Size bounds and query plans for relational joins

Albert Atserias
Universitat Politècnica de Catalunya
Barcelona, Spain

Martin Grohe
Humboldt Universität zu Berlin
Berlin, Germany

Dániel Marx
Budapest University of Technology and Economics
Budapest, Hungary

April 28, 2008

Abstract

Relational joins are at the core of relational algebra, which in turn is the core of the standard database query language SQL. As their evaluation is expensive and very often dominated by the output size, it is an important task for database query optimisers to compute estimates on the size of joins and to find good execution plans for sequences of joins. We study these problems from a theoretical perspective, both in the worst-case model, and in an average-case model where the database is chosen according to a known probability distribution. In the former case, our first key observation is that the worst-case size of a query is characterised by the fractional edge cover number of its underlying hypergraph, a combinatorial parameter previously known to provide an upper bound. We complete the picture by proving a matching lower bound, and by showing that there exist queries for which the join-project plan suggested by the fractional edge cover approach may be substantially better than any join plan that does not use intermediate projections. On the other hand, we show that in the average-case model, every join-project plan can be turned into a plan containing no projections in such a way that the expected time to evaluate the plan increases only by a constant factor independent of the size of the database. Not surprisingly, the key combinatorial parameter in this context is the maximum density of the underlying hypergraph. We show how to make effective use of this parameter to eliminate the projections.

1 Introduction

The join operation is one of the core operations of relational algebra, which in turn is the core of the standard database query language SQL. The two key components of a database system executing SQL-queries are the query optimiser and the execution engine. The optimiser translates the query into several possible execution plans, which are basically terms of the relational algebra (also called operator trees) arranging the operations that have to be carried out in a tree-like order. Using statistical information about the data, the optimiser estimates the execution cost of the different plans and passes the best one on to the execution engine, which then executes the plan and computes the result of the query. See [3] for a survey of query optimisation techniques.

Among the relational algebra operations, joins are usually the most costly, simply because a join of two relations, just like a Cartesian product of two sets, may be much larger than the relations. Therefore, query optimisers pay particular attention to the execution of joins, especially to the execution order of sequences of joins, and to estimating the size of joins. In this paper, we address the very fundamental questions of how to estimate the size of a sequence of joins and how to execute the sequence best from a theoretical point of view. While these questions have been intensely studied in practice, and numerous heuristics and efficiently solvable special cases are known (see, e.g., [3, 8, 7]), the very basic theoretical results we present here and their consequences apparently have not been noticed so far. Our key starting observation is that the size of a a sequence of joins is tightly linked to two combinatorial parameters of the underlying database schema, the *fractional edge cover number*, and the *maximum density*.

To make this precise, we need to get a bit more technical: A *join query* Q is an expression of the form

$$R_1(x_{11},\dots,x_{1r_1}) \bowtie \cdots \bowtie R_m(x_{m1},\dots,x_{mr_m}), \tag{1.1}$$

where the R_i are *relation names* with *attributes* $x_{i1},\dots,x_{ir_i}$. Let A be the set of all attributes occurring in Q and $n = |A|$. A *database instance* D for Q consists of relations $R_i(D)$ of arity r_i. It is common to think of the relation $R_i(D)$ as a table whose columns are labelled by the attributes $x_{i1},\dots,x_{ir_i}$ and whose rows are the tuples in the relation. Then *answer*, or *set of solutions*, of the query Q in D is the n-ary relation $Q(D)$ with attributes A consisting of all tuples t whose projection on the attributes of R_i belongs to the relation $R_i(D)$, for all i. Hence we are considering *natural joins* here (all of our results can easily be transferred to *equi-joins*, but not to general θ*-joins*). Now the most basic question is how large $Q(D)$ can get in terms of $|D|$. We address this question both in the worst case and the average case, and also subject to various constraints imposed on D.

An *execution plan* for a join query describes how to carry out the evaluation of the query by simple operations of the relational algebra such as joins of two relations or projections. The obvious execution plans for a join query break up the sequence of joins into pairwise joins and arrange these in a tree-like fashion. We call such execution plans *join plans*. As described in [3], most practical query engines simply arrange the joins in some linear (and not even a tree-like) order and then evaluate them in this order. However, it is also possible to use other operations, in particular projections, in an execution plan for a join query. We call execution plans that use joins and projections *join-project plans*. It is one of our main results that, even though projections are not necessary to evaluate join queries, their use may speed up the evaluation of a query super-polynomially.

Fractional covers, worst-case size, and join-project plans. Recall that an *edge cover* of a hypergraph H is a set C of edges of H such that each vertex is contained in at least one edge in C, and the *edge cover number* $\rho(H)$ of H is the minimum size of an edge cover of H. A *fractional edge cover* of H is a feasible solution for the linear programming relaxation of the integer linear program describing edge covers, and the *fractional edge cover number* $\rho^*(H)$ of H is the cost of an optimal solution. With a join query Q of the form (1.1) we can associate a hypergraph $H(Q)$ whose vertex set is the set of all attributes of Q and whose edges are the attribute sets of the relations R_i. The *(fractional) edge cover number* of Q is defined by $\rho(Q) = \rho(H(Q))$ and

$\rho^*(Q) = \rho^*(H(Q))$.

An often observed fact about edge covers is that for every given database D, the size of $Q(D)$ is bounded by $|D|^{\rho(Q)}$, where $|D|$ is the total number of tuples in D. Much less obvious is the fact that the size of $Q(D)$ can actually be bounded by $|D|^{\rho^*(Q)}$, as proved by the second and third author [9] in the context (and the language) of constraint satisfaction problems. This is a consequence to Shearer's Lemma [4], which is a combinatorial consequence of the submodularity of the entropy function, and is closely related to a result due to Friedgut and Kahn [6] on the number of copies of a hypergraph in another. Our first and most basic observation is that the fractional edge cover number $\rho^*(Q)$ also provides a lower bound to the worst-case answer size: we show that for every Q, there exist arbitrarily large databases D for which the size of $Q(D)$ is at least $|D|^{\rho^*(Q)} \cdot |Q|^{-1}$. The proof is a simple application of linear programming duality. Another result from [9] implies that for every join query there is a join-project plan, which can easily be obtained from the query and certainly be computed in polynomial time, that computes $Q(D)$ in time $O(|Q|^2 \cdot |D|^{\rho^*+1})$. Our lower bound shows that this is optimal up to a polynomial factor (of $|Q|^{2+\rho^*} \cdot |D|$, to be precise). In particular, we get the following equivalences giving an exact combinatorial characterisation of all classes of join queries that have polynomial size answers and can be evaluated in polynomial time.

Theorem 1. *Let $\mathcal{Q}$ be a class of join queries. Then the following statements are equivalent:*

(1) Queries in $\mathcal{Q}$ have answers of polynomial size.

(2) Queries in $\mathcal{Q}$ can be evaluated in polynomial time.

(3) Queries in $\mathcal{Q}$ can be evaluated in polynomial time by an explicit join-project plan.

(4) $\mathcal{Q}$ has bounded fractional edge cover number.

Note that it is not even obvious that the first two statements are equivalent, that is, that for every class of queries with polynomial size answers there is a polynomial time evaluation algorithm (the converse, of course, is trivial).

Hence with regards to worst-case complexity, join-project plans are optimal (up to a polynomial factor) for the evaluation of join queries. Our next result is that join plans are not: We prove that there are arbitrarily large join queries Q and database instances D such that our generic join-project plan computes $Q(D)$ in at most cubic time, whereas any join plan requires time $|D|^{\Omega(\log |Q|)}$ to compute $Q(D)$. We also observe that this bound is tight. Hence incorporating projections into a query plan may lead to a superpolynomial speed-up even if the projections are completely irrelevant for the query answer.

Maximum density, average-case size, and join plans. Consider the model $\mathscr{D}(N, (p_R))$ of random databases where the tuples in each relation R are chosen randomly and independently with probability $p_R = p_R(N)$ from a domain of size N. This is the analogue of the Erdős-Rényi model of random graphs adapted to our context. It is easy to see that for D from $\mathscr{D}(N, (p_R))$, the expected size of the query answer $Q(D)$ is $N^n \cdot \prod_R p_R$, where n is the number of attributes and the product ranges over all relation names R in Q. The question is whether $|Q(D)|$ will be concentrated around the expected value. This is governed by the *maximum density* $\overline{\delta}(Q, (p_R))$ of the query, a combinatorial parameter depending on the hypergraph of the query and the weights p_R. An application of the second moment method shows that if $\overline{\delta} = \log N - \omega(1)$, then $|Q(D)|$ is concentrated around its expected value, and if $\overline{\delta} = \log N + \omega(1)$, then $|Q(D)| = 0$ almost surely. We observe that the maximum density $\overline{\delta}$ can be computed in polynomial time using max-flow min-cut techniques. Interestingly, maximum density is closely related to the edge cover number of a matroid associated with the query Q (a generalisation of the bicircular matroid of a graph) — hence edge covers also show up in the average case scenario. Due to space limitation, we have to defer the details of this connection to the full version of this paper.

In view of the results about worst-case, it is a natural question whether join-project plans are more powerful than join plans in the average case setting as well. It turns out that this is not the case: We show that every join-project plan φ for Q can be turned into a join plan φ' for which the expected execution time increases only by a constant factor independent of the database. This may be viewed as our main technical result. The transformation of φ' into φ depends on a careful balance between delaying certain joins in order to reduce the number of attributes considered in each subquery occurring in the plan and keeping as many joins as possible in order to increase the density of the subquery. The choice of which subqueries to delay and which to keep is governed by a certain submodular function related to the density of the subqueries.

Size and integrity constraints. So far, we considered worst-case bounds which make no assumptions on the database, and average-case bounds which assume a known distribution on the database. However, practical query optimisers usually make use of known information about the databases when computing their size estimates. We consider the simplest such setting where the size of the relations is known, and we want to get a (worst case) estimate on the size of $Q(D)$ subject to the constraint that the relations in D have the given sizes. By suitably modifying the objective function of the linear program for edge covers, we obtain results analogous to those obtained for the unconstrained setting. A notable difference between the two results is that here the gap between upper and lower bound becomes 2^{-n}, where n is the number of attributes, instead of $|Q|^{-\rho^*}$. We give an example showing that the gap between upper and lower bound is essentially tight. However, this is not an inadequacy of our approach through fractional edge covers, but due to the inherent complexity of the problem: by a reduction from the maximum independent-set problem on graphs, we show that, unless NP $=$ ZPP, there is no polynomial time algorithm that approximates the worst case answer size $|Q(D)|$ for given Q and relation sizes N_R by a better-than-exponential factor. In a different direction, we extend our results to a setting where the database instance is required to satisfy integrity constraints. So far, we are able to deal with restricted types of functional dependencies; due to space limitations we defer these results to the full version of the paper as well.

The structure of the paper follows this introduction. Omitted proofs are provided in a technical appendix.

2 Preliminaries

For integers $m \leq n$, by $[m,n]$ we denote the set $\{m, m+1, \dots, n\}$ and by $[n]$ we denote $[1,n]$.

Our terminology is similar to that used in [1]: An *attribute* is a symbol a with an associated *domain* $\mathrm{dom}(a)$. If not specified otherwise, we assume $\mathrm{dom}(a)$ to be an arbitrary countably infinite set, say, $\mathbb{N}$. Sometimes, we will impose restrictions on the size of the domains. A *relation name* is a symbol R with an associated finite set of attributes A. For a set $A = \{a_1, \dots, a_n\}$ of attributes, we write $R(A)$ or $R(a_1, \dots, a_n)$ to denote that A is the set of attributes of R. The *arity* of $R(A)$ is $|A|$. A *schema* is a finite set of relation names.

For a set A of attributes, an *A-tuple* is a mapping t that associates an element $t(a)$ from $\mathrm{dom}(a)$ with each $a \in A$. Occasionally, we denote A-tuples in the form $t = (t_a : a \in A)$, with the obvious meaning that t is the A-tuple with $t(a) = t_a$. The set of all A-tuples is denoted by $\mathrm{tup}(A)$. An *A-relation* is a set of A-tuples. The *active domain* of an A-relation R is the set $\{t(a) : t \in R, a \in A\}$. The *projection* of an A-tuple t to a subset $B \subseteq A$ is the restriction $\pi_B(t)$ of t to B, and the *projection* of an A-relation R is the set $\pi_B(R) = \{\pi_B(t) : t \in R\}$.

A *database instance* D of schema σ, or a *σ-instance*, consists of an A-relation $R(D)$ for every relation name R in σ with set of attributes A. The *active domain* of D is the union of active domains of all its relations. The *size* of a σ-instance D is $|D| := \sum_{R \in \sigma} |R(D)|$.

A *join query* is an expression

$$Q := R_1(A_1) \bowtie \cdots \bowtie R_m(A_m),$$

where R_i is a relation name with attributes A_i. The *schema* of Q is the set $\{R_1, \dots, R_m\}$, and the *set of attributes* of Q is $\bigcup_i A_i$. We often denote the set of attributes of a join query Q by A_Q, and we write $\mathrm{tup}(Q)$ instead of $\mathrm{tup}(A_Q)$. We write $Q(A)$ to denote that A is the set of attributes of Q. The *size* of Q is $|Q| := \sum_i |A_i|$. We write $H(Q)$ for the (multi-)hypergraph that has vertex-set A and edge-(multi-)set $\{A_1, \dots, A_m\}$. If D is a

$\{R_1,\dots,R_m\}$-instance, the *answer* of Q on D is the A-relation

$$Q(D) = \big\{ t \in \mathrm{tup}(A) : \pi_{A_i}(t) \in R_i(D) \text{ for every } i \in [m] \big\}.$$

A *join plan* is a term built from relation names and binary join operators. For example, $(R_1 \bowtie R_2) \bowtie (R_3 \bowtie R_4)$ and $((R_1 \bowtie R_2) \bowtie R_3) \bowtie (R_1 \bowtie R_4)$ are two join plans corresponding to the same join query $R_1 \bowtie R_2 \bowtie R_3 \bowtie R_4$. A *join-project plan* is a term built from relation names, binary join operators, and unary project operators. For example, $(\pi_A(R_1) \bowtie R_2) \bowtie \pi_B(R_1)$ is a join-project plan. Join-project plans have a natural representation as labelled binary trees, where the leaves are labelled by relation names, the unary nodes are labelled by projections π_A, and the binary nodes by joins. Evaluating a join plan or join-project plan φ in a database instance D means substituting the relation names by the actual relations from D and carrying out the operations in the expression. We denote the resulting relation by $\varphi(D)$. A join(-project) plan φ is a plan *for* a query Q if $\varphi(D) = Q(D)$ for every database D. The *subplans* of a join(-project) plan are defined in the obvious way. For example, the subplans of $(R_1 \bowtie R_2) \bowtie \pi_A(R_3 \bowtie R_4)$ are R_1, R_2, R_3, R_4, $R_1 \bowtie R_2$, $R_3 \bowtie R_4$, $\pi_A(R_3 \bowtie R_4)$, $(R_1 \bowtie R_2) \bowtie \pi_A(R_3 \bowtie R_4)$. If φ is a join project plan, then we often use A_φ to denote the set of attributes of the query computed by φ (this only includes "free" attributes and not those projected away by some projection in φ), and we write $\mathrm{tup}(\varphi)$ instead of $\mathrm{tup}(A_\varphi)$.

3 Worst-case model

3.1. Size bounds. Let Q be a join query with schema σ. For every $R \in \sigma$, let A_R be the set of attributes of R, and $A = \bigcup_R A_R$. Then fractional edge covers are precisely the feasible solutions $(x_R : R \in \sigma)$ for the following linear program L_Q, and the fractional edge cover number $\rho^*(Q)$ is the cost of an optimal solution.

$$\begin{array}{llll} L_Q: & \text{minimise} & \sum_R x_R & \\ & \text{subject to} & \sum_{R\,:\,a \in A_R} x_R \geq 1 & \text{for all } a \in A, \\ & & x_R \geq 0 & \text{for all } R \in \sigma. \end{array} \tag{3.1}$$

By standard arguments, there always is an optimal fractional edge cover whose values are rational and of bit-length polynomial in $|Q|$.

Lemma 2 ([9]). *Let Q be a join query with schema σ and let D be a σ-instance. Then for every fractional edge cover $(x_R : R \in \sigma)$ of Q we have*

$$|Q(D)| \leq \prod_{R \in \sigma} |R(D)|^{x_R}.$$

For the reader's convenience, we give a proof of this lemma, which is actually a simplification of the proof in [9], in Appendix A. Note that the fractional edge cover in the statement of the lemma is not necessarily one of minimum cost. The next lemma shows that the upper bound of the previous lemma is tight:

Lemma 3. *Let Q be a join query with schema σ, and let $(x_R : R \in \sigma)$ be an optimal fractional edge cover of Q. Then for every $N_0 \in \mathbb{N}$ there is a σ-instance D such that $|D| \geq N_0$ and*

$$|Q(D)| \geq \prod_{R \in \sigma} |R(D)|^{x_R}.$$

Furthermore, we can choose D in such a way that $|R(D)| = |R'(D)|$ for all $R, R' \in \sigma$ with $x_R, x_{R'} > 0$.

Proof sketch. Let A_R be the set of attributes of $R \in \sigma$ and $A = \bigcup_R A_R$. Recall $(x_R : R \in \sigma)$ is an optimal solution for the linear program (3.1). By LP-duality, there is a solution $(y_a : a \in A)$ for the dual linear program

$$\begin{array}{lll} \text{maximise} & \sum_a y_a & \\ \text{subject to} & \sum_{a \in A_R} y_a \leq 1 & \text{for all } R \in \sigma, \\ & y_a \geq 0 & \text{for all } a \in A \end{array} \tag{3.2}$$

such that $\sum_a y_a = \sum_R x_R$. There even exists such a solution with rational values.

We take an optimal solution $(y_a : a \in A)$ with $y_a = p_a/q$, where $q \geq 1$ and $p_a \geq 0$ are integers. Let $N_0 \in \mathbb{N}$, and let $N = N_0^q$. We define a σ-instance D by letting $R(D) := \{t \in \mathrm{tup}(A_R) : t(a) \in [N^{p_a/q}] \text{ for all } a \in A_R\}$ for all $R \in \sigma$. It is not hard to prove that D has the desired properties. Details can be found in Appendix A. □

Combining Lemmas 2 and 3 we obtain that, for every join query Q, every instance D gives $|Q(D)| \leq |D|^{\rho^*(Q)}$, and that there exist arbitrarily large instances D such that $|Q(D)| \geq |D|^{\rho^*(Q)} \cdot |Q|^{-1}$. This yields the equivalence between statements (1) and (4) of Theorem 1

3.2. Execution plans. It was proved in [9] that there is an algorithm for evaluating a join query Q in a database D that runs in time $O\big(|Q|^2 \cdot |D|^{\rho^*(Q)+1}\big)$. An analysis of the proof shows that the algorithms can actually be casted as the evaluation of an explicit (and simple) join-project plan. For the reader's convenience, we give a proof in Appendix B. Combined with the bounds obtained in the previous section, this yields Theorem 1

We shall prove next that join plans perform significantly worse than join-project plans. Note that to evaluate a join plan one has to evaluate all its subplans. Hence for every subplan ψ of φ and every instance D, the size $|\psi(D)|$ is a lower bound for the time required to evaluate φ in D.

Theorem 4. *For every $m, N \in \mathbb{N}$ there are a join query Q and an instance D with $|Q| \geq m$ and $|D| \geq N$, and:*

(1) $\rho^*(Q) = 2$ and hence $|Q(D)| \leq |D|^2$ (actually, $|Q(D)| \leq |D|$).

(2) Every join plan φ for Q has a subplan ψ such that $|\psi(D)| \geq |D|^{\frac{1}{5}\log|Q|}$.

Proof sketch. Let $n = \binom{2m}{m}$. For every $s \subseteq [2m]$ with $|s| = m$, let a_s be an attribute with domain $\mathbb{N}$. For every $i \in [2m]$, let R_i be a relation name having as attributes all a_s such that $i \in s$. Let A_i be the set of attributes of R_i and $A = \bigcup_{i \in [2m]} A_i$. The arity of R_i is $|A_i| = \binom{2m-1}{m-1} = \frac{m}{2m} \cdot \binom{2m}{m} = \frac{n}{2}$. Let $Q := R_1 \bowtie \cdots \bowtie R_{2m}$. Then $|Q| = 2m \cdot n/2 = m \cdot n$. Furthermore, $\rho^*(Q) \leq 2$. To see this, let $x_{R_i} = 1/m$ for every $i \in [2m]$. This forms a fractional edge cover of Q, because for every $s \subseteq [2m]$ with $|s| = m$, the attribute a_s appears in the m atoms R_i with $i \in s$.

Next, we define an instance D by letting $R_i(D)$ be the set of all A_i-tuples that have an arbitrary value from $[N]$ in one coordinate and 1 in all other coordinates. It is not hard to prove that Q and D have the desired properties. Details can be found in Appendix B. □

Note that statement (2) of the theorem implies that any evaluation algorithm for the query Q based on evaluating join plans, which may even depend on the database instance, has a running time $O(|D|^{\Omega(\log|Q|)})$, in contrast with the running time $O(|D|^3)$ achieved by evaluating the join-project plan of Theorem 15. Using the well-known fact that the integrality gap of the linear program for edge covers is logarithmic in the number of vertices of the hypergraph (that is, attributes of the join query), it can be proved that for every query Q there is a join plan φ that can be evaluated in time $O(|D|^{2\rho^*(Q)\cdot\log|Q|})$, hence the lower bound is tight up to a small constant factor (see Proposition 17 in Appendix B). Furthermore, the proof of Proposition 17 shows that, for every join query Q, there is a join plan that can be evaluated in time $|D|^{\rho(Q)}$, where $\rho(Q)$ denotes the edge cover number of Q. However, note that not only $|D|^{\rho(Q)}$ is potentially superpolynomial over $|D|^{\rho^*(Q)}$, but finding this plan is in general NP-hard. Compare this with the fact that the join-project plan given by [9] can be found efficiently (see Theorem 15 in Appendix B).

4 Average-case model

In this section we assume that database is randomly generated according to the following model. Let σ be a schema and let A_R be the set of attributes of $R \in \sigma$. For every $R \in \sigma$, let $p_R : \mathbb{N} \to (0,1)$ be a function of N, and let $p(N) = (p_R(N) : R \in \sigma)$. We denote by $\mathscr{D}(N, p(N))$ the probability space on σ-instances with domain $[N]$ defined by placing each tuple $t \in [N]^{A_R}$ in $R(D)$ with probability $p_R(N)$, independently for each tuple t and each $R \in \sigma$. Typical p's of interest are $p_R(N) = 1/2$, $p_R(N) = C \cdot N^{1-|A_R|}$, or $p_R(N) = N^{1-|A_R|} \log N$. When each $p_R = 1/2$, this is the uniform distribution over σ-instances with domain $[N]$.

4.1. Size bounds and concentration. Let Q be a join query with schema σ, let n be the number of attributes of Q, and let m be the number of relation names in σ. Let X denote the size of the query answer $Q(D)$ when D is taken from $\mathscr{D}(N, p(N))$. The expectation of X is, trivially,

$$\mathrm{E}[X] = N^n \prod_{R \in \sigma} p_R(N). \tag{4.1}$$

We want to determine under what circumstances is $|Q(D)|$ concentrated around this value. For this we need to compute the variance of X, which depends on a parameter of Q defined next.

For every $R \in \sigma$, let w_R be a positive real weight, and let $w = (w_R : R \in \sigma)$. The *density* of Q with respect to w is defined as $\delta(Q,w) = \frac{1}{n}\sum_{R\in\sigma} w_R$. Note that if $w_R = 1$ for every R, then the density is m/n. For every $B \subseteq A$, let $Q[B]$ denote the subquery induced by B; that is, $Q[B]$ is the subquery formed by all the atoms $R \in \sigma$ that have all attributes in B. The *maximum density* of Q with respect to w is $\overline{\delta}(Q,w) = \max\{\delta(Q[B],w) : B \subseteq A, B \neq \emptyset\}$.

In applications to random instances, we typically fix $w_R(N)$ to $\log_2(1/p_R(N))$ and write $\delta(Q[B])$ and $\overline{\delta}(Q)$ instead of $\delta(Q[B],w)$ and $\overline{\delta}(Q,w)$. For this choice of weights a crucial distinction is made according to whether $\overline{\delta}$ is larger or smaller than $\log_2(N)$. In the first case, there exists subquery $Q[B]$ whose expected number of solutions is smaller than 1 and therefore, by Markov's inequality, Q itself has no solutions at all with probability bounded away from 0. In the second case, every subquery has at least one solution in expectation, and we can bound the variance of X as a function of $\overline{\delta}$. Since this will be of use later on, we derive it in detail. We show that, whenever $\overline{\delta} \leq \log_2(N)$, we have the following bound for the variance:

$$\mathrm{V}[X] \leq \mathrm{E}[X]^2 \cdot (2^n - 1)2^{\overline{\delta} - \log_2 N}. \tag{4.2}$$

Let A be the set of attributes of Q. For every $R \in \sigma$, let A_R be the set of attributes of R and for every $t \in [N]^{A_R}$, let $X(R,t)$ be the indicator for the event $t \in R(D)$. These are mutually independent random variables and the expectation of $X(R,t)$ is $p_R(N)$. For every $t \in [N]^A$, let $X(t)$ be the indicator for the event $t \in Q(D)$. Note that $X(t) = \prod_{R\in\sigma} X(R,t_R)$, where t_R denotes the projection of t to the attributes of R. Also $X = \sum_t X(t)$. Towards proving (4.2), let us bound

$$\mathrm{E}\left[X^2\right] = \sum_{s,t} \mathrm{E}\left[X(s)X(t)\right]. \tag{4.3}$$

For every fixed $B \subseteq A$, let F_B be the set of pairs $(s,t) \in [N]^A \times [N]^A$ such that $s(a) = t(a)$ for every $a \in B$ and $s(a) \neq t(a)$ for every $a \in A - B$. Clearly, $(F_B)_{B\subseteq A}$ is a partition of $[N]^A \times [N]^A$ and therefore

$$\sum_{s,t} \mathrm{E}\left[X(s)X(t)\right] = \sum_{B\subseteq A} \sum_{(s,t)\in F_B} \mathrm{E}\left[X(s)X(t)\right]. \tag{4.4}$$

Fix now some $B \subseteq A$ and $(s,t) \in F_B$, and let σ_B be the relations appearing in $Q[B]$. Observe that since s and t agree on B we have $t_R = s_R$ for every $R \in \sigma_B$ and therefore $X(R,s_R)X(R,t_R) = X(R,s_R)$ for every such R. Hence:

$$X(s)X(t) = \prod_{R\in\sigma} X(R,s_R) \prod_{R\in\sigma} X(R,t_R) = \prod_{R\in\sigma-\sigma_B} X(R,s_R)X(R,t_R) \prod_{R\in\sigma_B} X(R,s_R). \tag{4.5}$$

All variables in the right-hand side product are mutually independent because either they involve different relations or different tuples. Therefore,

$$\mathrm{E}\left[X(s)X(t)\right] = \prod_{R\in\sigma-\sigma_B} p_R^2 \prod_{R\in\sigma_B} p_R = \prod_{R\in\sigma} p_R^2 \prod_{R\in\sigma_B} p_R^{-1}. \tag{4.6}$$

The number of pairs (s,t) in F_B is bounded by $N^{2|A|-|B|}$. Therefore,

$$\sum_{(s,t)\in F_B} \mathrm{E}\left[X(s)X(t)\right] \leq N^{2|A|-|B|} \prod_{R\in\sigma} p_R^2 \prod_{R\in\sigma_B} p_R^{-1} = \mathrm{E}\left[X\right]^2 \cdot N^{-|B|} \prod_{R\in\sigma_B} p_R^{-1}. \tag{4.7}$$

For $B = \emptyset$, the second term in the right-hand side of (4.7) is 1 and we get $\mathrm{E}\left[X\right]^2$. For $B \neq \emptyset$, we have

$$N^{-|B|} \prod_{R\in\sigma_B} p_R^{-1} = N^{-|B|} 2^{|B|\,\delta(Q[B])} \leq \left(N^{-1} 2^{\overline{\delta}}\right)^{|B|} \leq 2^{\overline{\delta}-\log_2 N} \tag{4.8}$$

where the first inequality holds because $\delta(Q[B]) \leq \overline{\delta}$, and the second inequality holds because $|B| > 0$ and $\overline{\delta} \leq \log_2(N)$. Putting it all together we get

$$\mathrm{E}\left[X^2\right] = \sum_{B\subseteq A} \sum_{(s,t)\in F_B} \mathrm{E}\left[X(s)X(t)\right] \leq \mathrm{E}\left[X\right]^2 + (2^n - 1)\,\mathrm{E}\left[X\right]^2 2^{\overline{\delta}-\log_2 N}. \tag{4.9}$$

Since $\mathrm{V}\left[X\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}\left[X\right]^2$, this proves (4.2).

In the following, if X is a random variable defined on the probability space $\mathscr{D}(N, p(N))$, the expression "$X \sim A$ almost surely" means that for every $\varepsilon > 0$ and $\delta > 0$, there exists N_0 such that, for every $N \geq N_0$, we have $\Pr[\,|X - A| \leq \varepsilon A\,] \geq 1 - \delta$. With all this notation, we obtain the following threshold behaviour as an immediate consequence to Markov's and Chebyshev's inequalities:

Theorem 5. *Let Q be a join query with schema σ and n attributes. For every $R \in \sigma$, let $p_R : \mathbb{N} \to (0,1)$, $p(N) = (p_R(N) : R \in \sigma)$, and $\overline{\delta}(N) = \overline{\delta}(Q, w_R(N))$ for $w_R(N) = \log_2(1/p_R(N))$. Let D be drawn from $\mathscr{D}(N, p(N))$ and let X denote the size of $Q(D)$.*

(1) If $\overline{\delta}(N) = \log N - \omega(1)$, then $X \sim N^n \prod_{R\in\sigma} p_R(N)$ almost surely.

(2) If $\overline{\delta}(N) = \log N + \omega(1)$, then $X = 0$ almost surely.

In certain occasions, the concentration defined by "$X \sim A$ almost surely" is not enough. For example, it may sometimes be necessary to conclude that $\Pr[\,|X - A| \leq \varepsilon A\,] \geq 1 - N^{-d}$ for every $\varepsilon > 0$ and $d > 0$ in order to apply a union bound that involves a number of cases that grows polynomially with N. It turns out that such a strong concentration can also be guaranteed at the expense of a wider *threshold width* in Theorem 5: instead of $\log_2(N) - \omega(1)$ vs $\log_2(N) + \omega(1)$, we require $\log_2(N) - \omega(\log\log(N))$ vs $\log_2(N) + \omega(1)$. This does not follow from Chebyshev's inequality, and for the proof, which can be found in Appendix C, we use the polynomial concentration inequality from [11].

We conclude this section with the max-flow construction to compute $\overline{\delta}(Q, w)$. For every real number $\delta > 0$, we build a network $N(\delta)$ as follows. The network has a source s, a target t, and $|A| + |\sigma|$ intermediate nodes. There is a link of capacity δ between s and each $a \in A$. Each $a \in A$ has a link of infinite capacity to each $R \in \sigma$ with $a \in A_R$. And each $R \in \sigma$ is linked to t with capacity w_R. Recall that a cut in the network is a set of links that disconnects the target from the source. The capacity of the cut is the sum of the capacities of the links in it. Let $\gamma(Q, w, \delta)$ be the minimum capacity of all cuts of $N(\delta)$. It holds that $\gamma(Q, W, \delta) < \sum_{R\in\sigma} w_R$ if and only if $\overline{\delta}(Q, w) > \delta$. The proof of this is rather easy and can be found in Lemma 21 in Appendix C. By the max-flow min-cut algorithm, it follows that $\overline{\delta}(Q, w)$ is computable in polynomial time.

4.2. Execution plans. Theorem 4 shows that certain queries admit a join-project plan that cannot be converted into a join plan without causing a superpolynomial increase in the worst-case running time. The following result shows that when we are considering average-case running time, projections may be eliminated at a very small expected cost.

Theorem 6. *Let Q be a join query with schema σ and n attributes. For every $R \in \sigma$, let $p_R : \mathbb{N} \to (0,1)$, $p(N) = (p_R(N) : R \in \sigma)$. Let D be drawn from $\mathscr{D}(N, p(N))$. If there is a join-project plan φ for Q such that $E(|\varphi'(D)|) \leq T$ for every subplan φ' of φ, then there is a join plan ψ for Q such that $E(|\psi'(D)|) \leq c_\varphi T$ for every subplan ψ' of ψ, where c_φ is a constant depending only on φ.*

The join plan ψ is obtained by iteratively using a procedure that is capable of reducing the number of projections by one in such a way that the expected size of each subplan increases only by a factor depending only on the query. In each iteration, the procedure selects a subplan $\pi_X(\varphi_0)$ of φ such that φ_0 contains no projections, i.e. this projection π_X is lowest in the tree representation of φ. In the first step of the procedure, we replace φ_0 with a join plan φ_0' that contains only those relations appearing in φ_0 whose attributes are completely contained in X^*, where $X^* \supseteq X$ is an appropriate subset of attributes. In the second step, the projection π_X is removed (or, in other words, π_X is replaced by π_{X^*}, making it redundant). The key step of the algorithm is choosing the right X^*. If X^* is too small, then φ_0' is much less restrictive than φ_0, hence $|\pi_X(\varphi_0'(D))|$ can be much larger than $|\pi_X(\varphi_0(D))|$. On the other hand, if X^* is too large, then $|\pi_{X^*}(\varphi_0'(D))|$ can be much larger than $|\pi_X(\varphi_0'(D))|$.

Let $S \subseteq \sigma$ be a set of relations over the attributes A and denote by A_R the attributes of a relation R. For a subset $X \subseteq A$, let

$$f_S(X) := |X|(\log N - n - 1) - \sum_{R \in S[X]} w_R(N),$$

where the set $S[X]$ contains those relations $R \in S$ whose attributes are contained in X. It is easy to see that $f_S(X)$ is submodular (we assume that N is sufficiently large such that $\log N - n - 1$ is positive). It follows that X has a unique minimum-value extension:

Proposition 7. *For every $X \subseteq A$ and $S \subseteq \sigma$, there is a unique $C_S(X) \supseteq X$ such that $f_S(C_S(X))$ is minimal and, among such sets, $|C_S(X)|$ is maximal.*

The following two lemmas explain why $X^* = C_S(X)$ is the right choice. The first lemma shows that we do not get many additional tuples if we take the join of only those relations whose attributes are in $C_S(X)$.

Lemma 8. *Let $S \subseteq \sigma$ be a set of relation names and X a set of attributes. Let $X^* = C_S(X)$ and let $S' = S[X^*]$. For every X^*-tuple t,*

$$\Pr(t \in \pi_{X^*}(\bowtie_{R \in S} R(D)) \mid t \in \bowtie_{R \in S'} R(D)) \geq 1/2.$$

Proof. By definition, $t \in \pi_{X^*}(\bowtie_{R \in S} R(D))$ if and only if there is a $t' \in \bowtie_{R \in S} R(D)$ with $\pi_{X^*}(t') = t$. Note that if $t \in \bowtie_{R \in S'} R(D)$ and $\pi_{X^*}(t') = t$, then t' satisfies all the relations in S', hence the probability that such a t' is in $\bowtie_{R \in S} R(D)$ (assuming $t \in \bowtie_{R \in S'} R(D)$) depends only on the relations in $S \setminus S'$. We claim that this conditional probability is equal to the probability that a certain query Q' with schema σ' has at least one solution. The query Q' is over the attributes $A \setminus X^*$. The schema σ' contains a relational symbol R' for each $R \in S \setminus S'$; the set of attributes of R' is $A_R \setminus X^*$. We define the probability of placing a tuple into R' as $p_{R'}(N) = p_R(N)$ for every $R' \in \sigma'$. It is not difficult to see that $\Pr(t \in \pi_{X^*}(\bowtie_{R \in S} R(D)) \mid t \in \bowtie_{R \in S'} R(D))$ is equal to the probability that Q' has at least one solution.

Observe that if A' is a subset of the attributes in Q', then the relations in $\sigma'[A']$ were obtained from the relations in $S[X^* \cup A'] \setminus S[X^*]$, which means that the weight of these relations is counted in $f_S(X^* \cup A')$ but not in $f_S(X^*)$. If the weight of the relations in $\sigma'[A']$ is greater than $|A'|(\log N - n - 1)$, then $f_S(X^* \cup A') < f_S(X^*)$ would follow, contradicting the minimality of $X^* = C_S(X)$. This means that the maximum density $\overline{\delta}$ of Q' is at most $\log N - n - 1$. Thus by (4.2), the variance of the number of solutions is at most $(2^{n'} - 1) \cdot 2^{\overline{\delta} - \log N} < (2^{n'} - 1)2^{-(n+1)} \leq 1/2$ times the square of the expected number of solutions. Therefore, by Chebyshev's Inequality, the probability that there is no solution is at most $1/2$. $\square$

The second lemma shows that extending the projection from X to X^* does not increase the number of tuples too much: a tuple in the projection to X do not have too many extensions to X^*.

Lemma 9. *Let $S \subseteq \sigma$ be a set of relation names and X be a set of attributes. Let $X^* = C_S(X)$ and let $S' = S[X^*]$. For an X-tuple t, let L_t be the set of those X^*-tuples $t' \in \bowtie_{R\in S'} R(D)$ that have $\pi_X(t') = t$.*

(1) For every X-tuple t, $\mathrm{E}(|L_t| \mid t \in \pi_X(\bowtie_{R\in S'} R(D))) \leq 2^{n(n+2)}$.

(2) For every X^*-tuple t', $\Pr(t' \in \bowtie_{R\in S'} R(D) \mid \pi_X(t') \in \pi_X(\bowtie_{R\in S'} R(D))) \leq 2^{n(n+2)}N^{-|X^*\setminus X|}$.

Proof. Let t', t'' be two X^*-tuples with $\pi_X(t') = \pi_X(t'') = t$. The conditional probability $\Pr(t' \in L_t \mid t'' \in L_t)$ depends on the set $X \subseteq Y \subseteq X^*$ of attributes where t' and t'' are the same. Let w be the total weight of the relations in S' and let w_Y be the total weight of the relations in $S'[Y]$. Observe that $w - w_Y \geq |X^* \setminus Y|(\log N - n - 1)$: otherwise $f_S(Y)$ would be strictly less than $f_S(X^*)$, contradicting the minimality of $X^* = C_S(X)$. If $t'' \in L_t$, then every relation $R \in S'[Y]$ is satisfied by t' as well. Thus $\Pr(t' \in L_t \mid t'' \in L_t) = 2^{-(w-w_y)}$. The number of X^*-tuples that agree with t'' exactly on the attributes in Y is $(N-1)^{|X^*\setminus Y|}$, hence the expected number of such tuples in L_t, on the condition that $t'' \in L_t$, is

$$(N-1)^{|X^*\setminus Y|} \cdot 2^{-(w-w_Y)} \leq N^{|X^*\setminus Y|} \cdot 2^{-|X^*\setminus Y||(\log N - n-1)} \leq 2^{|X^*\setminus Y|(n+1)} \leq 2^{n(n+1)}.$$

Summing for every Y, we get that $\mathrm{E}(|L_t| \mid t'' \in L_t) \leq 2^n 2^{n(n+1)} = 2^{n(n+2)}$.

Let $t_1, \ldots, t_k$ be an ordering of the X^*-tuples whose projection to X is t; clearly, $k = N^{|X^*\setminus X|}$. Let X_i be the event $t_i \in \bowtie_{R\in S'} R(D)$ and Y_i be the event $\bigwedge_{j=1}^{i-1} t_j \notin \bowtie_{R\in S'} R(D)$. The event $t \in \pi_X(\bowtie_{R\in S'} R(D))$ is the disjoint union of the events $X_1Y_1, X_2Y_2, \ldots, X_kY_k$. Now we have

$$\mathrm{E}(|L_t| \mid t \in \pi_X(\bowtie_{R\in S'} R(D))) = \frac{\sum_{i=1}^k \Pr(X_iY_i)\,\mathrm{E}(|L_t| \mid X_iY_i)}{\sum_{i=1}^k \Pr(X_iY_i)} \leq \frac{\sum_{i=1}^k \Pr(X_iY_i)\,\mathrm{E}(|L_t| \mid X_i)}{\sum_{i=1}^k \Pr(X_iY_i)} \leq 2^{n(n+2)}.$$

The first inequality can be obtained as a consequence of the FKG Inequality (see Lemma 24 in the Appendix): $|L_t|$ is a monotone function of the random variables, the event X_i is the product of some random variables, and Y_i is an antimonotone function of the random variables.

To prove the second statement, observe first that if we fix a tuple t, then by symmetry, $\Pr(t' \in \bowtie_{R\in S'} R(D) \mid t \in \pi_X(\bowtie_{R\in S'} R(D)))$ has the same value for every X^*-tuple with $\pi_X(t') = t$. There are $N^{|X^*\setminus Y|}$ such tuples t' and the size of L_t is the sum of the indicator variables corresponding to these tuples. It follows that $\Pr(t' \in \bowtie_{R\in S'} R(D) \mid t \in \pi_X(\bowtie_{R\in S'} R(D))) = \mathrm{E}(|L_t| \mid t \in \pi_X(\bowtie_{R\in S'} R(D)))N^{-|X^*\setminus Y|} \leq 2^{n(n+2)}N^{-|X^*\setminus Y|}$. $\square$

While the statements of the previous two lemmas give the main intuition underlying the construction, the proof of Theorem 6 still requires a substantial amount of work, which we defer to Appendix D.2.

5 Size constraints

To estimate the size of joins, practical query optimisers use statistical information about the database instance such as the sizes of the relations, the sizes of some of their projections, or histograms. We consider the simplest such setting where the size of the relations is known, and we prove a (worst case) estimate on the size of $Q(D)$ subject to the constraint that the relations in D have the given sizes.

Let Q be a join query with schema σ. For every $R \in \sigma$, let A_R be the set of attributes of R, and $A = \bigcup_R A_R$. For every $R \in \sigma$, let N_R be a natural number, and let $L_Q(N_R : R \in \sigma)$ be the following linear program:

$$\begin{array}{lll} \text{minimise} & \sum_R x_R \cdot \log N_R & \\ \text{subject to} & \sum_{R:a\in A_R} x_R \geq 1 & \text{for all } a \in A, \\ & x_R \geq 0 & \text{for all } R \in \sigma. \end{array} \tag{5.1}$$

Note that the only difference with L_Q as defined in (3.1) is the objective function. This implies that every feasible solution of $L_Q(N_R : R \in \sigma)$ is also a fractional edge cover of Q.

Theorem 10. *Let Q be a join query with schema σ and let $N_R \in \mathbb{N}$ for all $R \in \sigma$. Let n be the number of attributes of Q, and let $(x_R : R \in \sigma)$ be an optimal solution of the linear program $L_Q(N_R : R \in \sigma)$.*

(1) For every σ-instance D with $|R(D)| = N_R$ for all R it holds that $|Q(D)| \leq \prod_R N_R^{x_R}$

(2) There is a σ-instance D such that $|R(D)| = N_R$ for all $R \in \sigma$ and $|Q(D)| \geq 2^{-n} \prod_R N_R^{x_R}$.

Proof sketch. Statement (1) is an immediate consequence of Lemma 2. Statement (2) is proved similarly to Lemma 3. The larger gap between upper and lower bounds is due to fact that we cannot choose the size N of the relations freely and thus in particular cannot guarantee that $N^{p_i/q}$ is integral. This causes a rounding error. The full proof can be found in Appendix E. □

In the next example we show that we cannot replace the lower bound of Theorem 10(2) by $2^{-(1-\varepsilon)n} \prod_R N_R^{x_R}$ for any $\varepsilon > 0$. This seems to indicate that maybe the approach to estimating the size of joins through fractional edge covers is no longer appropriate in the setting where the size of the relations is fixed. However, we shall then see that, in some sense, there is no better approach. In Theorem 12, we shall prove that there is no polynomial time algorithm that, given a query Q and relation sizes N_R, for $R \in \sigma$, approximates the worst case size of the query answer to a factor better than $2^{n^{1-\varepsilon}}$.

Example 11. Let $n \in \mathbb{N}$ be an integer, $0 < \varepsilon < 1$ a fixed constant, and $A = \{a_1, \ldots, a_n\}$ a set of attributes with domain $\mathbb{N}$. Let $r := \lfloor \varepsilon n / \log n \rfloor$. We assume that n is sufficiently large that $2^r > n$ holds. For every $B \in \binom{[n]}{r}$, let R_B be an r-ary relation with attributes B. Furthermore, for every $a \in A$, let R_a be a unary relation with the only attribute a. Let Q be the join of all these relations and let σ be the resulting schema.

For every $B \in \binom{[n]}{r}$, let $N_{R_B} = 2^r - 1$ and for every $a \in A$, let $N_{R_a} = 2$. Consider the linear program $L_Q(N_R : R \in \sigma)$. We obtain an optimal solution for this linear program by letting $x_{R_B} := n / \left(r\binom{n}{r}\right)$ and $x_{R_a} := 0$; optimality can easily be proved by considering the dual linear program.

We prove next that $\prod_R N_R^{x_R} = 2^n(1 - o(1))$:

$$\prod_{R \in \sigma} N_R^{x_R} = \left((2^r - 1)^{n/\left(r\binom{n}{r}\right)}\right)^{\binom{n}{r}} \geq (2^r - 1)^{n/r} \geq (2^r(1 - 1/n))^{n/r} = 2^n \,(1 - 1/n)^{n/r} = 2^n(1 - o(1)).$$

The second inequality follows from $2^r > n$ and the last equality follows from the fact if n tends to infinity, then $(1 - 1/n)^n$ goes to $1/e$ and r goes to infinity as well.

To complete the example, we prove that $|Q(D)| \leq 2^{\varepsilon n}$ for every instance D respecting the constraints N_R. Let D be a σ-instance with $|R(D)| = N_R$ for every $R \in \sigma$. From $N_{R_a} = 2$ it follows that in $Q(D)$ each attribute has at most two values, hence we can assume that $Q(D) \subseteq \{0,1\}^n$. Thus each tuple in $t \in Q(D)$ can be viewed as a subset $A_t = \{a \in A : t(a) = 1\}$ of A. For every $B \in \binom{[n]}{r}$, it holds $\pi_B(Q(D)) \leq N_{R_B} = 2^r - 1$, hence the Vapnik-Chervonenkis dimension of $Q(D)$ is less than r. Thus by Sauer's Lemma, we have

$$|Q(D)| \leq n^r \leq n^{\varepsilon n / \log n} = 2^{\varepsilon n},$$

as claimed.

Theorem 12. *For a given query Q with schema σ and a given set of size constraints $(N_R : R \in \sigma)$, denote by M the maximum of $|Q(D)|$ over databases satisfying $|R(D)| = N_R$ for every $R \in \sigma$. If for some $\varepsilon > 0$, there is a polynomial-time algorithm that, given a query Q with n attributes and size constraints N_R, computes two values M_L and M_U with $M_L \leq M \leq M_U$ and $M_U \leq M_L 2^{n^{1-\varepsilon}}$, then* NP = ZPP.

The proof, which is based on a reduction from the independent set problem, can be found in Appendix E.

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley, second edition, 1992.

[3] S. Chaudhuri. An overview of query optimization in relational systems. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 34–43, 1998.

[4] 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.

[5] J. Flum, M. Frick, and M. Grohe. Query evaluation via tree-decompositions. *Journal of the ACM*, 49(6):716–752, 2002.

[6] E. Friedgut and J. Kahn. On the number of copies of a hypergraph in another. *Israel Journal of Mathematics*, 105:251–256, 1998.

[7] H. Garcia-Molina, J. Widom, and J.D. Ullman. *Database System Implementation*. Prentice-Hall, 1999.

[8] G. Graefe. Query evaluation techniques for large databases. *ACM Computing Surveys*, 25, 1993.

[9] M. Grohe and D. Marx. Constraint solving via fractional edge covers. In *Proceedings of the of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 289–298, 2006.

[10] Johan Håstad. Clique is hard to approximate within $n^{1-\varepsilon}$. *Acta Math.*, 182(1):105–142, 1999.

[11] J. H. Kim and V. H. Vu. Concentration of multivariate polynomials and its applications. *Combinatorica*, 20(3):417–434, 2000.

[12] J. Rhadakrishnan. Entropy and counting. Available at http://www.tcs.tifr.res.in/∼jaikumar/mypage.html.

[13] V.V. Vazirani. *Approximation Algorithms*. Springer-Verlag, 2001.

A Proofs omitted in Section 3.1

The proof of Lemma 2 is based on a combinatorial lemma known as Shearer's lemma. The lemma appeared first in [4], where it is attributed to Shearer. The *entropy* of a random variable X with range U is

$$h[X] := -\sum_{x \in U} \Pr(X = x) \log \Pr(X = x)$$

Shearer's lemma gives an upper bound of a distribution on a product space in terms of its marginal distributions.

Lemma 13 (Shearer's Lemma). *Let $X = (X_i \mid i \in I)$ be a random variable, and let A_j, for $j \in J$, be (not necessarily distinct) subsets of the index set I such that each $i \in I$ appears in at least k of the sets A_j. For every $B \subseteq I$, let $X_B = (X_i \mid i \in B)$. Then*

$$\sum_{j=1}^{m} h[X_{A_j}] \geq k \cdot h[X].$$

A simple proof of the lemma can be found in [12].

Now we are ready to prove Lemma 2:

Proof of Lemma 2. Let A_R be the set of attributes of $R \in \sigma$ and $A = \bigcup_R A_R$. Without loss of generality we may assume that the fractional edge cover x_R only takes rational values, because the rationals are dense in the reals. Let p_R and q be nonnegative integers such that $x_R = p_R/q$. Let $m = \sum_R p_R$, and let $A_1, \dots, A_m$ be a sequence of subsets of A that contains precisely p_R copies of the set A_R, for all $R \in \sigma$. Then every attribute $a \in A$ is contained in at least q of the sets A_i, because

$$\left|\{i \in [m] : a \in A_i\}\right| = \sum_{R : a \in A_R} p_R \;=\; q \cdot \sum_{R : a \in A_R} x_R \;\geq\; q.$$

Let $X = (X_a \mid a \in A)$ be uniformly distributed on $Q(D)$, which we assume to be non-empty as otherwise the claim is obvious. That is, for every tuple $t \in Q(D)$ we have $\Pr(X = t) = 1/|Q(D)|$, and for all other A-tuples we have $\Pr(X = t) = 0$. Then $h[X] = \log|Q(D)|$. We apply Shearer's Lemma to the random variable X and the sets A_R, for $R \in \sigma$. (Thus we have $I = A$ and $J = \sigma$.) Note that for every $R \in \sigma$ the marginal distribution of X on A_R is 0 on all tuples not in $R(D)$. Hence the entropy of X_{A_R} is bounded by the entropy of the uniform distribution on $R(D)$, that is, $h[X_{A_R}] \leq \log|R(D)|$. Thus by Shearer's Lemma, we have

$$\sum_R p_R \cdot \log|R(D)| \geq \sum_{R \in \sigma} h[X_{A_R}] \geq q \cdot h[X] = q \cdot \log|Q(D)|.$$

It follows that

$$|Q(D)| \leq 2^{\sum_R (p_R/q) \cdot \log|R(D)|} = \prod_R |R(D)|^{x_R}.$$

□

Proof of Lemma 3. Let A_R be the set of attributes of $R \in \sigma$ and $A = \bigcup_R A_R$. Recall $(x_R : R \in \sigma)$ is an optimal solution for the linear program (3.1). By LP-duality, there is a solution $(y_a : a \in A)$ for the dual linear program

$$\begin{array}{lll} \text{maximise} & \sum_a y_a & \\ \text{subject to} & \sum_{a \in A_R} y_a \leq 1 & \text{for all } R \in \sigma, \\ & y_a \geq 0 & \text{for all } a \in A \end{array} \tag{1.1}$$

such that $\sum_a y_a = \sum_R x_R$. There even exists such a solution with rational values.

We take an optimal solution $(y_a : a \in A)$ with $y_a = p_a/q$, where $q \geq 1$ and $p_a \geq 0$ are integers. Let $N_0 \in \mathbb{N}$, and let $N = N_0^q$. We define a σ-instance D by letting

$$R(D) := \left\{ t \in \mathrm{tup}(A_R) : t(a) \in [N^{p_a/q}] \text{ for all } a \in A_R \right\}$$

for all $R \in \sigma$. Here we assume that $\mathrm{dom}(a) = \mathbb{N}$ for all attributes a. As there is at least one a with $y_a > 0$ and hence $p_a \geq 1$, we have $|D| \geq N^{1/q} = N_0$. Observe that

$$|R(D)| = \prod_{a \in A_R} N^{p_a/q} = N^{\sum_{a \in A_R} y_a}$$

for all $R \in \sigma$. Furthermore, $Q(D)$ is the set of all tuples $t \in \mathrm{tup}(A)$ with $t(a) \in [N^{p_a/q}]$ for every $a \in A$. Hence

$$|Q(D)| = \prod_{a \in A} N^{p_a/q} = N^{\sum_{a \in A} y_a} = N^{\sum_{R \in \sigma} x_R} = \prod_{R \in \sigma} N^{x_R}.$$

By complementary slackness of linear programming we have

$$\sum_{a \in A_R} y_a = 1 \quad \text{for all } R \in \sigma \text{ with } x_R = 0.$$

Thus $|R(D)| = N$ for all $R \in \sigma$ with $x_R > 0$ and

$$|Q(D)| = \prod_{R \in \sigma} N^{x_R} = \prod_{R \in \sigma} |R(D)|^{x_R}.$$

□

B Proofs omitted from Section 3.2

We define the *size* of a k-ary relation R to be the number $||R|| := |R| \cdot k$. The bounds stated in the following fact depend on the machine model; the fact is based on a standard random access machines with a uniform cost measure. Other models may require additional logarithmic factors. For details and a proof of the fact, we refer the reader to [5].

Fact 14. *The following hold:*

(1) The join $R \bowtie S$ of two relations R and S can be computed in time $O(||R|| + ||S|| + ||R \bowtie S||)$.

(2) The projection $\pi_B(R)$ of an A-relation R to a subset $B \subseteq A$ can be computed in time $O(||R||)$.

Theorem 15. *For every join query Q, there is a join-project plan for Q that can be evaluated in time $O\big(|Q|^2 \cdot |D|^{\rho^*(Q)+1}\big)$ on a given instance D. Moreover, there is a polynomial-time algorithm that, given Q, computes the join-project plan.*

Proof of Theorem 15. Let $Q = R_1(A_1) \bowtie \cdots \bowtie R_m(A_m)$ be a join query and D an instance for Q. Suppose that the attributes of Q are $\{a_1, \ldots, a_n\}$. For $i \in [n]$, let $B_i := \{a_1, \ldots, a_i\}$. Furthermore, let

$$\begin{aligned} \varphi_1 &:= \big(\cdots (\pi_{B_1}(R_1) \bowtie \pi_{B_1}(R_2)) \bowtie \cdots \bowtie \pi_{B_1}(R_m) \big), \\ \varphi_{i+1} &:= \Big(\cdots \big((\varphi_i \bowtie \pi_{B_{i+1}}(R_1)) \bowtie \pi_{B_{i+1}}(R_2) \big) \bowtie \cdots \bowtie \pi_{B_{i+1}}(R_m) \Big) \qquad \text{for all } i \geq 1. \end{aligned}$$

It is easy to see that for every $i \in [n]$ it holds that $\varphi_i(D) = \pi_{B_i}(Q(D))$ and hence $\varphi_n(D) = Q(D)$. Hence to compute $Q(D)$, we can evaluate the join-project plan φ_n.

To estimate the cost of the evaluating the plan, we need to establish the following claim:

For every $i \in [n]$ we have $|\varphi_i(D)| \leq |D|^{\rho^*(Q)}$.

To see this, we consider the join query

$$Q_i := R_1^i \bowtie \cdots \bowtie R_m^i,$$

where ${R_j}^i$ is a relation name with attributes $B_i \cap A_j$. The crucial observation is that $\rho^*(Q_i) \leq \rho^*(Q)$, because if $(x_R : R \in \sigma)$ is fractional edge cover of Q, then letting $x_{R^i} = x_R$ for every $R \in \sigma$ we get a fractional edge cover of Q_i of the same cost. If we let D_i be the database instance with ${R_j}^i(D_i) := \pi_{B_i}(R_j)$ for all $j \in [m]$, then we get

$$\varphi_i(D) = Q_i(D_i) \leq |D_i|^{\rho^*(Q_i)} \leq |D|^{\rho^*(Q)}.$$

This proves the claim.

We further observe that all intermediate results in the computation of $\varphi_{i+1}(D)$ from $\varphi_i(D)$ are contained in

$$\varphi_i(D) \times U,$$

where U is the active domain of D. Hence their size is bounded by $|\varphi_i(D)| \cdot |D| \leq |D|^{\rho^*(Q)+1}$, and by Fact 14 they can be computed in time $O(|D|^{\rho^*(Q)+1})$. Overall, we have to compute $n \cdot m$ projections, each requiring time $O(D)$, and $n \cdot m$ joins, each requiring time $O(|D|^{\rho^*(Q)+1})$. This yields the desired running time. □

Proof of Theorem 4. Let $n = \binom{2m}{m}$. For every $s \subseteq [2m]$ with $|s| = m$, let a_s be an attribute with domain $\mathbb{N}$. For every $i \in [2m]$, let R_i be a relation name having as attributes all a_s such that $i \in s$. Let A_i be the set of attributes of R_i and $A = \bigcup_{i \in [2m]} A_i$. The arity of R_i is

$$|A_i| = \binom{2m-1}{m-1} = \frac{m}{2m} \cdot \binom{2m}{m} = \frac{n}{2}.$$

Let $Q := R_1 \bowtie \cdots \bowtie R_{2m}$. Then $|Q| = 2m \cdot n/2 = m \cdot n$. Furthermore, $\rho^*(Q) \leq 2$. To see this, let $x_{R_i} = 1/m$ for every $i \in [2m]$. This forms a fractional edge cover of Q, because for every $s \subseteq [2m]$ with $|s| = m$, the attribute a_s appears in the m atoms R_i with $i \in s$.

Next, we define an instance D by letting $R_i(D)$ be the set of all A_i-tuples that have an arbitrary value from $[N]$ in one coordinate and 1 in all other coordinates. Formally,

$$R_i(D) := \bigcup_{a \in A_i} \bigcap_{b \in A_i \setminus a} \{t \in \mathrm{tup}(A_i) : t(a) \in [N], t(b) = 1\}.$$

Observe that $|R_i(D)| = (N-1)n/2 + 1$ for all $i \in [2m]$ and thus

$$|D| = (N-1)mn + 2m \geq N.$$

Furthermore, $Q(D)$ is the set of all A-tuples that have an arbitrary value from $[N]$ in one coordinate and 1 in all other coordinates. Hence $|Q(D)| = (N-1)n + 1 \leq |D|$. This completes the proof of (1).

To prove (2), we shall use the following simple (and well-known) combinatorial lemma:

Lemma 16. *Let T be a binary tree whose leaves are coloured with $2m$ colours, for some $m \geq 1$. Then there exists a node t of T such that at least $(m+2)/2$ and at most $m+1$ of the colours appear at leaves that are descendants of t.*

Proof. For every node t of T, let $c(t)$ be the number of colours that appear at descendants of T. The *height* of a node t is the length of the longest path from t to a leaf.

Let t be a node of minimum height such that $c(t) \geq m+2$, and let u_1, u_2 be the children of t. (Note that t cannot be a leaf because $c(t) \geq 2$.) Then $c(u_i) \leq m+1$ for $i = 1, 2$. Furthermore, $c(u_1) + c(u_2) \geq c(t)$, hence $c(u_i) \geq (m+2)/2$ for at least one i. □

Continuing the proof of the theorem, we let φ be a join plan for Q. We view the term φ as a binary tree T whose leaves are labelled by atoms R_i. We view the atoms as colours. Applying the lemma, we find a node t of T such that at least $(m+2)/2$ and at most $m+1$ of the colours appear at leaves that are descendants of t. Every inner node of the tree corresponds to a subplan of φ. We let ψ be the subplan corresponding to t. Then at least $(m+2)/2$ and at most $m+1$ atoms R_i appear in ψ. By symmetry, we may assume without loss of generality that the atoms of ψ are $R_1,\ldots,R_\ell$ for some $\ell \in \left[\lceil (m+2)/2 \rceil, m+1\right]$. Hence ψ is a plan for the join query

$$R_1 \bowtie \cdots \bowtie R_\ell.$$

Let $B := \bigcup_{i=1}^{\ell} A_i$ be the set of all attributes occurring in ψ. For $i \in [m+1]$, let $s_i = \{i\} \cup [m+2, 2m]$. Then for all $i, j \in [\ell]$ we have $a_{s_i} \in A_j$ if and only if $i = j$. Hence all tuples $t \in \mathrm{tup}(B)$ with $t(a_{s_i}) \in [N]$ for all $i \in [\ell]$ and $t(b) = 1$ for all $b \in B \setminus \{a_{s_1},\ldots,a_{s_\ell}\}$ are contained in $\psi(D)$. As there are N^ℓ such tuples, it follows that

$$|\psi(Q)| \geq N^\ell \geq N^{(m+2)/2}.$$

Statement (2) of the lemma follows, because

$$\log|Q| = \log m + \log n \leq \log m + \log 2^{2m} = \log m + 2m \leq 5 \cdot (m+2)/2,$$

provided m is large enough, which we may assume without loss of generality. ☐

Proposition 17. *For every query Q there is a join plan φ that can be evaluated in time $O(|D|^{2\rho^*(Q)\cdot\log|Q|})$.*

Proof. Let Q be a join query with schema σ and attributes A. An *edge cover* of Q is a subset $\gamma \subseteq \sigma$ such that $A \subseteq \bigcup_{R\in\gamma} A_R$, where A_R is the set of attributes of R. The *edge cover number* $\rho(Q)$ of Q is minimum size of an edge cover for Q. Observe that edge covers correspond to $\{0,1\}$-valued fractional edge covers and that the edge cover number is precisely the cost of the optimal integral fractional edge cover. It is well known that the integrality gap for the linear program defining fractional edge covers is H_n, where $n = |A|$ and H_n is the nth harmonic number (see, for example, [13], Chapter 13). It is known that $H_n \leq 2\log n$. Now the join plan consists in first joining the relations that form an edge cover of size $2\rho^*(Q)\cdot\log|Q|$ in arbitrary order, and then joining the result with the rest of relations in arbitrary order. ☐

C Proofs omitted from Section 4.1

For a random variable X, the expression "$X \sim A$ polynomially almost surely" means that for every $\varepsilon > 0$ and $d > 0$, there exists N_0 such that for every $N \geq N_0$ we have $\Pr\left[|X - A| \leq \varepsilon A\right] \geq 1 - N^{-d}$. The goal is to prove the following counterpart to Theorem 5.

Theorem 18. *Let Q be a join query with schema σ and n attributes. For every $R \in \sigma$, let $p_R : \mathbb{N} \to (0,1)$, $p(N) = (p_R(N) : R \in \sigma)$, and $\overline{\delta}(N) = \overline{\delta}(Q, w_R(N))$ for $w_R(N) = \log_2(1/p_R(N))$. Let D be drawn from $\mathscr{D}(N, p(N))$ and let X denote the size of $Q(D)$.*

(1) If $\overline{\delta}(N) = \log N - \omega(\log\log N)$, then $X \sim N^n \prod_{R\in\sigma} p_R(N)$ polynomially almost surely.

(2) If $\overline{\delta}(N) = \log N + \omega(1)$, then $X = 0$ almost surely.

For the proof, we will use the polynomial concentration method from [11]. Let $H = (V, E)$ be a hypergraph with $n = |V|$ and $k = \max_{e\in E}|e|$. For every $e \in E$, let $w(e)$ be a positive weight. Let $\{X_u : u \in V\}$ be a collection of mutually independent random variables, where each X_u is an indicator random variable with expected value p_u. Here $0 \leq p_u \leq 1$ for every $u \in V$. Let M be the following polynomial: $M = \sum_{e\in E} w(e)\prod_{u\in e} X_u$. For every $Y \subseteq V$, let M_Y be the partial derivative of M with respect to $\{X_u : u \in Y\}$, that is, $M_Y = \sum_{e\in E\colon Y\subseteq e} w(e)\prod_{u\in e\setminus Y} X_u$. Let $E_Y = \mathrm{E}[M_Y]$, and for every $i \in \{0,\ldots,k\}$, let $E_i = \max\{E_Y : Y \subseteq V, |Y| = i\}$. Note that $E_0 = \mathrm{E}[M]$. Let $E' = \max\{E_i : 1 \leq i \leq k\}$ and $E = \max\{E_0, E'\}$.

Theorem 19 (Theorem 7.8.1 in [2]). *For every $\lambda > 1$, it holds*

$$\Pr\left[\ |M - \mathrm{E}[M]| > a_k(EE')^{1/2}\lambda^k\ \right] < d_k e^{-\lambda} n^{k-1},$$

where $a_k = 8^k\sqrt{k!}$ and $d_k = 2e^2$.

For a complete proof of Theorem 18, it is easier to first state the following consequence to Theorem 19 (see also Corollary 4.1.3 in [11]):

Corollary 20. *For every two reals $\varepsilon > 0$ and $d > 0$, and every integer $k \geq 1$, there exists n_0 such that, in the setting of Theorem 19, if $n \geq n_0$ and $E_i/E_0 \leq (\log n)^{-4k}$ for every $i \in \{1,\dots,k\}$, then*

$$\Pr[\ |M - \mathrm{E}[M]| > \varepsilon\,\mathrm{E}[M]\ ] \leq n^{-d}$$

For a proof, it suffices to choose $\lambda(n) = (\varepsilon/a_k)^{1/k}(\log n)^2$ and let n_0 be such that $\lambda > 1$ and $d_k e^{-\lambda} n^{k-1} < n^{-d}$ for every $n \geq n_0$. Note that if $E_i/E_0 \leq (\log n)^{-4k}$ then $E = E_0$ and $E' \leq E_0(\log n)^{-4k}$, so $a_k(EE')^{1/2}\lambda^k \leq \varepsilon E_0$ and the claim follows.

Proof of Theorem 18. Let A be the set of attributes of Q. We start with (2). Suppose that $\overline{\delta}(N) = \log N + \omega(1)$ and fix a large N. Let $B \subseteq A$, $B \neq \emptyset$, be such that $\overline{\delta}(Q, w(N)) = \delta(Q[B], w(N))$. Let $Q_B = Q[B]$, let σ_B be the schema of Q_B, and let $M_B = |Q_B(D)|$. The expectation of M_B is

$$N^{|B|}\prod_{R\in\sigma_B} p_R(N) = 2^{|B|(\log N - |B|^{-1}\sum_{R\in\sigma_B}\log(1/p_R(N)))} = 2^{|B|(\log N - \delta(Q_B))}.$$

Since $\delta(Q_B) = \overline{\delta}(Q)$ and $|B| > 0$, the hypothesis $\overline{\delta}(N) = \log N + \omega(1)$ implies that this quantity approaches 0 as N grows. By Markov's inequality, $M_B = 0$ almost surely, and therefore $M = 0$ almost surely because every solution to Q gives a solution to Q_B.

For (1), suppose that $\overline{\delta}(N) = \log N - \omega(\log\log N)$ and fix a large N. For every $R \in \sigma$, let A_R be the set of attributes of R, and for every $t \in [N]^{A_R}$, let $X(R,t)$ be the indicator random variable for the event $t \in R(D)$. Note that these are mutually independent random variables and $\mathrm{E}[X(R,t)] = p_R(N)$ by the definition of the probability space. Note also that

$$M = \sum_t \prod_{R\in\sigma} X(R, t_R),$$

where t ranges over all tuples in $[N]^A$, and t_R denotes the projection of t to the attributes of R. We aim for an application of Corollary 20 with the random variables $X(R,t)$. We define the hypergraph $H = (V,E)$. The set of vertices V is the set of pairs (R,t) where $R \in \sigma$ and $t \in [N]^{A_R}$. There is one hyperedge e_t in E for every $t \in [N]^A$ that consists of all pairs (R, t_R) with $R \in \sigma$. Thus, the number of vertices is bounded by mN^r, where $m = |\sigma|$ and r is the maximum arity of the relations in σ. Furthermore, the maximum size of the edges in H, that is, the k in Corollary 20, is m.

We have

$$E_0 = N^n \prod_{R\in\sigma} p_R(N). \tag{3.1}$$

Let us bound E_i for $i > 0$. Fix a set of vertices of the hypergraph H, say

$$Y \subseteq \{(R,t) : R \in \sigma, t \in [N]^{r_R}\},$$

with $|Y| = i > 0$. We want to bound E_Y/E_0. Let σ' be the set of relations R that appear in Y. Let B be all the attributes of the relations in σ'. Note that $\sigma' \subseteq \sigma_B$, where σ_B is the schema of $Q[B]$. This will be of use later.

Let T be the set of all $t \in [N]^A$ such that $(R, t_R) \in Y$ for every $R \in \sigma'$, where as before, t_R denotes the projection of t to the attributes of R. If there exist t_1 and t_2 in T that disagree on some attribute of B, then automatically $E_Y = 0$ because then Y is not included in any hyperedge e_t of H and hence $M_Y = 0$. We may assume then that all $t \in T$ agree on B. This implies $|T| \leq N^{n-|B|}$. Under these conditions we have

$$E_Y = \mathrm{E}\left[\sum_{t \in T} \prod_{R \in \sigma - \sigma'} X(R, t_R)\right] \leq N^{n-|B|} \prod_{R \in \sigma - \sigma'} p_R(N).$$

Therefore, recalling (3.1), we bound E_Y / E_0 by

$$N^{-|B|} \prod_{R \in \sigma'} \frac{1}{p_R(N)} \leq N^{-|B|} \prod_{R \in \sigma_B} \frac{1}{p_R(N)} \leq N^{-|B|\left(1 - \frac{1}{\log N} |B|^{-1} \sum_{R \in \sigma_B} \log(1/p_R(N))\right)}, \tag{3.2}$$

where the first inequality holds because $\sigma' \subseteq \sigma_B$ and each p_R belongs to $(0,1)$. Using the hypothesis that $\overline{\delta}(N) = \log N - \omega(\log\log N)$, we bound (3.2) by $(\log N)^{-(4m+1)}$. We showed then that

$$E_Y / E_0 \leq (\log N)^{-(4m+1)},$$

and since this holds for an arbitrary Y of size $i > 0$, the bound is also valid for E_i / E_0. Recall now that the number of vertices $|V|$ of the hypergraph H is at most mN^r, and we can bound

$$(\log N)^{-(4m+1)} \leq (\log(mN^r))^{-4m} \leq (\log|V|)^{-4m}$$

for large N. The result follows from Corollary 20. $\square$

Lemma 21. *The following are equivalent:*

(1) $\gamma(Q, W, \delta) < \sum_{R \in \sigma} w_R$

(2) $\overline{\delta}(Q, w) > \delta$.

Proof of Lemma 21. Assume $\overline{\delta}(Q, w) > \delta$ and let $B \subseteq A$, $B \neq \emptyset$, be such that $\delta(Q[B], w) > \delta$. Let σ_B be the schema of $Q[B]$. Let S be the cut that consists of all links from the source to the nodes of B and the links from the nodes in $\sigma - \sigma_B$ to the target. The capacity of this cut is

$$|B|\delta + \sum_{R \in \sigma - \sigma_B} w_R < \sum_{R \in \sigma_B} w_R + \sum_{R \in \sigma - \sigma_B} w_R = \sum_{R \in \sigma} w_R,$$

where the inequality follows from $\delta(Q[B], w) > \delta$.

Suppose now $\gamma(Q, w, \delta) < \sum_R w_R$ and let S be a cut of minimum capacity. Let B be the set of $a \in A$ for which the link from the source to a is in S. Let σ_B be the schema of $Q[B]$. We claim that S does not contain any link from an $R \in \sigma_B$ to the target. For if it did, $S - \{(R, t)\}$ would also be a cut of smaller capacity. Also S contains all links from an $R \in \sigma - \sigma_B$ to t. For if it did not, S would not be a cut (we assume all R have a non-empty set of attributes). Finally, S does not contain any link from an $a \in A$ to an $R \in \sigma$ because those have infinite capacity. Therefore, the capacity of S is $\delta|B| + \sum_{R \in \sigma - \sigma_B} w_R$ and smaller than $\sum_{R \in \sigma} w_R$ by hypothesis. Hence $B \neq \emptyset$ and $\delta(Q[B], w) > \delta$. $\square$

D Proofs omitted from Section 4.2

D.1. Applications of the FKG-inequality. The FKG Inequality is a general tool for determining the correlation between monotone (antimonotone) events:

Fact 22. *Let V be a set of 0-1 random variables and let $f,g : \{0,1\}^V \to \mathbb{R}$ be monotone functions on these variables. Let $\mu : \{0,1\}^V \to \mathbb{R}^+$ be a function satisfying $\mu(x)\mu(y) \le \mu(x \vee y)\mu(x \wedge y)$ for every $x,y \in \{0,1\}^V$ (where $x \vee y$ and $x \wedge y$ denote the coordinate-wise disjunction and conjunction of the two tuples, respectively.) Then*

$$\left(\sum_{x\in\{0,1\}^V} f(x)g(x)\mu(x)\right)\left(\sum_{y\in\{0,1\}^V} \mu(y)\right) \ge \left(\sum_{x\in\{0,1\}^V} f(x)\mu(x)\right)\left(\sum_{y\in\{0,1\}^V} g(y)\mu(y)\right).$$

A proof can be found in [2].

In general, if A, B, C are three monotone 0-1 functions of a set of random variables, then $\Pr(A = 1 \mid B = 1) \le \Pr(A = 1 \mid BC = 1)$ is not necessarily true. However, it is true in the following special case:

Lemma 23. *Let V be a set of 0-1 random variables, let $M' \subseteq M \subseteq V$ be two subset of these variables, and let $F : \{0,1\}^V \to \{0,1\}$ be a monotone function. Then* $\Pr(F = 1 \mid \prod M = 1) \ge \Pr(F = 1 \mid \prod M' = 1)$.

Proof. For every $x \in \{0,1\}^V$, if every variable of M' is 1 in x, then let $\mu(x)$ be the probability of tuple x, otherwise let $\mu(x) = 0$. It is easy to verify that $\mu(x)\mu(y) \le \mu(x \vee y)\mu(x \wedge y)$ for every $x,y \in \{0,1\}^V$. Let $f = F$ and $g = \prod(M \setminus M')$. With these settings, Theorem 22 implies

$$\Pr(F = 1 \wedge \prod M = 1) \cdot \Pr(\prod M' = 1) \ge \Pr(F = 1 \wedge \prod M' = 1) \cdot \Pr(\prod M = 1)$$
$$\frac{\Pr(F = 1 \wedge \prod M = 1)}{\Pr(\prod M = 1)} \ge \frac{\Pr(F = 1 \wedge \prod M' = 1)}{\Pr(\prod M' = 1)},$$

what we had to show. ☐

The following lemma shows a similar statement, but this time an antimonotone event decreases the conditional expectation of a monotone function:

Lemma 24. *Let V be a set of 0-1 random variables, let $M \subseteq V$ be a subset of these variables, and let $F : \{0,1\}^V \to \mathbb{R}^+$ be a monotone function and $G : \{0,1\}^V \to \{0,1\}$ an antimonotone function. Then* $\mathrm{E}(F \mid \prod M = 1) \ge \mathrm{E}(F \mid \prod M = 1 \wedge G = 1)$.

Proof. The proof is similar to the proof of Lemma 23. For every $x \in \{0,1\}^V$, if every variable of M is 1 in x, then let $\mu(x)$ be the probability of tuple x, otherwise let $\mu(x) = 0$. Let $f = F$ and $g = 1 - G$. Then Theorem 22 implies

$$\mathrm{E}(F \cdot (1-G) \cdot \prod M) \cdot \Pr(\prod M = 1) \ge \mathrm{E}(F \cdot \prod M) \cdot \Pr((1-G) = 1 \wedge \prod M = 1)$$
$$(\mathrm{E}(F \cdot \prod M) - \mathrm{E}(F \cdot G \cdot \prod M)) \cdot \Pr(\prod M = 1) \ge \mathrm{E}(F \cdot \prod M) \cdot (\Pr(\prod M = 1) - \Pr(G = 1 \wedge \prod M = 1))$$
$$-\mathrm{E}(F \cdot G \cdot \prod M) \cdot \Pr(\prod M = 1) \ge -\mathrm{E}(F \cdot \prod M) \cdot \Pr(G = 1 \wedge \prod M = 1)$$
$$\frac{\mathrm{E}(F \cdot G \cdot \prod M)}{\Pr(G = 1 \wedge \prod M = 1)} \le \frac{\mathrm{E}(F \cdot \prod M)}{\Pr(\prod M = 1)}$$
$$\mathrm{E}(F \mid G = 1 \wedge \prod M = 1) \le \mathrm{E}(F \mid \prod M = 1),$$

what we had to prove. ☐

D.2. Proof of Theorem 6. The choice of the random database D can be thought of as a set of independent 0-1 variables (N^r variables describe an r-ary random relation). For a join-project plan φ and a tuple t, let $I_{\varphi(D),t}$ be the indicator random variable that is 1 if and only if $t \in \varphi(D)$; clearly $I_{\varphi(D),t}$ is a monotone function of the random variables describing the database. Since this function is monotone, it can be expressed as the disjunction of minterms, i.e., $I_{\varphi(D),t} = \bigvee_{i=1}^{M} I^{(i)}_{\varphi(D),t}$, where each minterm $I^{(i)}_{\varphi(D),t}$ is the product of a subset of the 0-1 random variables. The *rank* of a monotone function is the maximum size of a minterm of the function. The following two lemmas will be useful for determining conditional probabilities between these monotone functions.

Lemma 25. *Let φ be a join-project plan whose tree has ℓ leaves and let t be a tuple in* $\mathrm{tup}(\varphi)$.

(1) The rank of $I_{\varphi(D),t}$ is at most ℓ.

(2) If φ_0 is a subplan of φ, then $I_{\varphi(D),t}$ can be written as $I_{\varphi(D),t} = \bigvee_{t' \in \mathrm{tup}(\varphi_0)} (I_{\varphi_0(D),t'} \wedge J_{t'})$, where each $J_{t'}$ is a monotone function of the random variables. Moreover, if φ' is obtained from φ by replacing φ_0 with some other subplan φ'_0 satisfying $A_{\varphi_0} = A_{\varphi'_0}$, then $I_{\varphi'(D),t} = \bigvee_{t' \in \mathrm{tup}(\varphi'_0)} (I_{\varphi'_0(D),t'} \wedge J_{t'})$ with the same functions $J_{t'}$.

Proof. Statement 1 can be proved by a simple induction on the size of the tree of φ. First, observe that the rank of disjunction of functions is at most the maximum of the ranks of the functions, while the rank of the conjunction of functions is at most the sum of the ranks of the functions. If the tree consists of a single leaf (i.e., φ consists of a single relation symbol), then $I_{\varphi(D),t}$ is equal to one of the random variables, i.e., its rank is 1. If $\varphi = \pi_X(\varphi^*)$, then $I_{\varphi(D),t} = \bigvee_{t' \in \mathrm{tup}(\varphi^*), \pi_X(t') = \pi_X(t)} I_{\varphi^*(D),t'}$. By induction, the rank of each $I_{\varphi^*,t'}$ is at most ℓ, hence the rank of this disjunction is also at most ℓ. Finally, if $\varphi = \varphi_1 \bowtie \varphi_2$, then $I_{\varphi(D),t} = I_{\varphi_1(D),\pi_{A_{\varphi_1}}(t)} \wedge I_{\varphi_2(D),\pi_{A_{\varphi_2}}(t)}$. The number ℓ of leaves of φ is the sum of the number of leaves of φ_1 and φ_2, hence the rank of this conjunction is at most the number of leaves of φ.

To prove Statement 2, we build the function $I_{\varphi(D),t}$ using disjunctions and conjunctions as in the previous paragraph, but the functions $I_{\varphi_0(D),t'}$ are not decomposed any further. This way, $I_{\varphi(D),t}$ is expressed as a monotone function of the random variables and of the functions $I_{\varphi_0(D),t'}$ ($t' \in \mathrm{tup}(\varphi_0)$). Thus $I_{\varphi(D),t}$ can be written in the required form and it is clear that $J_{t'}$ does not depend on the structure of the subplan φ_0. □

Lemma 26. *Let f, g be two monotone 0-1 functions of rank at most ℓ on a set V of independent 0-1 random variables.*

(1) If the probability of 1 for each random variable is decreased by at most a factor $c > 1$, then the expected value of f is decreased by at most a factor c^{ℓ}.

(2) If $\Pr(g = 1 \mid f^{(i)} = 1) \geq p$ for every minterm $f^{(i)}$ of f, then $\Pr(g = 1 \mid f = 1) \geq p \cdot 2^{-2\ell}$.

Proof. To prove Statement 1, let $x_1 : V \to \{0,1\}$ be an assignment of the variables chosen according to the probabilities, and let x_2 be another assignment, where $x_2(v) = 1$ with probability $1/c$ independently. If $f(x_1) = 1$, then $f^{(i)}(x_1) = 1$ for at least one minterm $f^{(i)}$ of f, and with probability at least $c^{-\ell}$, $f^{(i)}(x_2) = 1$. Thus $\Pr(f(x_1 \wedge x_2) = 1) \geq c^{-\ell} \Pr(f(x_1) = 1)$, where $x_1 \wedge x_2$ is the random assignment defined as the conjunction of x_1 and x_2. Observe that for each variable $v \in V$, $\Pr(x_1(v) = 1) = c \cdot \Pr((x_1 \wedge x_2)(v) = 1)$, and the claim follows.

For Statement 2, let x_1, x_2 be two independent random assignments, chosen according to the probabilities of the random variables. To bound the conditional probability $\Pr(g(x_1) = 1 \mid f(x_1) = 1)$, we need to bound the probability $\Pr(g(x_1) = 1 \wedge f(x_1) = 1)$. To bound this probability, we calculate $\Pr(g(x_1 \vee x_2) = 1 \wedge f(x_1 \vee x_2) = 1)$ (where $x_1 \vee x_2$ is the assignment defined as the disjunction of x_1 and x_2). Observe that for each variable v, $\Pr((x_1 \vee x_2)(v) = 1) \leq 2\Pr(x_1(v) = 1)$. Thus by Statement 1, $\Pr(g(x_1 \vee x_2) = 1 \wedge f(x_1 \vee x_2) = 1) \leq$

$2^{2\ell}\Pr(g(x_1)=1 \wedge f(x_1)=1)$ (here we used that the rank of the conjunction of two rank ℓ functions is at most 2ℓ).

If $f(x_1)=1$, then $f^{(i)}(x_1)=1$ for at least one minterm $f^{(i)}$ of f. By assumption, with probability at least p, assignment x_2 sets the variables such that $g(x_1 \vee x_2)=1$ if $f^{(i)}(x_1)=1$. Thus $\Pr(g(x_1 \vee x_2)=1 \wedge f(x_1 \vee x_2)=1) \geq p\Pr(f(x_1)=1)$ and we have

$$\Pr(g(x_1)=1 \mid f(x_1)=1) = \frac{\Pr(g(x_1)=1 \wedge f(x_1)=1)}{\Pr(f(x_1)=1)} \geq \frac{2^{-2\ell}\Pr(g(x_1 \vee x_2)=1 \wedge f(x_1 \vee x_2)=1)}{\Pr(f(x_1)=1)}$$
$$\geq \frac{2^{-2\ell}p\Pr(f(x_1)=1)}{\Pr(f(x_1)=1)} \geq 2^{-2\ell}p.$$

□

Proof (of Theorem 6). First, we can assume that φ is of the form $(((\varphi^* \bowtie R_1) \bowtie R_2) \bowtie \ldots \bowtie R_{|\sigma|})$, where R_1, $\ldots$, $R_{|\sigma|}$ is an ordering of the relations in σ: if φ^* is a join-project plan for the query Q, then joining any relation R_i with φ^* does not change φ^*. However, this assumption will ensure that if we make any changes in φ^*, then φ will remain a join-project plan for the query Q. Furthermore, we assume that for every variable $a \in A$, there is a dummy unary relation R_a with $p_{R_a}(N)=1$ and hence $w_{R_a}(N)=0$.

The two steps described below reduce the number of projections in φ in such a way that the maximum expected size of a subplan is at most a constant factor larger in the new plan than in φ (with a constant depending only on φ). This procedure is repeated as many times as the number of projections, thus the total increase of the maximum expected size is only a constant c_φ.

Let $\pi_X(\varphi_0)$ be a subplan of $\varphi(D)$ such that φ_0 does not contain any projections. Let $S \subseteq \sigma$ be the relation names appearing in φ_0, which means that $\varphi_0(D) = \bowtie_{R \in S} R(D)$. Let $X^* = C_S(X)$.

Step 1. If a relation $R \in S \setminus S[X^*]$ appears in φ_0, then R is removed from φ_0. Removing a relation from φ_0 means deleting the leaf corresponding to the relation and replacing its parent with the sibling of the deleted leaf. Let φ_0' be the join-project plan obtained from φ_0 this way. We can assume that for every $a \in X^*$, the unary relation R_a appears in φ_0 (such relations can be joined with φ_0 without increasing the number of tuples). This ensures that $X^* \subseteq A_{\varphi_0'}$. Replacing subplan φ_0 of φ with φ_0' gives a new join-project plan φ'. By our initial assumption on the structure of φ, φ' is also a join-project plan for Q.

For every subplan ψ' of φ', there is a corresponding subplan ψ of φ. We show that

$$\mathrm{E}(|\psi'(D)|) \leq 2^{2\ell+1}\,\mathrm{E}(|\psi(D)|),$$

where ℓ is the number of leaves in ψ. If subplan ψ' is disjoint from φ_0', then ψ' and ψ are the same, and we are done. Thus we have to consider only two cases: ψ' is either completely contained in φ_0', or ψ' contains φ_0'.

Case 1: ψ' is contained in φ_0'. Let Y be the set of all attributes of the relations appearing in ψ'. If $Y \subseteq X^*$, then the attributes of each relation appearing in ψ' are fully contained in X^* and we are done: $\psi' = \psi$. Otherwise, let w (resp., w') be the total weight of the relations appearing in ψ (resp., ψ'). Observe that $w - w' \leq |Y \setminus X^*|(\log N - n - 1)$: otherwise we would have $f_S(X^* \cup Y) < f_S(X^*)$, contradicting the minimality of X^*. Thus the expected size of $|\psi'(D)|$ is

$$\mathrm{E}(|\psi'(D)|) = 2^{|Y \cap X^*|\log N - w'} \leq 2^{|Y \cap X^*|\log N - w' + |Y \setminus X^*|(\log N - n - 1) - (w - w')} \leq 2^{|Y|\log N - w} = \mathrm{E}(|\psi(D)|).$$

Case 2: ψ' is not contained in φ_0', which implies that ψ' contains $\pi_X(\varphi_0')$ as subplan. Let ℓ be the number of leaves of ψ (which is an upper bound on the number of leaves of ψ'). We claim that for every tuple $t \in \mathrm{tup}(\psi')$, $\Pr(t \in \psi'(D)) \leq 2^{2\ell+1}\Pr(t \in \psi(D))$, which implies $\mathrm{E}(|\psi'(D)|) \leq 2^{2\ell+1}\,\mathrm{E}(|\psi(D)|)$. To prove this, we show that for every minterm $I^{(i)}_{\psi'(D),t}$ of $I_{\psi'(D),t}$, we have $\Pr(I_{\psi(D),t}=1 \mid I^{(i)}_{\psi'(D),t}=1) \geq 1/2$, thus by Lemma 26(2), $\Pr(I_{\psi(D),t}=1 \mid I_{\psi'(D),t}=1) \geq 1/2^{2\ell+1}$, what we need.

By Lemma 25(2), $I_{\psi'(D),t}$ can be written as

$$I_{\psi'(D),t} = \bigvee_{t' \in \mathrm{tup}(\varphi_0')} (I_{\pi_X(\varphi_0'(D)),t'} \wedge J_{t'}). \tag{4.1}$$

Consider a particular minterm $I^{(i)}_{\psi'(D),t}$ for some i, which is the product of a subset V_i of the random variables. If $\prod V_i = 1$, then $I^{(i)}_{\psi'(D),t} = 1$, implying that $I_{\pi_X(\varphi_0'(D)),t'} \wedge J_{t'} = 1$ for some tuple $t' \in \mathrm{tup}(\pi_X(\varphi_0'))$, which further implies that there is a tuple $t'' \in \mathrm{tup}(\varphi_0')$ such that $\pi_X(t'') = t'$ and $I_{\varphi_0'(D),t''} = 1$. Since φ_0' does not contain projections, $I_{\varphi_0'(D),t''}$ is the product of a set V_i' of random variables. As $I^{(i)}_{\psi'(D),t} = 1$ implies $I_{\varphi_0'(D),t''} = 1$, we have $V_i' \subseteq V_i$. By a consequence of the FKG Inequality (see Lemma 23),

$$\Pr(I_{\pi_X(\varphi_0(D)),t'} = 1 \mid \prod V_i = 1) \geq \Pr(I_{\pi_X(\varphi_0(D)),t'} = 1 \mid \prod V_i' = 1) \tag{4.2}$$

(since $V_i' \subseteq V_i$ and $I_{\pi_X(\varphi_0(D)),t'}$ is monotone). Note that φ_0 and not φ_0' appears in (4.2). By Lemma 25(2), $I_{\psi(D),t}$ can be also written in the form (4.1), hence $I_{\pi_X(\varphi_0(D)),t'} \wedge J_{t'} = 1$ implies $I_{\psi(D),t} = 1$. Thus we have

$$\begin{aligned}
\Pr(I_{\psi(D),t} = 1 \mid \prod V_i = 1) &\geq \Pr(I_{\pi_X(\varphi_0(D)),t'} \wedge J_{t'} = 1 \mid \prod V_i = 1) \\
&= \Pr(I_{\pi_X(\varphi_0(D)),t'} = 1 \mid \prod V_i = 1) \\
&\geq \Pr(I_{\pi_X(\varphi_0(D)),t'} = 1 \mid \prod V_i' = 1) \\
&\geq \Pr(I_{\pi_{X^*}(\varphi_0(D)),t''} = 1 \mid \prod V_i' = 1) \\
&= \Pr(t'' \in \pi_{X^*}(\varphi_0(D)) \mid t'' \in \varphi_0'(D)) \\
&\geq 1/2.
\end{aligned}$$

what we had to show. (The first inequality follows from the fact that $I_{\pi_X(\varphi_0(D)),t'} \wedge J_{t'} = 1$ implies $I_{\psi(D),t} = 1$; the equality after that from the fact that $\prod V_i = 1$ implies $J_{t'} = 1$; the second inequality follows from (4.2); the third inequality follows from $\pi_X(t'') = t'$; the last inequality follows from Lemma 8.)

Step 2. In the second step of the procedure, we obtain a join-project plan φ'' from φ' by replacing the subplan $\pi_X(\varphi_0')$ with $\pi_{X^*}(\varphi_0')$. (This makes the projection redundant and can be eliminated, but it is more convenient to describe the step this way.) Note that this change does not have any effect on the size of $\psi(D)$ for any subplan ψ of φ_0'. Furthermore, if ψ is a subplan containing $\pi_X(\varphi_0')$ such that there is a projection node between the nodes corresponding to ψ and $\pi_X(\varphi_0')$, then it is easy to see that the change cannot increase $\psi(D)$. Thus the only situation that we have to verify is that if ψ' is a subplan of φ' containing $\pi_X(\varphi_0')$ and having no such projection node. We show that in this case $\mathrm{E}(|\psi''(D)|) \leq 2^{n(n+2)+2\ell}\mathrm{E}(|\psi'(D)|)$, where ψ'' is the subplan of φ'' corresponding to ψ'.

Note that $X^* \subseteq A_{\psi''}$, since ψ'' contains no projections above π_{X^*}. Let Z be $A_{\psi'} \setminus (X^* \setminus X)$. If a tuple t is in $\psi''(D)$, then clearly $t_1 := \pi_{X^*}(t)$ is in $\varphi_0'(D)$ and $t_2 := \pi_Z(t)$ is in $\pi_Z(\psi'(D))$. Thus

$$\begin{aligned}
&\Pr(t \in \psi''(D)) \\
&\leq \Pr(t_1 \in \varphi_0'(D) \wedge t_2 \in \pi_Z(\psi'(D)), t_2) \\
&= \Pr(\pi_X(t_1) \in \pi_X(\varphi_0'(D))) \cdot \Pr(t_1 \in \varphi_0'(D) \mid \pi_X(t_1) \in \pi_X(\varphi_0'(D))) \cdot \Pr(t_2 \in \pi_Z(\psi'(D)) \mid t_1 \in \varphi_0'(D)).
\end{aligned} \tag{4.3}$$

By Lemma 9(2), the second factor in (4.3) is at most $2^{n(n+2)} \cdot N^{-|X^* \setminus X|}$. To bound the third factor, we use that for every i there is a tuple $t_1' = t_1'(i)$ such that

$$I^{(i)}_{\pi_X(\varphi_0'(D)),\pi_X(t_1)} = I_{\varphi_0'(D),t_1'}.$$

Observe that if $t_1' \in \operatorname{tup}(\varphi_0')$ with $\pi_X(t_1) = \pi_X(t_1')$, then

$$\Pr(I_{\pi_Z(\psi'(D)),t_2} = 1 \mid I_{\varphi_0'(D),t_1'} = 1) = \Pr(I_{\pi_Z(\psi'(D)),t_2} = 1 \mid I_{\varphi_0'(D),t_1} = 1).$$

Hence for every i it holds that

$$\begin{aligned}\Pr(I_{\pi_Z(\psi'(D)),t_2} = 1 \mid I^{(i)}_{\pi_X(\varphi_0'(D)),\pi_X(t_1)} = 1) &= \Pr(I_{\pi_Z(\psi'(D)),t_2} = 1 \mid I_{\varphi_0'(D),t_1'} = 1)\\ &= \Pr(I_{\pi_Z(\psi'(D)),t_2} = 1 \mid I_{\varphi_0'(D),t_1} = 1)\end{aligned}$$

By Lemma 26(2), this means that

$$\Pr(I_{\pi_Z(\psi'(D)),t_2} = 1 \mid I_{\varphi_0'(D),t_1}) \leq 2^{2\ell} \Pr(I_{\pi_Z(\psi'(D)),t_2} = 1 \mid I_{\pi_X(\varphi_0'(D)),\pi_X(t_1)} = 1).$$

Therefore, continuing (4.3), we have

$$\begin{aligned}\Pr(t \in \psi''(D)) &\leq \Pr(\pi_X(t_1) \in \pi_X(\varphi_0'(D))) \cdot 2^{n(n+2)} N^{-|X^*\setminus X|} \cdot 2^{2\ell} \Pr(t_2 \in \pi_Z(\psi'(D)); |; \pi_X(t_1) \in \pi_X(\varphi_0'(D)))\\ &\leq 2^{n(n+2)+2\ell} N^{-|X^*\setminus Y|} \cdot \Pr(t_2 \in \pi_Z(\psi'(D)))\\ &= 2^{n(n+2)+2\ell} N^{-|X^*\setminus Y|} \cdot \mathrm{E}(|\pi_Z(\psi'(D))|) N^{-|Z|}.\end{aligned}$$

Thus the expected size of $\psi''(D)$ is

$$\begin{aligned}\mathrm{E}(|\psi''(D)|) &= N^{|A_{\psi''}|} \cdot 2^{n(n+2)+\ell} N^{-|X^*\setminus Y|} \cdot \mathrm{E}(|\pi_Z(\psi'(D))|) N^{-|Z|}\\ &= 2^{n(n+2)+\ell}\, \mathrm{E}(|\pi_Z(\psi'(D))|)\\ &\leq 2^{n(n+2)+\ell}\, \mathrm{E}(|\psi'(D)|).\end{aligned}$$

□

E Proofs omitted from Section 5

Proof of Theorem 10. Statement (1) is an immediate consequence of Lemma 2. To prove (2), let A be the attributes of Q, let A_R be the attributes of $R \in \sigma$. The LP-dual of $L_Q(N_R : R \in \sigma)$ is the following linear program $D_Q(N_R : R \in \sigma)$:

$$\begin{array}{lll} \text{maximise} & \sum_a y_a & \\ \text{subject to} & \sum_{a \in R_A} y_a \leq \log N_R & \text{for all } R \in \sigma,\\ & y_a \geq 0 & \text{for all } a \in A. \end{array}$$

Let $(y_a : a \in A)$ be an optimal solution for the dual. Then $\sum_{a\in A} y_a = \sum_{R\in\sigma} x_R \cdot \log N_R$.

For all $a \in A$, let $y_a' = \log \lfloor 2^{y_a} \rfloor$. We set

$$R' := \left\{ t \in \operatorname{tup}(A_R) : t(a) \in \left[2^{y_a'}\right] \text{ for all } a \in A_R \right\}.$$

Then

$$|R'| = \prod_{a\in A_R} 2^{y_a'} = \prod_{a \in A_R} \lfloor 2^{y_a} \rfloor \leq 2^{\sum_{a\in A_R} y_a} \leq 2^{\log N_R} = N_R.$$

We arbitrarily add tuples to R' to obtain a relation $R(D)$ of size N_R. In the resulting instance D, we have

$$|Q(D)| \geq \prod_{a\in A} 2^{y_a'} \geq \prod_{a\in A} \frac{2^{y_a}}{2} = 2^{-n} \cdot 2^{\sum_{a\in A} y_a} = 2^{-n} \cdot 2^{\sum_{R\in\sigma} x_R \cdot \log N_R} = 2^{-n} \cdot \prod_{R\in\sigma} N_R^{x_R}.$$

□

Lemma 27. *Let Q be a join query with schema σ and let $N_R := 2$ for all $R \in \sigma$. Let G be the primal graph of Q and let $\alpha(G)$ be the size of the maximum independent set in G. The maximum of $|Q(D)|$, taken over database instances satisfying $|R(D)| = N_R$ for every $R \in \sigma$, is exactly $2^{\alpha(G)}$.*

Proof. Let A_R be the attributes of $R \in \sigma$, and let A be the attributes of Q. First we give a database D with $|Q(D)| \geq 2^{\alpha(G)}$. Let $I \subseteq A$ be an independent set of size $\alpha(G)$. Since I is independent, $|A_R \cap I|$ is either 0 or 1 for every $R \in \sigma$. If $|A_R \cap I| = 0$, then we define $R(D)$ to contain a tuple that is 0 on every attribute. If $A_R \cap I = \{a\}$, then we define $R(D)$ to contain a tuple that is 0 on every attribute and a tuple that is 1 on a and 0 on every attribute in $A_R \setminus \{a\}$. We claim that

$$Q(D) = \{t \in \mathrm{tup}(A) : t(a) \in \{0,1\} \text{ for all } a \in I, t(a) = 0 \text{ for all } a \in A \setminus I\}.$$

Clearly, the value of an attribute in I is either 0 or 1, and every attribute in $A \setminus I$ is forced to 0. Furthermore, any combination of 0 and 1 on the attributes of I is allowed as long as all the other attributes are 0. Thus $|Q(D)| = 2^{\alpha(G)}$. Note that a relation R with $|A_R \cap I| = 0$ contains only one tuple in the definition above. To satisfy the requirement $|R(D)| = N_R = 2$, we can add an arbitrary tuple to each such relation R; this cannot decrease $|Q(D)|$.

Next we show that if $|R(D)| = 2$ for every relation $R \in \sigma$, then $|Q(D)| \leq 2^{\alpha(G)}$. Since $|R(D)| = 2$ for every relation, every attribute in A can have at most two values in $Q(D)$; without loss of generality it can be assumed that $Q(D) \subseteq \{0,1\}^{|A|}$. Furthermore, it can be assumed (by a mapping of the domain of the attributes) that the all-0 tuple is in $Q(D)$.

Let S be the set of those attributes that have two values in $Q(D)$, i.e.,

$$S = \{a \in A : |\pi_{\{a\}}(Q(D))| = 2\}.$$

For every $a \in S$, let S_a be the set of those attributes that are the same as a in every tuple of $Q(D)$, i.e.,

$$S_a = \{b \in S : t(a) = t(b) \text{ for every } t \in Q(D)\}.$$

We define a sequence a_1, a_2, $\dots$ of attributes by letting a_i be an arbitrary attribute in $S \setminus \bigcup_{j<i} S_{a_j}$. Let a_t be the last element in this sequence, which means that $\bigcup_{i=1}^{t} S_{a_i} = S$. We claim that $a_1, \dots, a_t$ are independent in G, implying $t \leq \alpha(G)$. Assume that a_i and a_j ($i < j$) are adjacent in G; this means that there is an $R \in \sigma$ with $a_i, a_j \in A_R$. By assumption, the all-0 tuple is in $R(D)$. As $a_i, a_j \in S$, there has to be a $t_1 \in R(D)$ with $t_1(a_i) = 1$ and a $t_2 \in R(D)$ with $t_2(a_j) = 1$. Since $|R(D)| = 2$ and the all-0 tuple is in $R(D)$, we have $t_1 = t_2$. But this means that $t(a_i) = t(a_j)$ for both tuples in $R(D)$, implying $a_j \in S_{a_i}$. However, this contradicts the way the sequence was defined.

Now it is easy to see that $|Q(D)| \leq 2^t \leq 2^{\alpha(G)}$: by setting the value of $a_1, \dots, a_t$, the value of every attribute in S is uniquely determined and the attributes in $A \setminus S$ are the same in every tuple of $Q(D)$. ☐

We get our inapproximability result by reduction from the following result by Hastad:

Theorem 28 ([10]). *If for some $\varepsilon_0 > 0$ there is a polynomial-time algorithm that, given an n-vertex graph G, can distinguish between the cases $\alpha(G) \leq n^{\varepsilon_0}$ and $\alpha(G) \geq n^{1-\varepsilon_0}$, then* NP = ZPP.

Proof of Theorem 12. We show that if such M_L and M_U could be determined in polynomial time, then we would be able to distinguish between the two cases of Theorem 28. Given an n-vertex graph $G = (V, E)$, we construct a query Q with attributes V and schema $\sigma = E$. For each edge $uv \in E$, there is a relation R_{uv} with attributes $\{u, v\}$. We set $N_R = 2$ for every relation $R \in \sigma$. Observe that the primal graph of Q is G. Thus by Lemma 27, $M = 2^{\alpha(G)}$.

Set $\varepsilon_0 := \varepsilon/2$. In case (1) of Theorem 28, $\alpha(G) \leq n^{\varepsilon_0}$, hence $M_L \leq M \leq 2^{n^{\varepsilon_0}}$ and

$$M_U \leq M_L 2^{n^{1-\varepsilon}} \leq 2^{n^{\varepsilon_0}+n^{1-\varepsilon}} < 2^{n^{1-\varepsilon_0}}$$

(if n is sufficiently large). On the other hand, in case (2) we have $\alpha(G) \geq M \geq n^{1-\varepsilon_0}$, which implies $M_U \geq 2^{n^{1-\varepsilon_0}}$. Thus we can distinguish between the two cases by comparing M_U with $2^{n^{1-\varepsilon_0}}$. $\square$