose by this approximation [35].

4.3 Probabilistic guarantees

Thus far, we have not made use of the 2nd frequency-moment ($\ell = 2$) in the information theoretic framework. An interesting open problem is to devise a natural way to incorporate ℓ-moments for $\ell \notin \{0, 1, +\infty\}$ – especially $\ell = 2$.

One possible way to make use of $\ell = 2$ is to start incorporating probabilistic guarantees into our estimator (instead of a guaranteed upper-bound with probability 1). For example, consider the $R(\mathsf{actor}, \mathsf{movie}, \mathsf{role})$ relation and suppose we have $F_2(\boldsymbol{d}_{\mathsf{actor}})$ as part of $s(\boldsymbol{D})$. Suppose the query is $R(\mathsf{actor}, \mathsf{movie}, \mathsf{role}) \wedge \mathsf{actor} =$ "KevinBacon".

- In the System-R approach, the estimate will be $\frac{|R|}{F_0(\boldsymbol{d}_{\mathsf{actor}})}$. This is the average actor-degree, which would be an under estimate if "Kevin Bacon" was a heavy hitter in the table.
- In the information theoretic approach, before taking $\ell = 2$ into account, our estimate would have been $F_\infty(\boldsymbol{d}_{\mathsf{actor}})$, which is good for a heavy hitter, but would be an over estimate if "Kevin Bacon" is *not* a heavy hitter.
- Having $F_2(\boldsymbol{d}_{\mathsf{actor}})$ allows us to take a probabilistic compromise. Recall Cantelli's inequality, which says that, for *any* random variable X

$$\mathsf{Pr}\left[X \geq \mathbb{E}[X] + \lambda\right] \leq \frac{\mathsf{Var}[X]}{\mathsf{Var}[X] + \lambda^2}$$

Let X be the frequency (degree) of "Kevin Bacon". We do not know the distribution of X. To be as model-free as possible, we follow the *maximum-entropy principle* [33] and assume "Kevin Bacon" is uniformly distributed among all actors. Then,

$$\mathbb{E}[X] = \frac{|R|}{F_0(\boldsymbol{d}_{\mathsf{actor}})} \tag{59}$$

$$\mathsf{Var}[X] = \frac{F_2(\boldsymbol{d}_{\mathsf{actor}})}{F_0(\boldsymbol{d}_{\mathsf{actor}})} - \mathbb{E}[X]^2. \tag{60}$$

From these quantities and Cantelli's inequality, we can strike a balance between the traditional approach and our approach: we can guarantee that our upper-bound estimate is correct with a certain probabilistic threshold.

The idea of applying the maximum entropy principle has been used successfully in selectivity estimation [43]. It can also be used to deal with other types of information one typically record in database catalogs: the (non-frequency) moments of continuous variables. For example, the System-R estimator for the predicate $A > c$ is $\frac{\max_A - c}{\max_A - \min_A}$, as shown in Table 1. This assumes a uniform distribution over the interval $[\min_A, \max_A]$. However, if we also collect the empirical mean and variance of the variable, then assuming uniformity may not make sense for these given statistics. Instead, following the maximum entropy principle, we should fit an exponential family distribution to model and bound this estimate. Then, Cantelli or Chebyshev inequality can be used to give the probabilistic guarantee at the desired level.

5 Conclusions

We presented a recent effort at RelationalAI to formulate and devise a solution to the classic cardinality estimation problem in query optimization. Our approach aims to be model-free, or as model-free as possible, in order to avoid well-documented shortcomings of the traditional selectivity estimation approach. Our formulation is only for the offline case: no sampling nor learning was incorporated.

The approach is promising, as a variant of it is working well in production. There remain highly interesting and non-trivial open questions, as presented. We sincerely hope this presentation inspires more database theorists to study the problem. There are enough deep connections to information theory, algorithms, optimization, and statistics for long-term research programs.

References

1 MS SQL server documentation, 2021. URL: https://docs.microsoft.com/en-us/sql/relational-databases/performance/cardinality-estimation-sql-server?view=sql-server-ver15.

2 Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. Profiling relational data: a survey. *VLDB J.*, 24(4):557–581, 2015. doi:10.1007/s00778-015-0389-y.

3 Serge Abiteboul, Richard Hull, and Victor Vianu. *Foundations of Databases*. Addison-Wesley, 1995. URL: http://webdam.inria.fr/Alice/.

4 Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, and Dan Suciu. Bag query containment and information theory. In Dan Suciu, Yufei Tao, and Zhewei Wei, editors, *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2020, Portland, OR, USA, June 14-19, 2020*, pages 95–112. ACM, 2020. doi:10.1145/3375395.3387645.

5 Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, and Dan Suciu. Decision problems in information theory. In Artur Czumaj, Anuj Dawar, and Emanuela Merelli, editors, *47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference)*, volume 168 of *LIPIcs*, pages 106:1–106:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.ICALP.2020.106.

6 Mahmoud Abo Khamis, Hung Q. Ngo, and Dan Suciu. Computing join queries with functional dependencies. In Tova Milo and Wang-Chiew Tan, editors, *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, pages 327–342. ACM, 2016. doi:10.1145/2902251.2902289.

7 Mahmoud Abo Khamis, Hung Q. Ngo, and Dan Suciu. What do shannon-type inequalities, submodular width, and disjunctive datalog have to do with one another? In Emanuel Sallinger, Jan Van den Bussche, and Floris Geerts, editors, *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2017, Chicago, IL, USA, May 14-19, 2017*, pages 429–444. ACM, 2017. doi:10.1145/3034786.3056105.

8 Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In Gary L. Miller, editor, *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing, Philadelphia, Pennsylvania, USA, May 22-24, 1996*, pages 20–29. ACM, 1996. doi:10.1145/237814.237823.

9 Albert Atserias, Martin Grohe, and Dániel Marx. Size bounds and query plans for relational joins. In *FOCS*, pages 739–748. IEEE Computer Society, 2008. doi:10.1109/FOCS.2008.43.

10 Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. doi:10.1017/CBO9780511804441.

11 Walter Cai, Magdalena Balazinska, and Dan Suciu. Pessimistic cardinality estimation: Tighter upper bounds for intermediate join cardinalities. In Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska, editors, *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 18–35. ACM, 2019. doi:10.1145/3299869.3319894.

12 Yu Chen and Ke Yi. Random sampling and size estimation over cyclic joins. In Carsten Lutz and Jean Christoph Jung, editors, *23rd International Conference on Database Theory, ICDT 2020, March 30-April 2, 2020, Copenhagen, Denmark*, volume 155 of *LIPIcs*, pages 7:1–7:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020. doi:10.4230/LIPIcs.ICDT.2020.7.

13 Zhiyuan Chen, Flip Korn, Nick Koudas, and S. Muthukrishnan. Selectivity estimation for boolean queries. In Victor Vianu and Georg Gottlob, editors, *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, May 15-17, 2000, Dallas, Texas, USA*, pages 216–225. ACM, 2000. doi:10.1145/335168.335225.

14 F. R. K. Chung, R. L. Graham, P. Frankl, and J. B. Shearer. Some intersection theorems for ordered sets and graphs. *J. Combin. Theory Ser. A*, 43(1):23–37, 1986. doi:10.1016/0097-3165(86)90019-1.

15 Graham Cormode, Minos N. Garofalakis, Peter J. Haas, and Chris Jermaine. Synopses for massive data: Samples, histograms, wavelets, sketches. *Found. Trends Databases*, 4(1-3):1–294, 2012. doi:10.1561/1900000004.

16 Saumya K. Debray and Nai-Wei Lin. Static estimation of query sizes in horn programs. In Serge Abiteboul and Paris C. Kanellakis, editors, *ICDT'90, Third International Conference on Database Theory, Paris, France, December 12-14, 1990, Proceedings*, volume 470 of *Lecture Notes in Computer Science*, pages 514–528. Springer, 1990. doi:10.1007/3-540-53507-1_99.

17 Kyle Deeds, Dan Suciu, Magda Balazinska, and Walter Cai. Degree sequence bound for join cardinality estimation, 2022. arXiv:2201.04166.

18 Alin Dobra. Histograms revisited: when are histograms the best approximation method for aggregates over joins? In Chen Li, editor, *Proceedings of the Twenty-fourth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 13-15, 2005, Baltimore, Maryland, USA*, pages 228–237. ACM, 2005. doi:10.1145/1065167.1065196.

19 Cristian Estan and Jeffrey F. Naughton. End-biased samples for join cardinality estimation. In Ling Liu, Andreas Reuter, Kyu-Young Whang, and Jianjun Zhang, editors, *Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, 3-8 April 2006, Atlanta, GA, USA*, page 20. IEEE Computer Society, 2006. doi:10.1109/ICDE.2006.61.

20 Ehud Friedgut and Jeff Kahn. On the number of copies of one hypergraph in another. *Israel J. Math.*, 105:251–256, 1998. doi:10.1007/BF02780332.

21 Allen Van Gelder. Multiple join size estimation by virtual domains. In Catriel Beeri, editor, *Proceedings of the Twelfth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 25-28, 1993, Washington, DC, USA*, pages 180–189. ACM Press, 1993. doi:10.1145/153850.153872.

22 Georg Gottlob, Stephanie Tien Lee, Gregory Valiant, and Paul Valiant. Size and treewidth bounds for conjunctive queries. *J. ACM*, 59(3):16, 2012. doi:10.1145/2220357.2220363.

23 Peter J. Haas, Jeffrey F. Naughton, S. Seshadri, and Arun N. Swami. Fixed-precision estimation of join selectivity. In Catriel Beeri, editor, *Proceedings of the Twelfth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 25-28, 1993, Washington, DC, USA*, pages 190–201. ACM Press, 1993. doi:10.1145/153850.153875.

24 Peter J. Haas, Jeffrey F. Naughton, and Arun N. Swami. On the relative cost of sampling for join selectivity estimation. In Victor Vianu, editor, *Proceedings of the Thirteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 24-26, 1994, Minneapolis, Minnesota, USA*, pages 14–24. ACM Press, 1994. doi:10.1145/182591.182594.

25 Hazar Harmouch and Felix Naumann. Cardinality estimation: An experimental survey. *Proc. VLDB Endow.*, 11(4):499–512, 2017. doi:10.1145/3186728.3164145.

26 Axel Hertzschuch, Claudio Hartmann, Dirk Habich, and Wolfgang Lehner. Simplicity done right for join ordering. In *11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021, Online Proceedings*. www.cidrdb.org, 2021. URL: http://cidrdb.org/cidr2021/papers/cidr2021_paper01.pdf.

27 Dawei Huang, Dong Young Yoon, Seth Pettie, and Barzan Mozafari. Join on samples: A theoretical guide for practitioners. *Proc. VLDB Endow.*, 13(4):547–560, 2019. doi:10.14778/3372716.3372726.

28 Sungjin Im, Ben Moseley, Hung Q. Ngo, Kirk Pruhs, and Alireza Samadian. On the complexity of computing the polymatroid bound, 2022. Manuscript.

29 Yannis E. Ioannidis. The history of histograms (abridged). In Johann Christoph Freytag, Peter C. Lockemann, Serge Abiteboul, Michael J. Carey, Patricia G. Selinger, and Andreas Heuer, editors, *Proceedings of 29th International Conference on Very Large Data Bases, VLDB 2003, Berlin, Germany, September 9-12, 2003*, pages 19–30. Morgan Kaufmann, 2003. doi:10.1016/B978-012722442-8/50011-2.

30 Yannis E. Ioannidis and Stavros Christodoulakis. On the propagation of errors in the size of join results. In James Clifford and Roger King, editors, *Proceedings of the 1991 ACM SIGMOD International Conference on Management of Data, Denver, Colorado, May 29-31, 1991.*, pages 268–277. ACM Press, 1991. doi:10.1145/115790.115835.

31 H. V. Jagadish, Raymond T. Ng, and Divesh Srivastava. Substring selectivity estimation. In Victor Vianu and Christos H. Papadimitriou, editors, *Proceedings of the Eighteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, May 31 - June 2, 1999, Philadelphia, Pennsylvania, USA*, pages 249–260. ACM Press, 1999. doi:10.1145/303976.304001.

32 Sai Vikneshwar Mani Jayaraman, Corey Ropell, and Atri Rudra. Worst-case optimal binary join algorithms under general ℓ_p constraints. *CoRR*, abs/2112.01003, 2021. arXiv:2112.01003.

33 E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev. (2)*, 106:620–630, 1957.

34 Stasys Jukna. *Extremal combinatorics*. Texts in Theoretical Computer Science. An EATCS Series. Springer, Heidelberg, second edition, 2011. With applications in computer science. doi:10.1007/978-3-642-17364-6.

35 Hossein Keshavarz, Hung Q. Ngo, and XuanLong Nguyen. Output size bounds under histogrammed frequency moment constraints, 2022. Manuscript.

36 Andreas Kipf, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter A. Boncz, and Alfons Kemper. Learned cardinalities: Estimating correlated joins with deep learning. In *9th Biennial Conference on Innovative Data Systems Research, CIDR 2019, Asilomar, CA, USA, January 13-16, 2019, Online Proceedings*. www.cidrdb.org, 2019. URL: http://cidrdb.org/cidr2019/papers/p101-kipf-cidr19.pdf.

37 Nick Koudas, S. Muthukrishnan, and Divesh Srivastava. Optimal histograms for hierarchical range queries. In Victor Vianu and Georg Gottlob, editors, *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, May 15-17, 2000, Dallas, Texas, USA*, pages 196–204. ACM, 2000. doi:10.1145/335168.335223.

38 Viktor Leis, Andrey Gubichev, Atanas Mirchev, Peter A. Boncz, Alfons Kemper, and Thomas Neumann. How good are query optimizers, really? *Proc. VLDB Endow.*, 9(3):204–215, 2015. doi:10.14778/2850583.2850594.

39 Viktor Leis, Bernhard Radke, Andrey Gubichev, Alfons Kemper, and Thomas Neumann. Cardinality estimation done right: Index-based join sampling. In *8th Biennial Conference on Innovative Data Systems Research, CIDR 2017, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*. www.cidrdb.org, 2017. URL: http://cidrdb.org/cidr2017/papers/p9-leis-cidr17.pdf.

40 Viktor Leis, Bernhard Radke, Andrey Gubichev, Atanas Mirchev, Peter A. Boncz, Alfons Kemper, and Thomas Neumann. Query optimization through the looking glass, and what we found running the join order benchmark. *VLDB J.*, 27(5):643–668, 2018. doi:10.1007/s00778-017-0480-7.

41 Richard J. Lipton and Jeffrey F. Naughton. Query size estimation by adaptive sampling. In Daniel J. Rosenkrantz and Yehoshua Sagiv, editors, *Proceedings of the Ninth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, April 2-4, 1990, Nashville, Tennessee, USA*, pages 40–46. ACM Press, 1990. doi:10.1145/298514.298540.

42 G. Lohman. Is query optimization a solved problem?, 2014. URL: http://wp.sigmod.org/?p=1075.

43 Volker Markl, Peter J. Haas, Marcel Kutsch, Nimrod Megiddo, Utkarsh Srivastava, and Tam Minh Tran. Consistent selectivity estimation via maximum entropy. *VLDB J.*, 16(1):55–76, 2007. doi:10.1007/s00778-006-0030-1.

44 Guido Moerkotte, Thomas Neumann, and Gabriele Steidl. Preventing bad plans by bounding the impact of cardinality estimation errors. *Proc. VLDB Endow.*, 2(1):982–993, 2009. doi:10.14778/1687627.1687738.

45 M. Muralikrishna and David J. DeWitt. Equi-depth histograms for estimating selectivity factors for multi-dimensional queries. In Haran Boral and Per-Åke Larson, editors, *Proceedings of the 1988 ACM SIGMOD International Conference on Management of Data, Chicago, Illinois, USA, June 1-3, 1988*, pages 28–36. ACM Press, 1988. doi:10.1145/50202.50205.

46 S. Muthukrishnan, Viswanath Poosala, and Torsten Suel. On rectangular partitionings in two dimensions: Algorithms, complexity, and applications. In Catriel Beeri and Peter Buneman, editors, *Database Theory - ICDT '99, 7th International Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings*, volume 1540 of *Lecture Notes in Computer Science*, pages 236–256. Springer, 1999. doi:10.1007/3-540-49257-7_16.

47 Hung Q. Ngo. Worst-case optimal join algorithms: Techniques, results, and open problems. In Jan Van den Bussche and Marcelo Arenas, editors, *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, Houston, TX, USA, June 10-15, 2018*, pages 111–124. ACM, 2018. doi:10.1145/3196959.3196990.

48 Hung Q. Ngo, Ely Porat, Christopher Ré, and Atri Rudra. Worst-case optimal join algorithms: [extended abstract]. In Michael Benedikt, Markus Krötzsch, and Maurizio Lenzerini, editors, *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 37–48. ACM, 2012.

49 Hung Q. Ngo, Ely Porat, Christopher Ré, and Atri Rudra. Worst-case optimal join algorithms: [extended abstract]. In *PODS*, pages 37–48, 2012.

50 J. Radhakrishnan. Entropy and counting. In J. C. Misra, editor, *Computational Mathematics, Modelling and Algorithms*, pages 146–168. Narosa Pub House, 2003.

51 Raghu Ramakrishnan and Johannes Gehrke. *Database management systems (3. ed.)*. McGraw-Hill, 2003.

52 Christopher Ré and Dan Suciu. Understanding cardinality estimation using entropy maximization. In Jan Paredaens and Dirk Van Gucht, editors, *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2010, June 6-11, 2010, Indianapolis, Indiana, USA*, pages 53–64. ACM, 2010. doi:10.1145/1807085.1807095.

53 P. Griffiths Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. Access path selection in a relational database management system. In *Proceedings of the 1979 ACM SIGMOD international conference on Management of data*, SIGMOD '79, pages 23–34, New York, NY, USA, 1979. ACM.

54 Patricia G. Selinger, Morton M. Astrahan, Donald D. Chamberlin, Raymond A. Lorie, and Thomas G. Price. Access path selection in a relational database management system. In Philip A. Bernstein, editor, *Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data, Boston, Massachusetts, USA, May 30 - June 1*, pages 23–34. ACM, 1979. doi:10.1145/582095.582099.

55 Richard P. Stanley. *Enumerative combinatorics. Volume 1*, volume 49 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2012.

56 Todd L. Veldhuizen. Triejoin: A simple, worst-case optimal join algorithm. In Nicole Schweikardt, Vassilis Christophides, and Vincent Leroy, editors, *Proc. 17th International Conference on Database Theory (ICDT), Athens, Greece, March 24-28, 2014*, pages 96–106. OpenProceedings.org, 2014. doi:10.5441/002/icdt.2014.13.

57 Zongheng Yang, Eric Liang, Amog Kamsetty, Chenggang Wu, Yan Duan, Xi Chen, Pieter Abbeel, Joseph M. Hellerstein, Sanjay Krishnan, and Ion Stoica. Deep unsupervised cardinality estimation. *Proc. VLDB Endow.*, 13(3):279–292, 2019. doi:10.14778/3368289.3368294.

58 Raymond W. Yeung. *Information Theory and Network Coding*. Springer Publishing Company, Incorporated, 1 edition, 2008.

59 Zhen Zhang and Raymond W. Yeung. On characterization of entropy function via information inequalities. *IEEE Transactions on Information Theory*, 44(4):1440–1452, 1998. doi:10.1109/18.681320.