Join Size Bounds using ℓ_p-Norms on Degree Sequences

MAHMOUD ABO KHAMIS, RelationalAI, USA
VASILEIOS NAKOS, RelationalAI, USA
DAN OLTEANU, University of Zurich, Switzerland
DAN SUCIU, University of Washington, USA

Estimating the output size of a query is a fundamental yet longstanding problem in database query processing. Traditional cardinality estimators used by database systems can routinely underestimate the true output size by orders of magnitude, which leads to significant system performance penalty. Recently, upper bounds have been proposed that are based on information inequalities and incorporate sizes and max-degrees from input relations, yet their main benefit is limited to cyclic queries, because they degenerate to rather trivial formulas on acyclic queries.

We introduce a significant extension of the upper bounds, by incorporating ℓ_p-norms of the degree sequences of join attributes. Our bounds are significantly lower than previously known bounds, even when applied to acyclic queries. These bounds are also based on information theory, they come with a matching query evaluation algorithm, are computable in exponential time in the query size, and are provably tight when all degrees are "simple".

CCS Concepts: • **Information systems** → **Join algorithms**.

Additional Key Words and Phrases: query output cardinality; degree sequence; worst-case optimal join

ACM Reference Format:
Mahmoud Abo Khamis, Vasileios Nakos, Dan Olteanu, and Dan Suciu. 2024. Join Size Bounds using ℓ_p-Norms on Degree Sequences. *Proc. ACM Manag. Data* 2, 2 (PODS), Article 96 (May 2024), 24 pages. https://doi.org/10.1145/3651597

1 INTRODUCTION

Cardinality estimation is a central yet longstanding open problem in database systems. It allows query optimizers to select a query plan that minimizes the size of the intermediate results and therefore the necessary time and memory to compute the query. Yet traditional estimators present in virtually all database management systems routinely underestimate the true cardinality by orders of magnitude, which can lead to inefficient query plans [14, 17, 21].

The past two decades introduced *worst-case upper bounds* on the output size of a join query. The first such bound is the *AGM bound*, which is a function of the sizes of the input tables [5]. It was further refined in the presence of functional dependencies [3, 13]. A more general bound is the *PANDA bound*, which is a function of both the sizes of the input tables and the max degrees of attributes in these tables [4]. These are powerful methods as they can be applied to arbitrary

Authors' addresses: Mahmoud Abo Khamis, mahmoud.abokhamis@relational.ai, RelationalAI, 2120 University Ave, Berkeley, CA, 94704, USA; Vasileios Nakos, vasileios.nakos@relational.ai, RelationalAI, 2120 University Ave, Berkeley, CA, 94704, USA; Dan Olteanu, olteanu@ifi.uzh.ch, University of Zurich, Department of Informatics, Andreasstrasse 15, Zurich, 8050, Switzerland; Dan Suciu, suciu@cs.washington.edu, University of Washington, Department of Computer Science & Engineering, Box 352350, Seattle, WA, 98195-2350, USA.

ACM 2836-6573/2024/5-ART96
https://doi.org/10.1145/3651597

joins and compute *provable* upper bounds on the query output size, unlike traditional cardinality estimators which often severely underestimate the query output size [20].

However, these theoretical bounds have not had practical impact. One reason is that most queries in practice are acyclic queries, where upper bounds become trivial: they simply multiply the size of one relation with the *maximum* degrees of the joining relations. This is not new for a practitioner: standard estimators do the same, but use the *average* degrees instead of the max degrees. A second, related reason, is that they use essentially the same statistics as existing cardinality estimators: cardinalities and max or average degrees. There have been a few implementations under the name *pessimistic cardinality estimators* [7, 15], but their empirical evaluation showed that they remain less accurate than other estimators [8, 14].

In this paper, we introduce new upper bounds on the query output size that use ℓ_p-norms of degree sequences. The *degree sequence* of a graph is the sorted list of the degrees of the nodes, $d_1 \geq d_2 \geq \cdots$, where d_1 the largest degree, d_2 the next largest, etc. The ℓ_p-norm of a degree sequence is defined as $(d_1^p + d_2^p + \cdots)^{1/p}$. Our method computes an upper bound in terms of ℓ_p-norms of the degree sequences of the join columns; to the best of our knowledge, these are the first upper bounds that use arbitrary ℓ_p-norms on the relations. They strictly generalize previous bounds based on cardinalities and max-degrees [4], because the ℓ_1-norm of an attribute $R.A$ is the size $\sum_i d_i$ of R, and the ℓ_∞-norm is the max degree d_1 of A. However, our method can use any other norm, which leads to a much tighter upper bound. We follow the standard assumption in cardinality estimation, and assume that several ℓ_p-norms are pre-computed, and available during cardinality estimation.

Like the AGM [5] and the PANDA [4] bounds, our method relies on information inequalities. The computed bound is the optimal solution of a linear program, and can be computed in time exponential in the size of the query. Our method applies to arbitrary join queries (cyclic or not), but, unlike AGM and PANDA, it leads to completely new bounds even for acyclic queries, and uses new kinds of statistics, which makes it more likely for these theoretical bounds to have impact in practical scenarios.

1.1 A Motivating Example

The standard illustration for size upper bounds is the triangle query:

$$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \wedge T(Z, X), \tag{1}$$

for which the AGM bound [5] (based on the ℓ_1-norm) is:

$$|Q| \leq (|R| \cdot |S| \cdot |T|)^{1/2} \tag{2}$$

and the PANDA bound [4] (based on the ℓ_1 and ℓ_∞ norms) is:

$$|Q| \leq |R| \cdot ||\mathsf{deg}_S(Z|Y)||_\infty \tag{3}$$

where $\mathsf{deg}_S(Z|Y) = (d_1, d_2, \ldots, d_m)$ is the degree sequence of Y in S, more precisely d_i is the number of occurrences of the i'th most frequent value $Y = y$. If the ℓ_2- and ℓ_3-norms of the degree sequences are also available, then we can derive new upper bounds, for example:

$$|Q| \leq \left(||\mathsf{deg}_R(Y|X)||_2^2 \cdot ||\mathsf{deg}_S(Z|Y)||_2^2 \cdot ||\mathsf{deg}_T(X|Z)||_2^2\right)^{1/3} \tag{4}$$

$$|Q| \leq \left(||\mathsf{deg}_R(Y|X)||_3^3 \cdot ||\mathsf{deg}_S(Y|Z)||_3^3 \cdot |T|^5\right)^{1/6} \tag{5}$$

Assuming the $\ell_1, \ell_2, \ell_3, \ell_\infty$ norms are precomputed, then all formulas above give us upper bounds on the query output size, and we can take the minimal one; which one is the smallest depends on the actual data.

1.2 Problem Definition

Before we define the problem investigated in this paper, we introduce the class of queries and the statistics under consideration.

For a number n, let $[n] \stackrel{\text{def}}{=} \{1, 2, \ldots, n\}$. We use upper case X for variable names, and lower case x for values of these variables. We use boldface for sets of variables, e.g., $\boldsymbol{X}$, and of constants, e.g., $\boldsymbol{x}$.

A full conjunctive (or join) query is defined by:

$$Q(\boldsymbol{X}) = \bigwedge_{j \in [m]} R_j(\boldsymbol{Y}_j) \tag{6}$$

where $\boldsymbol{Y}_j$ is the tuple of variables in R_j and $\boldsymbol{X} = \bigcup_{j \in [m]} \boldsymbol{Y}_j$ is the set of $n \stackrel{\text{def}}{=} |\boldsymbol{X}|$ variables in the query Q.

For a relation S and subsets $\boldsymbol{U}, \boldsymbol{V}$ of its attributes, let $\mathsf{deg}_S(\boldsymbol{V}|\boldsymbol{U})$ be the degree sequence of $\boldsymbol{U}$ in the projection $\Pi_{\boldsymbol{UV}} S$. Formally, let $G \stackrel{\text{def}}{=} (\Pi_{\boldsymbol{U}}(S), \Pi_{\boldsymbol{V}}(S), E)$ be the bipartite graph whose edges E are all pairs $(\boldsymbol{u}, \boldsymbol{v}) \in \Pi_{\boldsymbol{UV}}(S)$. Then $\mathsf{deg}_S(\boldsymbol{V}|\boldsymbol{U}) \stackrel{\text{def}}{=} (d_1, d_2, \ldots, d_m)$ is the degree sequence of the $\boldsymbol{U}$-nodes of the graph.

Fix $\boldsymbol{X}$ a set of variables. An *abstract conditional*, or simply *conditional*, is an expression of the form $\sigma = (\boldsymbol{V}|\boldsymbol{U})$. We say that σ is *guarded* by a relation $R(\boldsymbol{Y})$ if $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{Y}$; then we write $\mathsf{deg}_R(\sigma) \stackrel{\text{def}}{=} \mathsf{deg}_R(\boldsymbol{V}|\boldsymbol{U})$. An *abstract statistics* is a pair $\tau = (\sigma, p)$, where $p \in (0, \infty]$. If $B \geq 1$ is a real number, then we call the pair (τ, B) a *concrete statistics*, and call (τ, b), where $b \stackrel{\text{def}}{=} \log B$, a *concrete log-statistics*. If R is a relation guarding σ, then we say that R *satisfies* (τ, B) if $||\mathsf{deg}_R(\sigma)||_p \leq B$. When $p = 1$ then the statistics is a cardinality assertion on $|\Pi_{\boldsymbol{UV}}(R)|$, and when $p = \infty$ then it is an assertion on the maximum degree. We write $\Sigma = \{\tau_1, \ldots, \tau_s\}$ for a set of abstract statistics, and $\boldsymbol{B} = \{B_1, \ldots, B_s\}$ for an associated set of real numbers; thus, every pair (τ_i, B_i) is a concrete statistics. We will call the pair $(\Sigma, \boldsymbol{B})$ a *set of (concrete) statistics*, and call $(\Sigma, \boldsymbol{b})$, where $b_i \stackrel{\text{def}}{=} \log B_i$, a set of concrete *log-statistics*. We say that Σ is guarded by a relational schema $\boldsymbol{R} = (R_1, \ldots, R_m)$ if every $\tau_i \in \Sigma$ has a guard R_{j_i}, and we say that a database instance $\boldsymbol{D} = (R_1^D, \ldots, R_m^D)$ *satisfies* the statistics $(\Sigma, \boldsymbol{B})$, denoted by $\boldsymbol{D} \models (\Sigma, \boldsymbol{B})$, if $||\mathsf{deg}_{R_{j_i}^D}(\sigma)||_{p_i} \leq B_{\tau_i}$ for all $\tau_i = (\sigma_i, p_i) \in \Sigma$, where R_{j_i} is the guard of σ_i. We can now state the problem investigated in this paper:

Problem 1. *Given a join query Q and a set of statistics $(\Sigma, \boldsymbol{B})$ guarded by the (schema of the) query Q, find a bound $U \in \mathbb{R}$ such that for all database instances $\boldsymbol{D}$, if $\boldsymbol{D} \models (\Sigma, \boldsymbol{B})$, then $|Q(\boldsymbol{D})| \leq U$.*

The bound U is *tight*, if there exists a database instance $\boldsymbol{D}$ such that $\boldsymbol{D} \models (\Sigma, \boldsymbol{B})$ and $U = O(|Q(\boldsymbol{D})|)$.

1.3 Main Results

We solve Problem 1 for arbitrary join queries Q, databases $\boldsymbol{D}$ with relations of arbitrary arities, and statistics $(\Sigma, \boldsymbol{B})$ consisting of arbitrary ℓ_p-norms of degree sequences. We make the following contributions.

Contribution 1: ℓ_p Bounds on Query Output Size. Our key observation is that the concrete statistics $||\mathsf{deg}(\boldsymbol{V}|\boldsymbol{U})||_p \leq B$ implies the following inequality in information theory:

$$\frac{1}{p} h(\boldsymbol{U}) + h(\boldsymbol{V}|\boldsymbol{U}) \leq \log B \tag{7}$$

where h is the entropy of some probability distribution on R (reviewed in Sec. 3). Using (7) we prove the following general upper bound on the size of the query's output. Note that in this bound, the query variables become random variables when we take a probability distribution over the query's output: (The distribution will be specified later in the proof.)

Theorem 1.1. *Let Q be a full conjunctive query (6), $\boldsymbol{U}_i, \boldsymbol{V}_i \subseteq \boldsymbol{X}$ be sets of variables, for $i \in [s]$, and suppose that the following information inequality is valid for all entropic vectors $\boldsymbol{h}$ with variables $\boldsymbol{X}$:*

$$\sum_{i \in [s]} w_i \left(\frac{1}{p_i} h(\boldsymbol{U}_i) + h(\boldsymbol{V}_i | \boldsymbol{U}_i) \right) \geq h(\boldsymbol{X}) \tag{8}$$

where $w_i \geq 0$, and $p_i \in (0, \infty]$, for all $i \in [s]$. Assume that each conditional $(\boldsymbol{V}_i | \boldsymbol{U}_i)$ in (8) is guarded by some relation R_{j_i} in Q. Then, for any database instance $\boldsymbol{D} = (R_1^D, R_2^D, \ldots)$, the following upper bound holds on the query output size:

$$|Q(\boldsymbol{D})| \leq \prod_{i \in [s]} ||\mathsf{deg}_{R^D_{j_i}}(\boldsymbol{V}_i | \boldsymbol{U}_i)||_{p_i}^{w_i} \tag{9}$$

We prove the theorem in Sec. 4. Thus, one approach to find an upper bound on the query output is to find an inequality of the form (8), prove it using Shannon inequalities, then conclude that (9) holds. For example, the bounds (4)-(5) stated in our motivating example follow from the following inequalities:

$$(h(X) + 2h(Y|X)) + (h(Y) + 2h(Z|Y)) + (h(Z) + 2h(X|Z)) \geq 3h(XYZ) \tag{10}$$

$$(h(X) + 3h(Y|X)) + (h(Z) + 3h(Y|Z)) + 5h(XZ) \geq 6h(XYZ) \tag{11}$$

These can be proven by observing that they are sums of basic Shannon inequalities (reviewed in Sec. 3):

$$Eq.\ (10) \text{ is sum of } \begin{cases} h(X) + h(Y|X) + h(Z|Y) \geq & h(XYZ) \\ h(Y) + h(Z|Y) + h(X|Z) \geq & h(XYZ) \\ h(Z) + h(X|Z) + h(Y|X) \geq & h(XYZ) \end{cases}$$

$$Eq.\ (11) \text{ is sum of } \begin{cases} 2h(XZ) + 2h(Y|X) \geq & 2h(XYZ) \\ 2h(XZ) + 2h(Y|Z) \geq & 2h(XYZ) \\ h(X) + h(Y|X) + h(Z) \geq & h(XYZ) \\ h(Y|Z) + h(XZ) \geq & h(XYZ) \end{cases}$$

Contribution 2: Asymptotically Tighter Cardinality Upper Bounds. The AGM and PANDA's bounds also rely on an information inequality, but use only ℓ_1 and ℓ_∞. Our novelty is the extension to ℓ_p norms. We show in Sec. 2.1 that this leads to significantly better bounds. Quite surprisingly, we are able to improve significantly the bounds even for acyclic queries, and even for a single join.

Preliminary experiments [2] with cyclic queries on the SNAP graph datasets [22] and with acyclic queries on the JOB benchmark [20] show that the upper bounds based on ℓ_p-norms can be orders of magnitude closer to the true cardinalities than the traditional cardinality estimators (e.g., used by DuckDB) and the theoretical upper bounds based on the ℓ_1 and ℓ_∞ norms only. To achieve the best upper bound with our method, a variety of norms are used in the experiments.

Contribution 3: New Algorithm Meeting the New Bounds. The celebrated Worst Case Optimal Join algorithm runs in time bounded by the AGM bound [23, 24]. A more complex algorithm [4] runs in time bounded by the PANDA bound. In Sec. 2.2 we describe an algorithm that runs in time bounded by our new ℓ_p-bounds. Any such algorithm must include PANDA's as a special case, because our bounds strictly generalize PANDA's. Our new algorithm in Sec. 2.2 consists of reducing the general case to PANDA. We do this by repeatedly partitioning each relation R such that a constraint on

$||\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_p$ can be replaced by two constraints, on $|\Pi_{\boldsymbol{U}}(R)|$ and $||\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_\infty$. The original query becomes a union of queries, one per combination of parts of different relations. The algorithm then evaluates each of these queries using PANDA's algorithm.

Contribution 4: Computing the bounds. One way to describe the solution to Problem 1 is as follows. Consider a set of statistics $(\Sigma, \boldsymbol{B})$. Any valid information inequality (8) implies some bound on the query output size, namely $|Q| \leq \prod_{i\in[s]} B_{j_i}^{w_i}$. The best bound is their *minimum*, over all valid inequalities (8); we denote the log of this minimum by *Log-U-Bound*. This describes the solution to Problem 1 as a minimization problem. This approach is impractical, because the number of valid inequalities is infinite. In Sec. 5 we describe an alternative, dual characterization of the upper bound, as a maximization problem, by considering the following quantity:

$$\textit{Log-L-Bound} = \sup_{\boldsymbol{h}\models(\Sigma,\boldsymbol{b})} h(\boldsymbol{X}) \tag{12}$$

where $\boldsymbol{X}$ is the set of all variables in the query Q, and $\boldsymbol{h}$ is required to "satisfy" the concrete log-statistics $(\Sigma, \boldsymbol{b})$, meaning that inequality (7) is satisfied for every statistics in Σ. Equation (12) defines a *maximization* problem. Our fourth contribution is:

Theorem 1.2 (Informal). *If $\boldsymbol{h}$ ranges over the same closed, convex cone K in both* (8) *and* (12)*, then Log-U-Bound = Log-L-Bound.*

We explain the theorem. K is used implicitly in (8) to define when the inequality is *valid*, namely when it holds $\forall \boldsymbol{h} \in K$, and also in (12), as the range of $\boldsymbol{h}$. The theorem says that, if K is topologically closed and convex, then the two quantities coincide. The special case of the theorem when $K \stackrel{\text{def}}{=} \Gamma_n$ is the set of *polymatroids* and (8) are the Shannon inequalities appeared implicitly in [4]; the general statement is new, and it includes the non-trivial case when $K \stackrel{\text{def}}{=} \bar{\Gamma}_n^*$ is the closure of entropic vectors and (8) are all entropic inequalities. To indicate which cone was used, we will use the subscript K in (12). Theorems 1.1 and 1.2 and the fact that $\bar{\Gamma}_n^* \subseteq \Gamma_n$ imply:

$$\log|Q| \leq \textit{Log-U-Bound}_{\bar{\Gamma}_n^*} \leq \textit{Log-U-Bound}_{\Gamma_n} \tag{13}$$

Theorem 1.2 has two important applications. First, it gives us an effective method for solving Problem 1, when (8) are restricted to Shannon inequalities, because in that case (12) is the optimal value of a linear program. Second, it allows us to study the tightness of the bound, by taking a deeper look at (13). We prove [2] that the *entropic bound*, $\textit{Log-U-Bound}_{\bar{\Gamma}_n^*}$, is asymptotically tight (which is a weaker notion than tightness), while, in general, the *polymatroid bound*, $\textit{Log-U-Bound}_{\Gamma_n}$, is not even asymptotically tight.

Contribution 5: Simple degree sequences. The tightness analysis leaves us with a dilemma: the entropic bound is tight but not computable, while the polymatroid bound is computable but not tight. We reconcile them in Sec. 6: For *simple degree sequences*, the two bounds coincide, i.e., they become equal. A degree sequence $\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})$ is *simple* if $|\boldsymbol{U}| \leq 1$. Moreover, in this case the bound is tight, in our usual sense: there exists a database $\boldsymbol{D}$ such that the size of the query output is $|Q(\boldsymbol{D})| \geq c \cdot 2^{\textit{Log-U-Bound}_{\Gamma_n}}$, where c is a constant that depends only on the query Q. The database $\boldsymbol{D}$ can be restricted to have a special form, called a *normal database.*

Closely related work. Jayaraman et al. [16] present a new algorithm for evaluating a query Q and prove a runtime in terms of ℓ_p-norms on degree sequences. Their result is limited to binary relations (thus all degrees are simple), to a single value p for a given query, and to queries with girth $\geq p + 1$. (The girth is the length of the minimal cycle.) While their work concerns only the algorithm, not a bound on the output, one can derive a bound from the runtime of the algorithm, since the output size cannot exceed the runtime. In the full version [2], we describe their bound explicitly, and show that it is a special case of our inequality (8). For example, for the triangle query (1) their runtime is (4), but they cannot derive (5), because the query graph has girth 3, hence they cannot use ℓ_3. The authors also notice that the worst-case instance is not always a product database, as in the AGM bound, but don't characterize it: our paper shows that this is always a *normal database*.

The Degree Sequence Bound (DSB) [9] is a tight upper bound of a query Q in terms of the degree sequences of its join attributes. The query Q is restricted to be Berge-acyclic, which also implies that all degree sequences are simple. There exists a 1-to-1 mapping between a degree sequence $d_1 \geq \cdots \geq d_m$ and its first m norms $\ell_1, \ldots, \ell_m$ (see Appendix A), therefore the DSB and our new bound could have access to the same information. Somewhat surprisingly, the DSB bound can be asymptotically better: the reason is that the 1-to-1 mapping is monotone only in one direction. We describe this analysis in Appendix B.1. In practice, both methods have access to fewer statistics than m: the DSB bound uses lossy compression [10], while our bound will have access to only a few ℓ_p-norms, making the two methods incomparable.

2 APPLICATIONS

Before we present the technical details of our results, we discuss two applications: cardinality estimation and query evaluation.

2.1 Cardinality Estimation

Our main intended application of Theorem 1.1 is for pessimistic cardinality estimation: given a query and statistics on the database, compute an upper bound on the query output size. A bound is good if it is as small as possible, i.e. as close as possible to the true output size. We follow the common assumption in cardinality estimation that the statistics are precomputed[1] and available at estimation time. For example the system may have precomputed the $\ell_2, \ell_5, \ell_\infty$-norms of $\mathsf{deg}_R(Y|X)$ and the ℓ_1, ℓ_{10}-norms of $\mathsf{deg}_S(Z|Y)$. We give several examples of upper bounds of the from (9) that improve significantly previously known bounds. For presentation purposes we describe all bounds in this section using (9). A system would instead rely on (12), i.e. it will compute the numerical value of the upper bound by optimizing a linear program, as we explain in Sec. 5. To reduce clutter, in this section we abbreviate $|Q(\boldsymbol{D})|$ with $|Q|$, and drop the superscript D from an instance R^D when no confusion arises.

Example 2.1. *As a warmup we start with a single join:*

$$Q(X, Y, Z) = R(X, Y) \wedge S(Y, Z) \tag{14}$$

Traditional cardinality estimators (as found in textbooks [25], see also [20]) use the formula

$$|Q| \approx \frac{|R| \cdot |S|}{\max(|\Pi_Y(R)|, |\Pi_Y(S)|)} \tag{15}$$

Since $\frac{|R|}{|\Pi_Y(R)|}$ *is the average degree of* $R(X|Y)$, (15) *is equivalent to*

$$|Q| \approx \min\left(|S| \cdot avg(\mathsf{deg}_R(X|Y)), |R| \cdot avg(\mathsf{deg}_S(Z|Y))\right) \tag{16}$$

Turning our attention to upper bounds, we note that the AGM bound is $|R| \cdot |S|$. *A better bound is the PANDA bound, which replaces avg with* max *in* (16)*:*

$$|Q| \leq \min\left(|S| \cdot ||\mathsf{deg}_R(X|Y)||_\infty, \ \ |R| \cdot ||\mathsf{deg}_S(Z|Y)||_\infty\right) \tag{17}$$

[1]It takes $O(N \log N)$ time to compute the degree sequence of an attribute X of a relation R of size N: sort R by X, group-by X, count, then sort again by the count.

Our framework derives several new upper bounds, by using ℓ_p-statistics other than ℓ_1 and ℓ_∞. We start with the simplest:

$$|Q| \leq ||\mathsf{deg}_R(X|Y)||_2 \cdot ||\mathsf{deg}_S(Z|Y)||_2 \tag{18}$$

The reader may notice that this inequality is Cauchy-Schwartz, but, in the framework of Th. 1.1, it follows from a Shannon inequality:

$$\frac{1}{2}\left(h(Y) + 2h(X|Y)\right) + \frac{1}{2}\left(h(Y) + 2h(Z|Y)\right) \geq h(XYZ)$$

The inequality can be simplified to $h(Y) + h(X|Y) + h(Z|Y) \geq h(XYZ)$, which holds because $h(Y) + h(X|Y) = h(XY)$, $h(Z|Y) \geq h(Z|XY)$, and $h(XY) + h(Z|XY) = h(XYZ)$; we review Shannon inequalities in Sec. 3. Depending on the data, (18) *can be asymptotically better than* (17)*. A simple example where this happens is when Q is a self-join, i.e. $R(X,Y) \wedge R(Z,Y)$. Then, the two degree sequences are equal, $\mathsf{deg}_R(X|Y) = \mathsf{deg}_R(Z|Y)$, and* (18) *becomes an equality, because $|Q| = ||\mathsf{deg}_R(X|Y)||_2^2$. Thus,* (18) *is exactly $|Q|$, while* (17) *continues to be an over approximation of $|Q|$, and can be asymptotically worse (see Appendix B.1).*

A more sophisticated inequality for the join query is the following, which holds for all $p, q \geq 0$ s.t. $\frac{1}{p} + \frac{1}{q} \leq 1$:

$$|Q| \leq ||\mathsf{deg}_R(X|Y)||_p^p \cdot ||\mathsf{deg}_S(Z|Y)||_q^{\frac{q}{p(q-1)}} |S|^{1-\frac{q}{p(q-1)}} \tag{19}$$

Depending on the concrete statistics on the data, this new bound can be much better than both (17) *and* (18)*. We prove this bound in Appendix B.1, where we also use this bound to study the connection between our ℓ_p-bounds on the Degree Sequence Bound in [9].*

The new bounds (18)-(19) *are just two examples, and other inequalities exist. In the full version [2], we provide some empirical evidence showing that, even for a single join, these new formulas indeed give better bounds on real data.*

Example 2.2. *In real applications most queries are acyclic. In the full version [2], we conducted some preliminary empirical evaluation on the JOB benchmark consisting of 33 acyclic queries over the IMDB real dataset, and found that the new ℓ_p-bounds are significantly better than both traditional estimators (e.g., used by DuckDB) and pessimistic estimators (AGM, PANDA). We give here a taste of how such a bound might look for a path query of length $n \geq 3$:*

$$Q(X_1, \ldots, X_n) = \bigwedge_{i \in [n-1]} R_i(X_i, X_{i+1})$$

Traditional cardinality estimators apply (15) *repeatedly; similarly PANDA relies on a straightforward extension of* (17)*. Our new approach leads, for example, to:*

$$|Q|^p \leq |R_1|^{p-2} \cdot ||\mathsf{deg}_{R_2}(X_1|X_2)||_2^2 \cdot \prod_{i=2,n-2} ||\mathsf{deg}_{R_i}(X_{i+1}|X_i)||_{p-1}^{p-1} \cdot ||\mathsf{deg}_{R_{n-1}}(X_n|X_{n-1})||_p^p$$

This bound holds for any $p \geq 2$, because of the following Shannon inequality:

$$\begin{aligned}&(p-2)h(X_1X_2) + (h(X_2) + 2h(X_1|X_2)) + \sum_{i=2,n-2}(h(X_i) + (p-1)h(X_{i+1}|X_i)) \\ &+ (h(X_{n-1}) + ph(X_n|X_{n-1})) \geq ph(X_1 \ldots X_n)\end{aligned} \tag{20}$$

In particular, the above inequality holds because it is a sum of the following basic Shannon inequalities:

$$(p-2)h(X_1X_2) + \sum_{i=2,n-2}(p-2)h(X_{i+1}|X_i) + (p-2)h(X_n|X_{n-1}) \geq (p-2)h(X_1\dots X_n)$$
$$h(X_2) + h(X_1|X_2) + \sum_{i=2,n-2} h(X_{i+1}|X_i) + h(X_n|X_{n-1}) \geq h(X_1\dots X_n)$$
$$h(X_1|X_2) + \sum_{i=2,n-2} h(X_i) + h(X_{n-1}) + h(X_n|X_{n-1}) \geq h(X_1\dots X_n)$$

Several other bounds for the path query exist [2]. To our surprise, when we conducted our empirical evaluation in the full version [2], we found that the system used ℓ_p-norms from a wide range, $p \in \{1, 2, \dots, 29, \infty\}$. This shows the utility of having a large variety of ℓ_p-norm statistics for the purpose of cardinality estimation. It also raises a theoretical question: is it the case that, for every p, there exists a query/database, for which the optimal bound uses the ℓ_p-norm? We answer this positively next.

Example 2.3. *For every p, there exists a query and a database instance where the ℓ_p-norm on degree sequences leads to the best upper bound. Consider the cycle query of length $p+1$:*

$$Q(X_0, \dots, X_p) = R_0(X_0, X_1) \wedge \dots \wedge R_{p-1}(X_{p-1}, X_p) \wedge R_p(X_p, X_0)$$

For every $q \in [p]$, the following is an upper bound (generalizing (4)*):*

$$|Q| \leq \prod_{i=0,p} ||\mathsf{deg}_{R_i}(X_{(i+1)\,mod(p+1)}|X_i)||_q^{\frac{q}{q+1}} \tag{21}$$

To prove (21)*, we show the following Shannon inequality, where the arithmetic on the indices is taken modulo $p+1$, i.e. $i+1$ means $(i+1) \mod (p+1)$ etc:*

$$\sum_{i=0,p}(h(X_i) + qh(X_{i+1}|X_i)) \geq (q+1)h(X_0\dots X_p) \tag{22}$$

To prove the inequality, we proceed as follows. First, we observe that, for each $i = 0, p$, the following is a Shannon inequality:

$$h(X_i) + h(X_{i+1}|X_i) + \dots + h(X_{i+q}|X_{i+q-1}) \geq h(X_iX_{i+1}\dots X_{i+q})$$

All indices are taken modulo $p+1$, for example $i+q$ means $(i+q) \mod (p+1)$. Each inequality above can be easily checked. Next, we add up these $p+1$ inequalities, and make two observations. First, the sum of their LHS is precisely the LHS of (22)*. Second, after adding up their RHS, we use the following Shannon inequality $\sum_{i=0,p} h(X_i \dots X_{i+q}) \geq (q+1)h(X_0X_1\dots X_p)$ (which holds because each variable X_k occurs exactly $q+1$ times on the left, hence this is a Shearer-type inequality). Together, these observations prove* (22)*. We compare now the upper bound* (21) *to the AGM and PANDA bounds. To reduce the clutter we will assume that $R_0 = R_1 = \dots = R_p$. Then the AGM bound and the PANDA bounds are:*

$$|Q| \leq |R|^{\frac{p+1}{2}} \qquad\qquad |Q| \leq |R| \cdot ||\mathsf{deg}_R(Y|X)||_\infty^{p-1} \tag{23}$$

They follow from the following straightforward Shannon inequalities:

$$h(X_0X_1) + h(X_1X_2) + \dots + h(X_pX_0) \geq 2h(X_0X_1X_2\dots X_p)$$
$$h(X_0X_1) + h(X_2|X_1) + \dots + h(X_p|X_{p-1}) \geq h(X_0X_1X_2\dots X_p)$$

In Appendix B.2, we prove that, for any p, there exists a database instance where the bound (21) *for $q := p$ is the theoretically optimal bound that one can obtain by using the statistics on all $\ell_1, \ell_2, \ldots, \ell_p, \ell_\infty$-norms of all degree sequences.*

Example 2.4. *Previous examples used only binary relations. We illustrate here some examples with relations of higher arity. More precisely, we derive general upper bounds for the class of queries, called Loomis-Whitney, that have relational atoms with more than two join variables. A Loomis-Whitney query has n variables and n relational atoms, such that there is one atom for each set of $n-1$ variables. The triangle query is the Loomis-Whitney query with $n = 3$.*

The Loomis-Whitney query with $n = 4$ is:

$$Q(X, Y, Z, W) = A(X, Y, Z) \wedge B(Y, Z, W) \wedge C(Z, W, X) \wedge D(W, X, Y)$$

One bound that can be obtained with our framework is the following:

$$|Q|^4 \leq ||\mathsf{deg}_A(YZ|X)||_2^2 \cdot |B| \cdot ||\mathsf{deg}_C(WX|Z)||_2^2 \cdot |D|$$

This bound follows from the following information inequality:

$$4h(XYZW) \leq (h(X) + 2h(YZ|X)) + h(YZW) + (h(Z) + 2h(WX|Z)) + h(WXY)$$

The inequality holds because it is a sum of 4 Shannon inequalities:

$$\begin{aligned} h(XYZW) &\leq h(X) + h(YZ|X) + h(W|YZ) \\ h(XYZW) &\leq h(Z) + h(WX|Z) + h(Y|WX) \\ h(XYZW) &\leq h(YZ|X) + h(WX) \\ h(XYZW) &\leq h(WX|Z) + h(YZ) \end{aligned}$$

2.2 Query Evaluation

The second application is to query evaluation: we show that, if inequality (8) holds for all polymatroids, then we can evaluate the query in time bounded by (9) times a polylogarithmic factor in the data and an exponential factor in the sum of the p values of the statistics. Our algorithm generalizes the PANDA's algorithm [4] from ℓ_1 and ℓ_∞ norms to arbitrary norms. Recall that PANDA starts from an inequality of the form (8), where every p_i is either 1 or ∞, and computes $Q(\boldsymbol{D})$ in time $O\left(\prod_i B_i^{w_i}\right)$ if the database satisfies $|\Pi_{\boldsymbol{U}_i \boldsymbol{V}_i}(R_{j_i})| \leq B_i$ when $p_i = 1$ and $||\mathsf{deg}_{R_{j_i}}(\boldsymbol{V}_i|\boldsymbol{U}_i)||_\infty \leq B_i$ when $p_i = \infty$.

Our algorithm uses PANDA as a black box, as follows. It first partitions the relations on the join columns so that, within each partition, all degrees are within a factor of two, and each statistics defined by some ℓ_p-norm on the degree sequence of the join column can be expressed alternatively using only ℓ_1 and ℓ_∞. The original query becomes a union of queries, one per combination of parts of different relations. The algorithm then evaluates each of these queries using PANDA. We describe next the details of data partitioning and the reduction to PANDA.

Consider a relation R with attributes $\boldsymbol{X}$, and consider a concrete statistics (τ, B), where $\tau = ((\boldsymbol{V}|\boldsymbol{U}), p)$. We say that R *strongly satisfies* (τ, B), in notation $R \models_s (\tau, B)$, if there exists a number $d > 0$ such that $||\mathsf{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_\infty \leq d$ and $|\Pi_U(R)| \leq B^p/d^p$. If $R \models_s (\tau, B)$ then $R \models (\tau, B)$ because:

$$||\mathsf{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_p^p \leq |\Pi_U(R)| \cdot ||\mathsf{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_\infty^p \leq \frac{B^p}{d^p} d^p = B^p \tag{24}$$

In other words, R strongly satisfies the ℓ_p statistics (τ, B) if it satisfies an ℓ_1 and an ℓ_∞ statistics that imply (τ, B). We prove:

Lemma 2.5. *Fix a join query Q, and suppose that inequality* (8) *holds for all polymatroids. Let $\Sigma = \{(\boldsymbol{V}_i|\boldsymbol{U}_i, p_i) \mid i \in [s]\}$ be the abstract statistics and $w_i \geq 0$ be the coefficients in* (8). *If a database*

$\boldsymbol{D}$ strongly satisfies the concrete statistics $(\Sigma, \boldsymbol{B})$, then the query output $Q(\boldsymbol{D})$ can be computed in time $O\left(\prod_{i\in[s]} B_i^{w_i} \text{polylog } N\right)$, where N is the size of the active domain of $\boldsymbol{D}$.

Proof. Since $\boldsymbol{D}$ strongly satisfies the concrete statistics $(\Sigma, \boldsymbol{B})$, we can use (24) and replace each ℓ_p-statistics with an ℓ_1 and an ℓ_∞ statistics. We write B_i as $B_i = B_{i,1}^{\frac{1}{p}} \cdot B_{i,\infty}$, such that both $|\Pi_{U_i}(R_{j_i}^D)| \leq B_{i,1}$ and $||\text{deg}_{R_{j_i}}(\boldsymbol{V}|\boldsymbol{U})||_\infty \leq B_{i,\infty}$ hold. Expand the inequality (8) to $\sum_i \frac{w_i}{p_i} h(\boldsymbol{U}_i) + \sum_i w_i h(\boldsymbol{V}_i|\boldsymbol{U}_i) \leq h(\boldsymbol{X})$. This can be viewed as an inequality of the form (8) with $2s$ terms, where half of the terms have $p_i = 1$ and the others have $p_i = \infty$. Therefore, PANDA's algorithm can use this inequality and run in time: and

$$O\left(\prod_{i\in[s]} B_{i,1}^{\frac{w_i}{p_i}} \cdot \prod_{i\in[s]} B_{i,\infty}^{w_i} \cdot \text{polylog } N\right) = O\left(\prod_{i\in[s]} B_i^{w_i} \text{polylog } N\right)$$

□

In order to use the lemma, we prove the following:

Lemma 2.6. *Let R be a relation that satisfies an ℓ_p-statistics, $R \models (((\boldsymbol{V}|\boldsymbol{U}), p), B)$. Then we can partition R into $\lceil 2^p \rceil \log N$ disjoint relations, $R = R_1 \cup R_2 \cup \ldots$, such that each R_i strongly satisfies the ℓ_p-statistics, $R_i \models_s (((\boldsymbol{V}|\boldsymbol{U}), p), B)$.*

Proof. By assumption, $||\text{deg}_R \boldsymbol{V}|\boldsymbol{U}||_p^p \leq B^p$. First, partition R into $\log N$ buckets $R_i, i = 1, \ldots, \lceil \log N \rceil$, where R_i contains the tuples t whose $\boldsymbol{U}$-component $\boldsymbol{u}$ satisfies:

$$2^{i-1} \leq \text{deg}_R(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) = \text{deg}_{R_i}(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \leq 2^i$$

Then $|\Pi_{\boldsymbol{U}}(R_i)| \leq B^p / 2^{p(i-1)}$, because:

$$B^p \geq ||\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_p^p \geq ||\text{deg}_{R_i}(\boldsymbol{V}|\boldsymbol{U})||_p^p \geq |\Pi_{\boldsymbol{U}}(R_i)| \cdot 2^{p(i-1)}$$

Second, partition R_i into $\lceil 2^p \rceil$ sets $R_{i,1}, R_{i,2}, \ldots$ such that $|\Pi_{\boldsymbol{U}}(R_i)| \leq B^p / 2^{pi}$. Then, each $R_{i,j}$ strongly satisfies the concrete statistics $(((\boldsymbol{V}|\boldsymbol{U}), p), B)$, and their union is R. □

Our discussion implies:

Theorem 2.7. *There exists an algorithm that, given a join query Q, an inequality* (8) *that holds for all polymatroids, and a database $\boldsymbol{D}$ satisfying the concrete statistics $(\Sigma, \boldsymbol{B})$, computes the query output $Q(\boldsymbol{D})$ in time $O\left(c \cdot \prod_{i\in[s]} B_i^{w_i} \text{polylog} N\right)$; here $c = \prod_{i\in[s]} \lceil 2^{p_i} \rceil$, where $p_1, \ldots, p_s$ are the norms occurring in Σ.*

Proof. Using Lemma 2.6, for each ℓ_{p_i}-norm, we partition $\boldsymbol{D}$ into a union of 2^{p_i} databases $\boldsymbol{D}_1 \cup \boldsymbol{D}_2 \cup \ldots$, where each $\boldsymbol{D}_j$ strongly satisfies $(\Sigma, \boldsymbol{B})$. Resolving s such norms like this partitions $\boldsymbol{D}$ into c parts. We then apply Lemma 2.5 to each part. □

3 BACKGROUND ON INFORMATION THEORY

Consider a finite probability space (D, P), where $P : D \to [0, 1]$, $\sum_{x\in D} P(x) = 1$, and denote by X the random variable with outcomes in D. The *entropy* of X is:

$$H(X) \stackrel{\text{def}}{=} -\sum_{x\in D} P(x) \log P(x) \tag{25}$$

If $N \stackrel{\text{def}}{=} |D|$, then $0 \leq H(X) \leq \log N$, the equality $H(X) = 0$ holds iff X is deterministic, and $H(X) = \log N$ holds iff X is uniformly distributed. Given n jointly distributed random variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$, we denote by $\boldsymbol{h} \in \mathbb{R}_+^{2^{[n]}}$ the following vector: $h_\alpha \stackrel{\text{def}}{=} H(\boldsymbol{X}_\alpha)$ for $\alpha \subseteq [n]$, where

$\boldsymbol{X}_\alpha$ is the joint random variable $(X_i)_{i\in\alpha}$, and $H(\boldsymbol{X}_\alpha)$ is its entropy; such a vector $\boldsymbol{h} \in \mathbb{R}_+^{2^{[n]}}$ is called *entropic*. We will blur the distinction between a vector in $\mathbb{R}_+^{2^{[n]}}$, a vector in $\mathbb{R}_+^{2^{\boldsymbol{X}}}$, and a function $2^{\boldsymbol{X}} \to \mathbb{R}_+$, and write interchangeably $\boldsymbol{h}_\alpha$, $\boldsymbol{h}_{\boldsymbol{X}_\alpha}$, or $h(\boldsymbol{X}_\alpha)$. A *polymatroid* is a vector $\boldsymbol{h} \in \mathbb{R}_+^{2^{[n]}}$ that satisfies the following *basic Shannon inequalities*:

$$h(\emptyset) = 0 \tag{26}$$

$$h(\boldsymbol{U} \cup \boldsymbol{V}) \geq h(\boldsymbol{U}) \tag{27}$$

$$h(\boldsymbol{U}) + h(\boldsymbol{V}) \geq h(\boldsymbol{U} \cup \boldsymbol{V}) + h(\boldsymbol{U} \cap \boldsymbol{V}) \tag{28}$$

The last two inequalities are called called *monotonicity* and *submodularity* respectively.

For any set $\boldsymbol{V} \subseteq \{X_1, \ldots, X_n\}$, the *step function* $\boldsymbol{h}^{\boldsymbol{V}}$ is:

$$h^{\boldsymbol{V}}(\boldsymbol{U}) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } \boldsymbol{V} \cap \boldsymbol{U} \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

There are $2^n - 1$ non-zero step functions (since $\boldsymbol{h}^{\emptyset} \equiv 0$). A *normal polymatroid* is a positive linear combination of step functions. When $\boldsymbol{V}$ is a singleton set, $\boldsymbol{V} = \{X_i\}$ for some $i = 1, n$, then we call $\boldsymbol{h}^{X_i}$ a *basic modular function*. A *modular* function is a positive linear combination of $\boldsymbol{h}^{X_1}, \ldots, \boldsymbol{h}^{X_n}$. The following notations are used in the literature: M_n is the set of modular functions, N_n is the set of normal polymatroids, Γ_n^* is the set of entropic vectors, $\bar{\Gamma}_n^*$ is its topological closure, and Γ_n is the set of polymatroids. It is known that $M_n \subset N_n \subset \Gamma_n^* \subset \bar{\Gamma}_n^* \subset \Gamma_n \subset \mathbb{R}_+^{2^{[n]}}$, that M_n, N_n, Γ_n are polyhedral cones, $\bar{\Gamma}_n^*$ is a closed, convex cone, and Γ_n^* is not a cone.[2]

The *conditional* of a vector $\boldsymbol{h}$ is defined as:

$$h(\boldsymbol{V}|\boldsymbol{U}) \stackrel{\text{def}}{=} h(\boldsymbol{U}\boldsymbol{V}) - h(\boldsymbol{U})$$

where $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$. If $\boldsymbol{h}$ is a polymatroid, then $h(\boldsymbol{V}|\boldsymbol{U}) \geq 0$. If $\boldsymbol{h}$ is entropic and realized by some probability distribution, then:

$$h(\boldsymbol{V}|\boldsymbol{U}) = \mathop{\mathbb{E}}_{\boldsymbol{u}}[h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})] \tag{30}$$

where $h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u})$ is the standard entropy of the random variable $\boldsymbol{V}$ conditioned on $\boldsymbol{U} = \boldsymbol{u}$.

An *information inequality* is a linear inequality of the form:

$$\boldsymbol{c} \cdot \boldsymbol{h} \geq 0 \tag{31}$$

where $\boldsymbol{c} \in \mathbb{R}^{2^{[n]}}$. Given a set $K \subseteq \mathbb{R}_+^{2^{[n]}}$, we say that the inequality is *valid for* K if it holds for all $\boldsymbol{h} \in K$; in that case we write $K \models \boldsymbol{c} \cdot \boldsymbol{h} \geq 0$. *Entropic inequalities* are those valid for Γ^* or, equivalently, for $\bar{\Gamma}_n^*$: it is an open problem whether they are decidable. *Shannon inequalities* are those valid for Γ_n and are decidable in exponential time.

4 PROOF OF THEOREM 1.1

In this section we prove Theorem 1.1, by showing that the information inequality (8) implies an upper bound on the query output size. The crux of the proof is inequality (7), which we prove below in Lemma 4.1. It establishes a new connection between information measures and the ℓ_p-norm, Eq. (34) below.

We briefly review connections that are known between database statistics and information measures. Let R be a relation instance with attributes $\boldsymbol{X}$ and with N tuples. Let $P : R \to [0, 1]$ be any probability distribution whose outcome consists of the tuples in R, in particular $\sum_{t\in R} P(t) = 1$,

[2]We refer to [6] for the definitions.

and let $h : 2^{\boldsymbol{X}} \rightarrow \mathbb{R}_+$ be its entropic vector. The following two inequalities connect $\boldsymbol{h}$ to statistics on R:

$$\forall \boldsymbol{V} \subseteq \boldsymbol{X} : \qquad h(\boldsymbol{V}) \leq \log |\Pi_{\boldsymbol{V}}(R)| \tag{32}$$

$$\forall \boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X} : \qquad h(\boldsymbol{V}|\boldsymbol{U}) \leq \log ||\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_\infty \tag{33}$$

Eq. (33) follows from (32), from the fact that, for all $\boldsymbol{u} \in \Pi_{\boldsymbol{U}}(R)$,

$$h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \leq \log \text{deg}_R(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}) \leq \log \max_{\boldsymbol{u}'} \text{deg}_R(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}') = \log ||\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_\infty$$

and from (30). In addition to these two connections, Lee [19] also proved a connection between conditional mutual information and multivalued dependencies, which is unrelated to our paper. We prove here a new connection:

Lemma 4.1. *With the notations above, the following holds:*

$$\forall p \in (0, \infty] : \qquad \frac{1}{p} h(\boldsymbol{U}) + h(\boldsymbol{V}|\boldsymbol{U}) \leq \log ||\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_p \tag{34}$$

Proof. When $p = \infty$, then (34) becomes (33), so we can assume $p < \infty$ and rewrite (34) to:

$$h(\boldsymbol{U}) + p h(\boldsymbol{V}|\boldsymbol{U}) \leq \log ||\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_p^p$$

Assume that $\Pi_{\boldsymbol{U}}(R)$ has N distinct values $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$, and that each $\boldsymbol{u}_i$ occurs with d_i distinct values $\boldsymbol{V} = \boldsymbol{v}$. In particular, $\text{deg}_R(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}_i) = d_i$. Let $P_i \stackrel{\text{def}}{=} P(\boldsymbol{U} = \boldsymbol{u}_i)$ be the marginal probability of $\boldsymbol{u}_i$. We use the definition of the entropy (25) and the formula for the conditional (30) and derive:

$$\begin{aligned} h(\boldsymbol{U}) + p h(\boldsymbol{V}|\boldsymbol{U}) &= \sum_i P_i \log \frac{1}{P_i} + p \sum_i P_i h(\boldsymbol{V}|\boldsymbol{U} = \boldsymbol{u}_i) \leq \sum_i P_i \log \frac{1}{P_i} + p \sum_i P_i \log d_i \\ &= \sum_i P_i \log \frac{d_i^p}{P_i} \leq \log \left(\sum_i P_i \frac{d_i^p}{P_i} \right) = \log \sum_i d_i^p = \log ||\text{deg}_R(\boldsymbol{V}|\boldsymbol{U})||_p^p \end{aligned}$$

where the last inequality is Jensen's inequality. □

Proof of Theorem 1.1. Assume that inequality (8) holds for all entropic vectors $\boldsymbol{h}$. Fix a database instance $\boldsymbol{D} = (R_1, \ldots, R_m)$.

Consider the uniform probability distribution over the output $Q(\boldsymbol{D})$, and let $\boldsymbol{h}$ be its entropic vector. By uniformity, $h(\boldsymbol{X}) = \log |Q(\boldsymbol{D})|$. By assumption, every conditional term $h(\boldsymbol{V}_i|\boldsymbol{U}_i)$, $i \in [s]$ that occurs in (8) has a witness R_{j_i}. From Lemma 4.1, we have

$$\frac{1}{p_i} h(\boldsymbol{U}_i) + h(\boldsymbol{V}_i|\boldsymbol{U}_i) \leq \log ||\text{deg}_{R_{j_i}}(\boldsymbol{V}_i|\boldsymbol{U}_i)||_{p_i}$$

Using inequality (8), we derive:

$$\log |Q(\boldsymbol{D})| = h(\boldsymbol{X}) \leq \sum_{i \in [s]} w_i \left(\frac{1}{p_i} h(\boldsymbol{U}_i) + h(\boldsymbol{V}_i|\boldsymbol{U}_i) \right) \leq \sum_{i \in [s]} w_i \log ||\text{deg}_{R_{j_i}}(\boldsymbol{V}_i|\boldsymbol{U}_i)||_{p_i}$$

This immediately implies the upper bound (9). □

5 COMPUTING THE BOUND

In this section we prove Theorem 1.2. Recall that the main problem in our paper, problem 1, asks for an upper bound to the query's output, given a set of concrete statistics on the database. So far we have proven Theorem 1.1, which says that, for any valid information inequality of the form (8), we can infer *some* bound. The best bound is their minimum, over all valid inequalities (8), and depends on the concrete statistics of the database. In this section we describe how to compute the best bound, by using the dual of information inequalities.

Given a vector $\boldsymbol{h} \in \mathbb{R}_+^{2^{[n]}}$ an abstract conditional $\sigma = (\boldsymbol{V}|\boldsymbol{U})$, and an abstract statistics $\tau = (\sigma, p)$, we denote by:

$$h(\sigma) \stackrel{\text{def}}{=} h(\boldsymbol{V}|\boldsymbol{U}) \qquad\qquad h(\tau) \stackrel{\text{def}}{=} \frac{1}{p}h(\boldsymbol{U}) + h(\boldsymbol{V}|\boldsymbol{U}) \tag{35}$$

We say that a vector $\boldsymbol{h}$ *satisfies* a concrete log-statistics (τ, b) if $h(\tau) \leq b$. Similarly, $\boldsymbol{h} \in \mathbb{R}_+^{2^{[n]}}$ *satisfies* a set of concrete log-statistics $(\Sigma, \boldsymbol{b})$, in notation $\boldsymbol{h} \models (\Sigma, \boldsymbol{b})$, if $h(\tau_i) \leq b_i$ for all $\tau_i \in \Sigma, b_i \in \boldsymbol{b}$.

Definition 5.1. *If $\Sigma = \{\tau_1, \ldots, \tau_s\}$ is a set of abstract statistics, then a Σ-*inequality *is an information inequality of the form:*

$$\sum_{i \in [s]} w_i h(\tau_i) \geq h(\boldsymbol{X}) \tag{36}$$

where $w_i \geq 0$. Notice that (8) *in Theorem 1.1 is a Σ-inequality.*

For $K \subseteq \mathbb{R}_+^{2^{[n]}}$, the log-upper bound and log-lower bound of a set of log-statistics $(\Sigma, \boldsymbol{b})$ are:

$$\textit{Log-U-Bound}_K(\Sigma, \boldsymbol{b}) \stackrel{\textit{def}}{=} \inf_{\boldsymbol{w}: K \models \textit{Eq. } (36)} \sum_{i \in [s]} w_i b_i \tag{37}$$

$$\textit{Log-L-Bound}_K(\Sigma, \boldsymbol{b}) \stackrel{\textit{def}}{=} \sup_{\boldsymbol{h} \in K: \boldsymbol{h} \models (\Sigma, \boldsymbol{b})} h(\boldsymbol{X}) \tag{38}$$

Fix a query $Q(\boldsymbol{X}) = \bigwedge_j R_j(\boldsymbol{Y}_j)$ that guards Σ, and assume $K = \bar{\Gamma}_n^*$: by Theorem 1.1, if a database $\boldsymbol{D}$ satisfies the statistics $(\Sigma, \boldsymbol{B})$, then $\log |Q(\boldsymbol{D})| \leq \text{Log-U-Bound}_K$, but it is an open problem whether this bound is computable. On the other hand, Log-L-Bound_K is not a bound, but it has two good properties. First, when $K = \Gamma_n$, then Log-L-Bound_K *is* computable, as the optimal value of a linear program: we show this in Example 5.3. Second, when the optimal vector $\boldsymbol{h}^*$ of the maximization problem (38) is the entropy of some relation, then we can construct a "worst-case database instance" $\boldsymbol{D}$: we use this in Sec. 6. We prove that (37) and (38) are equal:

Theorem 5.2. *If K is any closed, convex cone, and $N_n \subseteq K \subseteq \Gamma_n$ then $\textit{Log-U-Bound}_K = \textit{Log-L-Bound}_K$.*

The special case of this theorem when $K = \Gamma_n$ was already implicit in [4]. The proof of the general case is more difficult, and we defer it to [2]. Both $\bar{\Gamma}_n^*$ and Γ_n are closed, convex cones, hence the theorem applies to both. We call the corresponding bounds the *almost-entropic bound* (when $K = \bar{\Gamma}_n^*$) and the *polymatroid bound* (when $K = \Gamma_n$) respectively.

There are two important applications of Theorem 5.2. First, it gives us an effective algorithm for computing the polymatroid bound, by computing the optimal value of a linear program: we used this method in all experiments in the full version [2]. We illustrate here with a simple example.

Example 5.3. *Consider the triangle query Q in* (1)*. Assume that we have the following statistics for the relations R, S, T: (a) their cardinalities, denoted by B_R, B_S, B_T, whose logarithms are b_R, b_S, b_T, (b) the ℓ_2-norms of all degree sequences: (c) the ℓ_3 norms of all degree sequences. Then the polymatroid bound* (38) *can be computed by optimizing the following linear program, with 8 variables*

$h(\emptyset), h(X), \ldots, h(XYZ)$:

$$
\begin{aligned}
&\textit{maximize } h(XYZ), \textit{ subject to:} && \\
&h(XY) \leq b_R,\ h(YZ) \leq b_S\ h(XZ) \leq b_T && \textit{// cardinality stats} \\
&\frac{1}{2}h(X) + h(Y|X) \leq b_{((Y|X),2)} \quad \ldots && \textit{// } \ell_2\textit{-norm stats} \\
&\frac{1}{3}h(X) + h(Y|X) \leq b_{((Y|X),3)} \quad \ldots && \textit{// } \ell_3\textit{-norms stats} \\
&h(X) + h(XYZ) \leq h(XY) + h(XZ) && \textit{// Shannon inequalities} \\
&h(Y) + h(XYZ) \leq h(XY) + h(YZ) && \textit{// i.e. } (26)\text{-}(28) \\
&\ldots &&
\end{aligned}
$$

The second application of Theorem 5.2 is that it allows us to reason about the tightness of the bounds. If we can convert the optimal $\boldsymbol{h}^*$ in the lower bound (38) into a database, then we have a worst-case instance witnessing the fact that the bound is tight. We show in the full version [2] that the almost-entropic bound is *asymptotically* tight (a weaker form of tightness), while the polymatroid bound is not tight. However, we show in the next section that the polymatroid bound is tight in the special case of simple degrees.

6 SIMPLE DEGREE SEQUENCES

Call a conditional $\sigma = (\boldsymbol{V}|\boldsymbol{U})$ *simple* if $|\boldsymbol{U}| \leq 1$; call a set of abstract statistics Σ *simple* if, for all $(\sigma, p) \in \Sigma$, σ is simple. Simple conditionals were introduced in [1] to study query containment under bag semantics. We prove here that, when all statistics are simple, then the polymatroid bound is tight, meaning that there exists a worst case database $\boldsymbol{D}$ such that the size $|Q(\boldsymbol{D})|$ of the query output is within a query-dependent constant of the polymatroid bound. Recall (Sec. 3) that N_n is the set of normal polymatroids.

Theorem 6.1. *If Σ is simple, then*

$$\textit{Log-U-Bound}_{N_n}(\Sigma, \boldsymbol{b}) = \textit{Log-U-Bound}_{\bar{\Gamma}_n^*}(\Sigma, \boldsymbol{b}) = \textit{Log-U-Bound}_{\Gamma_n}(\Sigma, \boldsymbol{b})$$

The proof relies on the following result in [1]:

Lemma 6.2. *[[1]] Let Σ be a simple set of LP-statistics. Consider the Σ-inequality* (36)*. Then the following are equivalent:*

- *Eq.* (36) *is valid for all $\boldsymbol{h} \in \Gamma_n$.*
- *Eq.* (36) *is valid for all $\boldsymbol{h} \in \bar{\Gamma}_n^*$.*
- *Eq.* (36) *is valid for all $\boldsymbol{h} \in N_n$.*

In other words, if the inequality (36) is simple, then it is valid for all polymatroids iff it is valid for all (almost-) entropic vectors, iff it is valid for all normal polymatroids. This immediately implies 6.1.

In the rest of the section, we will use Theorem 6.1 to prove that the polymatroid bound is tight. For that we prove a lemma. If $T(\boldsymbol{X})$ is any relation instance with attributes $\boldsymbol{X}$, then *its entropy*, $\boldsymbol{h}_T$, is the entropic vector defined by the uniform probability distribution on T. Call the relation T *totally uniform* if, for all $\boldsymbol{V} \subseteq \boldsymbol{X}$, the marginal distribution on $\Pi_{\boldsymbol{V}}(T)$ is also uniform. Equivalently, it is totally uniform if $\log|\Pi_{\boldsymbol{V}}(T)| = h_T(\boldsymbol{V})$ for all $\boldsymbol{V} \subseteq \boldsymbol{X}$. The lemma below proves that, if $\boldsymbol{h}$ is normal, then it can be approximated by the entropy of a totally uniform T, which we will call a *normal* relation. Recall from Sec. 3 that $\boldsymbol{h}$ is *normal* if it is a positive, linear combination of step functions:

$$\boldsymbol{h} = \sum_{\boldsymbol{V} \subseteq \boldsymbol{X}} \alpha_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V}} \tag{39}$$

where $\boldsymbol{\alpha_V} \geq 0$.

Lemma 6.3. *Let $\boldsymbol{h}$ be the normal polymatroid in (39), and let c is the number of non-zero coefficients $\alpha_{\boldsymbol{V}}$. Then there exist a totally uniform relation $T(\boldsymbol{X})$ such that $|T| \geq \frac{1}{2^c}2^{h(\boldsymbol{X})}$, whose entropy $\boldsymbol{h}_T$ satisfies $\forall \boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$, $h_T(\boldsymbol{V}|\boldsymbol{U}) \leq h(\boldsymbol{V}|\boldsymbol{U})$.*

The lemma implies tightness of the polymatroid bound:

Corollary 6.4. *If all statistics in Σ are simple, the polymatroid bound U-Bound$_{\Gamma_n}$ ($\stackrel{def}{=} 2^{\text{Log-U-Bound}_{\Gamma_n}}$) is tight.*

Proof. Since N_n is polyhedral, we have:

$$\begin{aligned}
\text{Log-U-Bound}_{\Gamma_n} =& \text{Log-U-Bound}_{N_n} && \text{by Th. 6.1}\\
=& \text{Log-L-Bound}_{N_n} && \text{by Th. 5.2}\\
=& \max_{\boldsymbol{h} \in N_n : (\Sigma, \boldsymbol{b}) \models \boldsymbol{h}} h(\boldsymbol{X}) && \text{by (38)}\\
=& h^*(\boldsymbol{X})
\end{aligned}$$

where $\boldsymbol{h}^* \in N_n$ is optimal solution to the maximization problem. Let $T(\boldsymbol{X})$ be the totally uniform relation given by Lemma 6.3. Define the database instance $\boldsymbol{D} = (R_1^D, \ldots, R_m^D)$ by setting $R_j^D \stackrel{\text{def}}{=} \Pi_{Y_j}(T)$, for $j = 1, m$. Then $\boldsymbol{D}$ satisfies the constraints $(\Sigma, \boldsymbol{B})$, because, by total uniformity:

$$\begin{aligned}
\log ||\mathsf{deg}_{R_{j_i}}(\boldsymbol{V}_i|\boldsymbol{U}_i)||_p^p =& \log\left(|\Pi_{\boldsymbol{U}_i}(R_{j_i})| \cdot \left(\text{avg}(\mathsf{deg}_{R_{j_i}}(\boldsymbol{V}_i|\boldsymbol{U}_i))\right)^p\right) = \log\left(|\Pi_{\boldsymbol{U}_i}(T)| \cdot \left(\frac{|\Pi_{\boldsymbol{U}_i\boldsymbol{V}_i}(T)|}{|\Pi_{\boldsymbol{U}_i}(T)|}\right)^p\right)\\
=& h_T(\boldsymbol{U}_i) + p h_T(\boldsymbol{V}_i|\boldsymbol{U}_i) \leq h^*(\boldsymbol{U}_i) + p h^*(\boldsymbol{V}_i|\boldsymbol{U}_i) \leq b_i
\end{aligned}$$

The corollary follows from $|Q(\boldsymbol{D})| = |T| \geq \frac{1}{2^c}2^{h^*(\boldsymbol{X})} = \frac{1}{2^c}\text{U-Bound}_{\Gamma_n}$. proving that the bound is tight. □

In the rest of the section we prove Lemma 6.3. Given two $\boldsymbol{X}$-tuples $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{x}' = (x_1', \ldots, x_n')$ their *domain product* is $\boldsymbol{x} \otimes \boldsymbol{x}' \stackrel{\text{def}}{=} ((x_1, x_1'), \ldots, (x_n, x_n'))$: it has the same *n* attributes, and each attribute value is a pair consisting of a value from $\boldsymbol{x}$ and a value from $\boldsymbol{x}'$. Given two relations $T(\boldsymbol{X}), T'(\boldsymbol{X})$, with the same attributes, their *domain product* is $T \otimes T' \stackrel{\text{def}}{=} \{\boldsymbol{x} \otimes \boldsymbol{x}' \mid \boldsymbol{x} \in T, \boldsymbol{x}' \in T'\}$. The following hold:

$$\begin{aligned}
|T \otimes T'| =& |T| \cdot |T'|\\
\boldsymbol{h}_{T \otimes T'} =& \boldsymbol{h}_T + \boldsymbol{h}_{T'}
\end{aligned} \tag{40}$$

Domain products were first introduced by Fagin [11] (under the name *direct product*), and appear under various names in [1, 12, 18].

Definition 6.5. *For $\boldsymbol{V} \subseteq \boldsymbol{X}$, the* basic normal relation $T_N^{\boldsymbol{V}}$ *is:*

$$T_N^{\boldsymbol{V}} \stackrel{def}{=} \{(\underbrace{k, \cdots, k}_{\textit{attributes in } \boldsymbol{V}}, \underbrace{0, \cdots, 0}_{\boldsymbol{X}-\boldsymbol{V}}) \mid k = 0, N-1\} \tag{41}$$

A normal relation *is a domain product of basic normal relations.*

Proposition 6.6. *(1) $T_N^{\boldsymbol{V}}$ is totally uniform. (2) Its entropy is $\boldsymbol{h}_{T_N^{\boldsymbol{V}}} = (\log N) \cdot \boldsymbol{h}^{\boldsymbol{V}}$, where $\boldsymbol{h}^{\boldsymbol{V}}$ is the step function.*

The proof is immediate and omitted. It follows that every normal relation is totally uniform, because $|\Pi_{\boldsymbol{V}}(T \otimes T')| = |\Pi_{\boldsymbol{V}}(T)| \cdot |\Pi_{\boldsymbol{V}}(T')| = 2^{h_T(\boldsymbol{V})} \cdot 2^{h_{T'}(\boldsymbol{V})} = 2^{h_{T\otimes T'}(\boldsymbol{V})}$, and the entropy of a normal relation is a normal polymatroid, because it is the sum of some step functions. We illustrate normal relations with an example.

Example 6.7. *The following is a basic normal relation:*

$$T_N^{X,Z} = \begin{array}{|ccc|} \hline X & Y & Z \\ \hline 0 & 0 & 0 \\ 1 & 0 & 1 \\ 2 & 0 & 2 \\ & \cdots & \\ N-1 & 0 & N-1 \\ \hline \end{array}$$

Its entropy is $(\log N)\boldsymbol{h}^{X,Z}$*. The following are normal relations:*

$$\begin{aligned} T_1 =& \{(i,j,k) \mid i,j,k \in [0:N-1]\} && = T_N^X \otimes T_N^Y \otimes T_N^Z \\ T_2 =& \{(i,i,i) \mid i \in [0:N-1]\} && = T_N^{XYZ} \\ T_3 =& \{(i,(i,j),j) \mid i,j \in [0:N-1]\} && = T_N^{X,Y} \otimes T_N^{Y,Z} \end{aligned}$$

Their cardinalities are $|T_1| = N^3$, $|T_2| = N$, $|T_3| = N^2$.

Proof. (of Lemma 6.3) Fix a normal polymatroid $\boldsymbol{h}$ given by (39). For each $\boldsymbol{V} \subseteq \boldsymbol{X}$, define $\beta_{\boldsymbol{V}} \stackrel{\text{def}}{=} \log\lfloor 2^{\alpha_V} \rfloor$. Then 2^{β_V} is an integer, and satisfies the following: (a) $\beta_{\boldsymbol{V}} \le \alpha_{\boldsymbol{V}}$, (b) $2^{\beta_V} \ge \frac{1}{2} 2^{\alpha_V}$ when $\alpha_{\boldsymbol{V}} \neq 0$ and $\beta_{\boldsymbol{V}} = \alpha_{\boldsymbol{V}}$ when $\alpha_{\boldsymbol{V}|} = 0$. Define the normal relation $T \stackrel{\text{def}}{=} \bigotimes_{\boldsymbol{V}\subseteq \boldsymbol{X}} T^{\boldsymbol{V}}_{2^{\beta_V}}$; thus, T is uniform. We check that T satisfies the lemma. Its entropy is

$$\boldsymbol{h}_T = \sum_{\boldsymbol{V}\subseteq\boldsymbol{X}} \beta_{\boldsymbol{V}} \boldsymbol{h}^{\boldsymbol{V}}$$

Condition (1) follows form property (a). For all $\boldsymbol{U}, \boldsymbol{W} \subseteq \boldsymbol{X}$:

$$h_T(\boldsymbol{W}|\boldsymbol{U}) = \sum_{\boldsymbol{V}\subseteq\boldsymbol{X}} \beta_{\boldsymbol{V}} h^{\boldsymbol{V}}(\boldsymbol{W}|\boldsymbol{U}) \le \sum_{\boldsymbol{V}\subseteq\boldsymbol{X}} \alpha_{\boldsymbol{V}} h^{\boldsymbol{V}}(\boldsymbol{W}|\boldsymbol{U}) = h(\boldsymbol{W}|\boldsymbol{U})$$

Condition (2) follows from property (b):

$$2^{h_T(\boldsymbol{X})} = |T| = \prod_{\boldsymbol{V}\subseteq\boldsymbol{X}} |T^{\boldsymbol{V}}_{2^{\beta_V}}| = \prod_{\boldsymbol{V}\subseteq\boldsymbol{X}} 2^{\beta_V} \ge \frac{1}{2^c} \prod_{\boldsymbol{V}\subseteq\boldsymbol{X}} 2^{\alpha_V} = \frac{1}{2^c} 2^{h(\boldsymbol{X})}$$

□

Example 6.8. *Recall that tightness of the AGM bound (ℓ_1-bound) is achieved by a* product *database, where each relation is the cartesian product of its attributes. We show a query where no product database matches the ℓ_p-upper bound, instead a normal database is needed:*

$$Q(X,Y,Z) = R_1(X,Y) \wedge R_2(Y,Z) \wedge R_3(Z,X) \wedge S_1(X) \wedge S_2(Y) \wedge S_3(Z)$$

Assume the statistics assert that each of $||\mathsf{deg}_{R_1}(Y|X)||_4^4$, $||\mathsf{deg}_{R_2}(Z|Y)||_4^4$, $||\mathsf{deg}_{R_3}(X|Z)||_C^4$, $|S_1|$, $|S_2|$, $|S_3|$ *is* $\le B \stackrel{def}{=} 2^b$*. The log-statistics are:*

$$\begin{aligned} &h(X) \le b \quad h(Y) \le b \quad h(Z) \le b \\ &h(X) + 4h(Y|X) \le b \;\; h(Y) + 4h(Z|Y) \le b \;\; h(Z) + 4h(X|Z) \le b \end{aligned} \tag{42}$$

Consider the following Shannon inequality:

$$h(X) + h(Y) + h(Z) + (h(X) + 4h(Y|X)) + \\ (h(Y) + 4h(Z|Y)) + (h(Z) + 4h(X|Z)) \geq 6h(XYZ) \tag{43}$$

In particular, the above inequality is a sum of the following basic Shannon inequalities:

$$2h(X) + 2h(Y|X) + 2h(Z|Y) \geq 2h(XYZ)$$
$$2h(Y) + 2h(Z|Y) + 2h(X|Z) \geq 2h(XYZ)$$
$$2h(Z) + 2h(X|Z) + 2h(Y|X) \geq 2h(XYZ)$$

Inequality (43) *implies that* $|Q(\boldsymbol{D})| \leq B$. *To compute the worst-case instance* $\boldsymbol{D}$, *observe that* $\boldsymbol{h}^* = b \cdot \boldsymbol{h}^{\{X,Y,Z\}}$ *is the optimal solution to* (42), *since it satisfies* (42) *and* $h^*(XYZ) = b$, *and define:*

$$T \stackrel{def}{=} \{(k, k, k) \mid k = 0, \lfloor 2^b \rfloor - 1\}$$

Then $\boldsymbol{D}$ *consists of projections of* T, *e.g.* $R_1^D = \Pi_{XY}(T)$, $S_1^D = \Pi_X(T)$, *etc, and* $|Q(\boldsymbol{D})| = |T| = \lfloor 2^b \rfloor \geq \frac{1}{2} 2^b = \frac{1}{2} B$. *On the other hand, for any product database* $\boldsymbol{D}$, *the output* $Q(\boldsymbol{D})$ *is asymptotically smaller than* B. *Such a database has* $R_1^D = [N_X] \times [N_Y]$ *and* $||\text{deg}_{R_1}(Y|X)||_4^4 = N_X N_Y^4$. *The concrete* ℓ_4*-statistics become:*

$$N_X N_Y^4 \leq B \qquad N_Y N_Z^4 \leq B \qquad N_Z N_X^4 \leq B$$

By multiplying them we derive $N_X N_Y N_Z \leq B^{3/5}$. *Since* $Q(\boldsymbol{D}) = [N_X] \times [N_Y] \times [N_Z]$ *we derive* $|Q(\boldsymbol{D})| \leq B^{3/5}$, *which is asymptotically smaller than the upper bound* B.

7 CONCLUSIONS

We have described a new upper bound on the size of the output of a multi-join query, using ℓ_p-norms of degree sequences. Our techniques are based on information inequalities, and extend prior results in [3–5, 13]. This is complemented by a query evaluation algorithm whose runtime matches the size bound. The bound can be computed by optimizing a linear program whose size is exponential in the size of the query. This bound is tight in the case when all degree sequences are simple.

Our new bounds significantly extend the previously known upper bounds, especially for acyclic queries. We have also conducted some very preliminary experiments on real datasets in the full version [2], which showed significantly better upper bounds for acyclic queries than the AGM and PANDA bounds from prior work.

In future work, we will incorporate our ℓ_p-bounds into a cardinality estimation system.

ACKNOWLEDGMENTS

The authors would like to acknowledge Luis Torrejón Machado for their help with the preliminary experiments reported in the full version of this paper [2].

This work was partially supported by NSF-BSF 2109922, NSF-IIS 2314527 and NSF-SHF 2312195. Part of this work was conducted while some of the authors participated in the Simons Program on Logic and Algorithms in Databases and AI.

REFERENCES

[1] Mahmoud Abo Khamis, Phokion G. Kolaitis, Hung Q. Ngo, and Dan Suciu. 2021. Bag Query Containment and Information Theory. *ACM Trans. Database Syst.* 46, 3 (2021), 12:1–12:39. https://doi.org/10.1145/3472391

[2] Mahmoud Abo Khamis, Vasileios Nakos, Dan Olteanu, and Dan Suciu. 2023. Join Size Bounds using Lp-Norms on Degree Sequences. *arXiv e-prints* (June 2023). https://doi.org/10.48550/arXiv.2306.14075 arXiv:2306.14075

[3] Mahmoud Abo Khamis, Hung Q. Ngo, and Dan Suciu. 2016. Computing Join Queries with Functional Dependencies. In *Proc. ACM PODS 2016*. ACM, 327–342. https://doi.org/10.1145/2902251.2902289

[4] Mahmoud Abo Khamis, Hung Q. Ngo, and Dan Suciu. 2017. What Do Shannon-type Inequalities, Submodular Width, and Disjunctive Datalog Have to Do with One Another?. In *Proc. ACM PODS 2017*. ACM, 429–444. https://doi.org/10.1145/3034786.3056105
[5] Albert Atserias, Martin Grohe, and Dániel Marx. 2013. Size Bounds and Query Plans for Relational Joins. *SIAM J. Comput.* 42, 4 (2013), 1737–1767. https://doi.org/10.1137/110859440
[6] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization.* Cambridge University Press. https://doi.org/10.1017/CBO9780511804441
[7] Walter Cai, Magdalena Balazinska, and Dan Suciu. 2019. Pessimistic Cardinality Estimation: Tighter Upper Bounds for Intermediate Join Cardinalities. In *SIGMOD, 2019*. ACM, 18–35. https://doi.org/10.1145/3299869.3319894
[8] Jeremy Chen, Yuqing Huang, Mushi Wang, Semih Salihoglu, and Kenneth Salem. 2022. Accurate Summary-based Cardinality Estimation Through the Lens of Cardinality Estimation Graphs. *Proc. VLDB Endow.* 15, 8 (2022), 1533–1545. https://www.vldb.org/pvldb/vol15/p1533-chen.pdf
[9] Kyle Deeds, Dan Suciu, Magda Balazinska, and Walter Cai. 2023. Degree Sequence Bound for Join Cardinality Estimation. In *ICDT 2023 (LIPIcs, Vol. 255)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 8:1–8:18. https://doi.org/10.4230/LIPIcs.ICDT.2023.8
[10] Kyle B. Deeds, Dan Suciu, and Magdalena Balazinska. 2023. SafeBound: A Practical System for Generating Cardinality Bounds. *Proc. ACM Manag. Data* 1, 1 (2023), 53:1–53:26. https://doi.org/10.1145/3588907
[11] Ronald Fagin. 1982. Horn clauses and database dependencies. *J. ACM* 29, 4 (1982), 952–985.
[12] Dan Geiger and Judea Pearl. 1993. Logical and Algorithmic Properties of Conditional Independence and Graphical Models. *The Annals of Statistics* 21, 4 (1993), 2001–2021. http://www.jstor.org/stable/2242326
[13] Georg Gottlob, Stephanie Tien Lee, Gregory Valiant, and Paul Valiant. 2012. Size and Treewidth Bounds for Conjunctive Queries. *J. ACM* 59, 3 (2012), 16:1–16:35. https://doi.org/10.1145/2220357.2220363
[14] Yuxing Han, Ziniu Wu, Peizhi Wu, Rong Zhu, Jingyi Yang, Liang Wei Tan, Kai Zeng, Gao Cong, Yanzhao Qin, Andreas Pfadler, Zhengping Qian, Jingren Zhou, Jiangneng Li, and Bin Cui. 2021. Cardinality Estimation in DBMS: A Comprehensive Benchmark Evaluation. *Proc. VLDB Endow.* 15, 4 (2021), 752–765. https://doi.org/10.14778/3503585.3503586
[15] Axel Hertzschuch, Claudio Hartmann, Dirk Habich, and Wolfgang Lehner. 2021. Simplicity Done Right for Join Ordering. In *CIDR 2021*. www.cidrdb.org. http://cidrdb.org/cidr2021/papers/cidr2021_paper01.pdf
[16] Sai Vikneshwar Mani Jayaraman, Corey Ropell, and Atri Rudra. 2021. Worst-case Optimal Binary Join Algorithms under General ℓ_p Constraints. *CoRR* abs/2112.01003 (2021). https://arxiv.org/abs/2112.01003
[17] Kyoungmin Kim, Jisung Jung, In Seo, Wook-Shin Han, Kangwoo Choi, and Jaehyok Chong. 2022. Learned Cardinality Estimation: An In-depth Study. In *SIGMOD, 2022*. ACM, 1214–1227. https://doi.org/10.1145/3514221.3526154
[18] Swastik Kopparty and Benjamin Rossman. 2011. The homomorphism domination exponent. *Eur. J. Comb.* 32, 7 (2011), 1097–1114. https://doi.org/10.1016/j.ejc.2011.03.009
[19] Tony T. Lee. 1987. An Information-Theoretic Analysis of Relational Databases - Part I: Data Dependencies and Information Metric. *IEEE Trans. Software Eng.* 13, 10 (1987), 1049–1061. https://doi.org/10.1109/TSE.1987.232847
[20] Viktor Leis, Andrey Gubichev, Atanas Mirchev, Peter A. Boncz, Alfons Kemper, and Thomas Neumann. 2015. How Good Are Query Optimizers, Really? *Proc. VLDB Endow.* 9, 3 (2015), 204–215. https://doi.org/10.14778/2850583.2850594
[21] Viktor Leis, Bernhard Radke, Andrey Gubichev, Atanas Mirchev, Peter A. Boncz, Alfons Kemper, and Thomas Neumann. 2018. Query optimization through the looking glass, and what we found running the Join Order Benchmark. *VLDB J.* 27, 5 (2018), 643–668. https://doi.org/10.1007/s00778-017-0480-7
[22] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data.
[23] Hung Q. Ngo, Ely Porat, Christopher Ré, and Atri Rudra. 2018. Worst-case Optimal Join Algorithms. *J. ACM* 65, 3 (2018), 16:1–16:40. https://doi.org/10.1145/3180143
[24] Hung Q. Ngo, Christopher Ré, and Atri Rudra. 2013. Skew strikes back: new developments in the theory of join algorithms. *SIGMOD Rec.* 42, 4 (2013), 5–16. https://doi.org/10.1145/2590989.2590991
[25] Raghu Ramakrishnan and Johannes Gehrke. 2003. *Database management systems (3. ed.)*. McGraw-Hill.
[26] Doron Zeilberger. 1984. A combinatorial proof of Newton's identities. *Discrete mathematics* 49, 3 (1984).

A EQUIVALENCE OF ℓ_p-NORMS AND DEGREE SEQUENCES

The following is a standard result, establishing a 1-to-1 correspondence between a sequence of length m and its first m norms. We include it here for completeness.

Lemma A.1. *Denote by $S \subseteq \mathbb{R}_+^m$ the set of sorted sequences $d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$. The mapping $\varphi : S \rightarrow \mathbb{R}_+^m$ defined by $\varphi(\boldsymbol{d}) \stackrel{def}{=} (||\boldsymbol{d}||_1, ||\boldsymbol{d}||_2^2, \ldots, ||\boldsymbol{d}||_m^m)$ is injective.*

In other words, having the full degree sequence $d_1 \geq d_2 \geq \cdots \geq d_m$ is equivalent to having the ℓ_p-norms for $p = 1, 2, \ldots, m$.

Proof. We make use of the elementary symmetric polynomials

$$\begin{aligned}
e_0(\boldsymbol{d}) &= 1 \\
e_1(\boldsymbol{d}) &= d_1 + d_2 + \ldots + d_m \\
e_2(\boldsymbol{d}) &= \sum_{1 \leq i < j \leq m} d_i d_j \\
&\ldots \\
e_m(\boldsymbol{d}) &= d_1 \cdot d_2 \cdot \ldots \cdot d_m \\
e_k(\boldsymbol{d}) &= 0, k > m.
\end{aligned}$$

Using Newton's identities (see [26] for a simple proof) we can express the elementary symmetric polynomials using the L_p-norms as follows

$$k \cdot e_k(\boldsymbol{d}) = \sum_{p=1}^{k} (-1)^{p-1} e_{k-p}(\boldsymbol{d}) \cdot \|\boldsymbol{d}\|_p^p.$$

Thus, given the values of $\|\boldsymbol{d}\|_p$ for $p \in [m]$, the first m values of the elementary symmetric polynomials inductively, by:

$$\begin{aligned}
\boldsymbol{e}_0(\boldsymbol{d}) =& 1 \\
1 \cdot e_1(\boldsymbol{d}) =& e_0(\boldsymbol{d})||\boldsymbol{d}||_1^1 \\
2 \cdot e_2(\boldsymbol{d}) =& e_1(\boldsymbol{d})||\boldsymbol{d}||_1^1 - e_0(\boldsymbol{d})||\boldsymbol{d}||_2^2 \\
&\ldots \\
m \cdot e_m(\boldsymbol{d}) =& e_{m-1}(\boldsymbol{d})||d||_1^1 - e_{m-2}(\boldsymbol{d})||d||_2^2 + \cdots + (-1)^{m-1} e_0(\boldsymbol{d})||d||_m^m
\end{aligned}$$

This uniquely determines the values $e_1(\boldsymbol{d}), e_2(\boldsymbol{d}), \ldots, e_m(\boldsymbol{d})$. Using Vieta's formulas we have that the polynomial with roots $d_1, d_2, \ldots, d_m$ corresponds to the polynomial

$$\lambda^m - e_1(\boldsymbol{d})\lambda^{m-1} + e_2(\boldsymbol{d})\lambda^{m-2} + \ldots + (-1)^m e_m(\boldsymbol{d}).$$

Thus, the first m symmetric polynomials uniquely determine the degree vector $\boldsymbol{d}$. □

B EXAMPLES

B.1 A Single Join (Example 2.1)

We discuss here in depth our new bounds applied to the single join query in Example 2.1. For convenience, we repeat here the query (14):

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

Inequality (18). We start by describing a simple example where the bound (18) is asymptotically better the PANDA bound (17). For this purpose we define a type of database instance that we will also use in the rest of the section.

Definition B.1. *An (α, β)-sequence is a degree sequence of the form:*

$$(\underbrace{M^\beta, \ldots, M^\beta}_{M^\alpha \text{ values}}, \quad \underbrace{1, \ldots, 1}_{M-M^\alpha \text{ values}} \quad) \tag{44}$$

where $\alpha, \beta > 0$ and $\alpha + \beta \leq 1$. An (α, β) relation is a binary relation $R(X, Y)$ where both $\mathsf{deg}_R(Y|X)$ and $\mathsf{deg}_R(X|Y)$ are a (α, β)-sequence. [3]

In other words, there are M^α nodes with degree M^β, and $M - M^\alpha$ nodes with degree 1.

Let both R and S be (α, β)-instances with $\alpha = \beta = 1/3$. Then the PANDA bound (17) is $M^{4/3}$, while our bound (18) is $O(M)$, which is asymptotically better.

The inequality (18) is a special case of a more general inequality, which is of independent interest and we show it here. This new inequality uses the number of distinct values in the columns $R.Y$ and $S.Y$. Such statistics are often available in database systems, and they are captured by our framework because any cardinality statistics is a special case of an ℓ_1-statistics, e.g. $|\Pi_Y(R)|$ is the same as $||\mathsf{deg}_R(X|\emptyset)||_1$. PANDA also uses such cardinalities: for example, denoting $M \stackrel{\text{def}}{=} \min(|\Pi_Y(R)|, |\Pi_Y(S)|)$, PANDA also considers the following inequality:

$$|Q| \leq ||\mathsf{deg}_R(X|Y)||_\infty \cdot ||\mathsf{deg}_S(Z|Y)||_\infty \cdot M, \tag{45}$$

Yet the best PANDA bound remains (17), because it is always better than (45).

Our new inequality uses M in the following bound, which holds for all $p, q > 0$ satisfying $\frac{1}{p} + \frac{1}{q} \leq 1$:

$$|Q| \leq ||\mathsf{deg}_R(X|Y)||_p \cdot ||\mathsf{deg}_S(Z|Y)||_q \cdot M^{1-\frac{1}{p}-\frac{1}{q}} \tag{46}$$

Inequality (18) is the special case of (46) for $p = q = 2$, while the PANDA bound (17) is the special case $p = 1, q = \infty$ and $p = \infty, q = 1$.

We prove (46), by using the following Shannon inequality (which is of the form (8)):

$$\left(\frac{1}{p}h(Y) + h(X|Y)\right) + \left(\frac{1}{q}h(Y) + h(Z|Y)\right) + \left(1 - \frac{1}{p} - \frac{1}{q}\right)h(Y) \geq h(XYZ)$$

The inequality simplifies to $h(Y) + h(X|Y) + h(Z|Y) \geq h(XYZ)$, which holds because: $h(Z|Y) \geq h(Z|XY)$; $h(Y) + h(X|Y) = h(XY)$; and $h(XY) + h(Z|XY) = h(XYZ)$.

Examining closer (46), we also prove that it can only be optimal when $\frac{1}{p} + \frac{1}{q} = 1$, because, whenever $p \leq p_1$ and $q \leq q_1$, then the bound (46) using (p, q) is better than that using (p_1, q_1). In particular, the only integral values of p, q for which (46) could be optimal are $(\infty, 1)$, $(1, \infty)$, and $(2, 2)$: other potentially optimal pairs (p, q) exists, e.g. $(6/5, 6)$, but they require fractional p or q. To prove this claim, it suffices to prove the following: if $p \leq p_1$ then

$$\frac{||\mathsf{deg}_R(X|Y)||_p}{M^{\frac{1}{p}}} \leq \frac{||\mathsf{deg}_R(X|Y)||_{p_1}}{M^{\frac{1}{p_1}}}$$

We rearrange the inequality as:

$$||\mathsf{deg}_R(X|Y)||_p \leq ||\mathsf{deg}_R(X|Y)||_{p_1} \cdot M^{\frac{1}{p}-\frac{1}{p_1}}$$

[3]Such a relation exists, either by Gale–Ryser theorem, or by direct construction: take R the disjoint union of $\{(i, (i, j)) \mid i \in [M^\alpha], j \in [M^\beta]\}$, $\{((i, j), i) \mid i \in [M^\alpha], j \in [M^\beta]\}$, and $\{(i, i) \mid i \in [M - 2M^{\alpha+\beta}]\}$.

Denoting $\text{deg}_R(X|Y) = (d_1, d_2, \ldots)$ the inequality becomes:

$$\left(\sum_i d_i^p\right)^{\frac{1}{p}} \leq \left(\sum_i d_i^{p_1}\right)^{\frac{1}{p_1}} M^{\frac{1}{p}-\frac{1}{p_1}}$$

We raise both sides to the power p, and denote by $a_i \stackrel{\text{def}}{=} d_i^p$ and $q \stackrel{\text{def}}{=} \frac{p_1}{p}$. Then the inequality becomes:

$$\sum_i a_i \leq \left(\sum_i a_i^q\right)^{\frac{1}{q}} M^{1-\frac{1}{q}}$$

which is Hölder's inequality. This proves the claim.

Inequality (19). Next, we provide the proof of (19), by establishing the following Shannon inequality:

$$\left(\frac{1}{p}h(Y) + h(X|Y)\right) + \left(1 - \frac{q}{p(q-1)}\right)h(YZ) + \frac{q}{p(q-1)}\left(\frac{1}{q}h(Y) + h(Z|Y)\right) \geq h(XYZ)$$

The coefficient $1 - \frac{q}{p(q-1)}$ is ≥ 0 because $\frac{1}{p} + \frac{1}{q} \leq 1$. We expand the LHS of inequality and obtain:

$$\begin{aligned}
&\frac{1}{p}h(Y) + h(X|Y) + h(YZ) - \frac{q}{p(q-1)}h(YZ) + \frac{1}{p(q-1)}h(Y) + \frac{q}{p(q-1)}h(Z|Y) \\
=&\frac{q}{p(q-1)}h(Y) + h(X|Y) + h(YZ) - \frac{q}{p(q-1)}h(YZ) + \frac{q}{p(q-1)}h(Z|Y) \\
=&h(X|Y) + h(YZ) \geq h(XYZ)
\end{aligned}$$

which proves the claim. We will show below that (19) can be strictly better than (46).

Comparison to the DSB. A method for computing an upper bound on the query's output using degree sequences was described in [9], which uses the full degree sequence $d_1 \geq d_2 \geq \cdots$ instead of its $\ell_1, \ell_2, \ldots$ norms. We compare it here to our method, on our single join query. It turns out that (19) play a key role in this comparison.

Suppose R, S have the following degree sequences:

$$\begin{aligned}
\text{deg}_R(X|Y) =& a_1 \geq a_2 \geq \cdots \geq a_M \\
\text{deg}_S(Z|Y) =& b_1 \geq b_2 \geq \cdots \geq b_M
\end{aligned}$$

If the system has full access to both degree sequences, then the Degree-Sequence Bound (DSB) defined in [9] is the following quantity:

$$DSB \stackrel{\text{def}}{=} \sum_{i=1,M} a_i b_i \tag{47}$$

In general the degree sequences are too large to store, and the DSB bound needs to use compression [10], but for the purpose of our discussion we will assume that we know both degree sequences, and DSB is given by the formula above. It is easy to check that $|Q| \leq DSB$. Our bound (18) becomes:

$$|Q| \leq ||\text{deg}_R(X|Y)||_2 \cdot ||\text{deg}_S(Z|Y)||_2 = \sqrt{(\sum_i a_i^2)(\sum_i b_i^2)}$$

Thus, the DSB and the ℓ_2-bound above are the two sides of the Cauchy-Schwartz inequality; DSB is obviously the better one. DSB is also better than the PANDA bound (17), which in our notation is $\min(a_1 \sum_i b_i, b_1 \sum_i a_i)$ (assuming a_1 and b_1 are the largest degrees). Can we compute a better

ℓ_p-bound? We will show that (19) can improve over both (17), and (18), however, it remains strictly weaker than the *DSB* bound. This may be surprising, given the 1-1 correspondence between the statistics and the ℓ_p-bounds that we described in Appendix A. The mapping between a degree sequence of length M and its $\ell_1, \ell_2, \ldots, \ell_M$-norms is 1-to-1, and, moreover, both bounds are tight: tightness of the DSB bound was proven in [9], while tightness of the polymatroid bound holds because both degrees are simple, and it follows from our discussion in Sec. 6. So, one expects that some ℓ_p-bounds will match the *DSB* expression (47). However, this is not the case, for a rather subtle reason: it is because the set of databases to which these two bounds apply, differ. The 1-to-1 mapping from degrees to ℓ_p-norms is monotone in one direction, but not in the other. For example, consider the degree sequence $\boldsymbol{d} = (d_1, d_2) = (a + \varepsilon, a - \varepsilon)$, where $||\boldsymbol{d}||_1 = 2a$, $||\boldsymbol{d}||_2^2 = 2a^2 + 2\varepsilon^2$. A database with degree sequence $\boldsymbol{d'} = (d'_1, d'_2) = (a, a)$ satisfies the ℓ_p-constraints, because $||\boldsymbol{d'}||_1 = 2a$, $||\boldsymbol{d'}||_2^2 = 2a^2$, but it does not satisfy the degree sequence, because $d'_2 > d_2$. We show next that the polymatroid bound that we can obtain from the ℓ_p-norms can be strictly worse than the DSB. However, we note that, for practical applications, the degree sequences in the *DSB* bound need to be compressed, leading to a different loss of precision, which makes it incomparable to the ℓ_p-bound.

We describe now an instance where there exists a gap between the DSP bound and the polymatroid bound: the relation R is a $(0, 1/3)$-relation, while S is a $(0, 2/3)$-relation, see Def. B.1. More precisely, the two relations $R(X, Y), S(Y, Z)$ will have the following degree sequences:

$$\mathsf{deg}_R(X|Y) = \left(M^{\frac{1}{3}}, 1, 1, \ldots, 1\right) \qquad M \text{ values}$$
$$\mathsf{deg}_S(Z|Y) = \left(M^{\frac{2}{3}}, 1, 1, \ldots, 1\right) \qquad M \text{ values}$$

There are M degrees equal to 1 in both sequences. The value $DSB = O(M)$ is asymptotically tight, because $|Q| = O(M)$. Assume that we access to all statistics $||\mathsf{deg}_R(X|Y)||_p$, $||\mathsf{deg}_S(Z|Y)||_p$, for $p = 1, 2, \ldots, M, \infty$. We prove:

Claim 1. *The polymatroid bound is* $M^{\frac{10}{9}}$.

Normally, the polymatroid bound is computed as the optimal solution of a linear program, as described in Sec. 5. However, to prove the claim, we proceed differently. First, we describe an inequality proving that the polymatroid bound is $O(M^{\frac{10}{9}})$. Second, we describe a database instance that satisfies all the given ℓ_p-statistics, for which the query output has size $|Q| = \Omega(M^{\frac{10}{9}})$. These two steps prove the claim. We start by computing the ℓ_p-norms for our instance:

$$||\mathsf{deg}_R(X|Y)||_p^p = \begin{cases} O(M) & \text{when } p \le 2 \\ O\left(M^{\frac{p}{3}}\right) & \text{when } p \ge 3 \end{cases}$$
$$||\mathsf{deg}_S(Z|Y)||_q^q = \begin{cases} O(M) & \text{when } q = 1 \\ O\left(M^{\frac{2q}{3}}\right) & \text{when } q \ge 2 \end{cases}$$
$$|Q| = DSB = M^{\frac{1}{3}} \cdot M^{\frac{2}{3}} + M = O(M)$$

For the first step, we use the inequality (19) specialized for $p = 3, q = 2$, which we show here:[4]

$$|Q| \le ||\mathsf{deg}_R(X|Y)||_3 \cdot |S|^{\frac{1}{3}} \cdot ||\mathsf{deg}_S(Z|Y)||_2^{\frac{2}{3}} \tag{48}$$

[4]A direct proof follows from the following Shannon inequality:

$$\frac{1}{3}\left(h(Y) + 3h(X|Y)\right) + \frac{1}{3}h(YZ) + \frac{1}{3}\left(h(Y) + 2h(Z|Y)\right) \ge h(XYZ)$$

Since $|S| = ||\text{deg}_S(Z|X)||_1 = O(M)$, we obtain

$$|Q| \leq O\left(M^{\frac{1}{3}} \cdot M^{\frac{1}{3}} \cdot M^{\frac{2}{3}\cdot\frac{2}{3}}\right) = O\left(M^{\frac{10}{9}}\right)$$

The other upper bound (46) leads to strictly larger upper bounds, for any choice of p, q.

For the second step we construct a new database instance R', S' that satisfies all the ℓ_p-statistics that we computed for R, S. We describe them using their degrees:

$$\text{deg}_{R'}(X|Y) = O\left(M^{\frac{1}{9}}, \dots, M^{\frac{1}{9}}\right) \qquad M^{\frac{2}{3}} \text{ values}$$
$$\text{deg}_{S'}(Z|Y) = O\left(M^{\frac{1}{3}}, \dots, M^{\frac{1}{3}}\right) \qquad M^{\frac{2}{3}} \text{ values}$$

Then the following hold:

$$||\text{deg}_{R'}(X|Y)||_p^p = O\left(M^{\frac{p}{9}+\frac{2}{3}}\right)$$
$$||\text{deg}_{S'}(Z|Y)||_q^q = O\left(M^{\frac{q}{3}+\frac{2}{3}}\right)$$
$$|Q'| = M^{\frac{1}{9}} \cdot M^{\frac{1}{3}} \cdot M^{\frac{2}{3}} = M^{\frac{10}{9}}$$

We check that the ℓ_p-norms of the degrees of R', S' are no larger than those of R, S:

$$p \leq 2: \qquad ||\text{deg}_{R'}(X|Y)||_p^p = O\left(M^{\frac{p}{9}+\frac{2}{3}}\right) \leq O(M) = ||\text{deg}_R(X|Y)||_p^p$$
$$p \geq 3: \qquad ||\text{deg}_{R'}(X|Y)||_p^p = O\left(M^{\frac{p}{9}+\frac{2}{3}}\right) \leq O\left(M^{\frac{p}{3}}\right) = ||\text{deg}_R(X|Y)||_p^p$$

$$q = 1: \qquad ||\text{deg}_{S'}(Z|Y)||_1 = O\left(M^{\frac{1}{3}+\frac{2}{3}}\right) \leq O(M) = ||\text{deg}_S(Z|Y)||_1$$
$$q \geq 2: \qquad ||\text{deg}_{S'}(Z|Y)||_q^q = O\left(M^{\frac{q}{3}+\frac{2}{3}}\right) \leq O\left(M^{\frac{2q}{3}}\right) = ||\text{deg}_S(Z|Y)||_q^q$$

Similarly, $|R'.Y| = |S'.Y| = M^{\frac{2}{3}} \leq M$. It follows that the relations R', S' satisfy all constraints on the ℓ_p-norms, including those on $|R'.Y|, |S'.Y|$ (assuming the latter are available). Yet the size of the output of the query on R', S' is $M^{\frac{10}{9}}$.

As explained earlier, the issue stems from the fact that the DSB bound does not permit the instance R', S', since its degree sequences are not dominated by those of R, S.

B.2 The Cycle Query (Example 2.3)

We show that, for every $p \geq 1$ there exists a database instance where the bound (21) for $q := p$ is the theoretically optimal bound that can be derived using all statistics on $\ell_1, \ell_2, \dots, \ell_p, \ell_\infty$ norms.

First, we describe a database instance for which bound (21) for $q := p$ is better than the bounds in (23). The instance consists of the (α, β)-relation R for $\alpha = \beta = \frac{1}{p+1}$ (see Def. B.1); to simplify the notations here we will rename M to N. Thus, we have $|R| = N$, $||\text{deg}_R(Y|X)||_q^q = N$ for $q \in [p]$, $||\text{deg}_R(Y|X)||_\infty = N^{\frac{1}{p+1}}$, and the bounds in (23) and (21) become $N^{\frac{p+1}{2}}$, $N^{\frac{2p}{p+1}}$, and $N^{\frac{p+1}{q+1}}$ respectively. The best bound among them is the latter, when $q = p$, which gives us $|Q| \leq ||\text{deg}_R(Y|X)||_p^p = (1 + o(N))N$. All other bounds are asymptotically worse. Thus, among these three formulas, (21) is the best, namely for $q := p$. However, this does not yet prove that these formulas provide the best bounds if we have access to the given statistics.

We show now that these bounds are tight. In other words, we show that there exists relation instances R for which the bounds are tight, up to constant factors. We already know this for the

$\{1\}$-bound (the AGM bound), since the AGM bound is N^{ρ^*}, where ρ^* is the optimal fractional edge covering number of the $(p+1)$-cycle, which is $\rho^* = \frac{p+1}{2}$.

Consider now the $\{1,\infty\}$-bound, in other words we have only the statistics for $||\text{deg}_R(Y|X)||_1$ (which is $|R|$) and $||\text{deg}_R(Y|X)||_\infty$. We prove that the PANDA bound in (23) is indeed optimal. In fact we prove a more general claim: the $\{1,\infty\}$-bound of the cycle query is $|Q| \leq N \cdot D^{p-1}$, whenever $|R| \leq N$, $||\text{deg}_R Y|X||_\infty \leq D$, and N, D are numbers satisfying $D^2 \leq N$. In our case we have $D = N^{\frac{1}{p+1}}$, and the claim implies that the $\{1,\infty\}$-bound is $N^{1+\frac{p-1}{p+1}} = N^{\frac{2p}{p+1}}$. To prove this claim, we will refer to the polymatroid upper bound, and polymatroid lower bound in Def. 5.1. The Shannon inequality that we proved in Example 2.3 implies $\text{Log-U-Bound}_{\Gamma_n}(Q) \leq \log N + (p-1)\log D$. We also have $\text{Log-U-Bound}_{\Gamma_n}(Q) = \text{Log-U-Bound}_{N_n}(Q)$ (by Theorem 6.1), where N_n are the normal polymatroids, and $\text{Log-U-Bound}_{N_n}(Q) = \text{Log-L-Bound}_{N_n}(Q)$ by Theorem 5.2. We claim that there exists a normal polymatroid that satisfies the $\{1,\infty\}$-statistics and where $h(X_0 \ldots X_p) = \log N + (p-1)\log D$: the claim implies $\log N + (p-1)\log D \leq \text{Log-L-Bound}_{\Gamma_n}(Q)$, which proves that the $\{1,\infty\}$-bound is $N \cdot D^{p-1}$. To prove the claim, consider the following polymatroid:

$$h(\emptyset) = 0, \qquad \forall \boldsymbol{W} \neq \emptyset,\ h(\boldsymbol{W}) \stackrel{\text{def}}{=} \log N + (|\boldsymbol{W}| - 2)\log D$$

Then h satisfies the required statistics:

$$\forall i: \quad h(X_i X_{i+1}) \leq \log N \qquad\qquad h(X_{i+1}|X_i) \leq \log D$$

and $h(X_0 X_1 \ldots X_p) = \log N + (p-1)\log D$. It remains to observe that h is a normal polymatroid, which follows by writing it as $\boldsymbol{h} = (\log N - 2\log D) \cdot h^{\boldsymbol{X}} + \log D \cdot \sum_{i=0,p} h^{X_i}$.

Finally, we prove that, if we have available all statistics $||\text{deg}_R(Y|X)||_q$ for $q = 1, 2, \ldots, p, \infty$, then the best query upper bound is (21). Fix a number $q \in [p]$, and let N, L, D be three positive numbers satisfying $L \leq N$ and $L \leq D^{q+1}$. Then we claim that the $\{1, 2, \ldots, q, \infty\}$-bound of the cyclic query in Example 2.3, when the input relation satisfies the statistics $|R| \leq N$, $||\text{deg}_R(Y|X)||_r^r \leq L$, for all $r \leq q$, and $||\text{deg}_R(Y|X)||_\infty \leq D$, is $|Q| \leq L^{\frac{(p+1)q}{q+1}}$. The claim applies to our database instance (α, β) for $\alpha = \beta = \frac{1}{p+1}$, because we have $L = (1 + o(N))N$ and $D = N^{\frac{1}{p+1}}$, and implies that the $\{1, 2, \ldots, q\}$-bound is $L^{\frac{(p+1)q}{q+1}}$. To prove the claim, we use the same reasoning as above: it suffices to describe a polymatroid satisfying the statistics

$$\begin{aligned} h(X_i X_{i+1}) &\leq \log N \\ \forall r = 2, q: \quad h(X_i X_{i+1}) + (r-1)h(X_{i+1}|X_i) &\leq \log L \\ h(X_{i+1}|X_i) &\leq \log D \end{aligned}$$

The desired polymatroid is the following modular function: $h(\boldsymbol{W}) \stackrel{\text{def}}{=} \frac{|\boldsymbol{W}| \cdot \log L}{q+1}$. In other words, $\boldsymbol{h} = \frac{1}{q+1} \sum_{i=0,p} \boldsymbol{h}^{X_i}$. Then, the first inequality above is $\frac{2\log L}{q+1} \leq \log N$, and it holds because $L \leq N$. The second inequality is $(r+1)\frac{\log L}{q+1} \leq \log L$, which holds because $r \leq q$. And the third inequality is $\frac{\log L}{q+1} \leq \log D$, which holds by the assumption $L \leq D^{q+1}$.

Received December 2023; revised February 2024; accepted March 2024