Synopses for Query Optimization: A Space-Complexity Perspective

RAGHAV KAUSHIK
Microsoft Research
JEFFREY F. NAUGHTON and RAGHU RAMAKRISHNAN
University of Wisconsin-Madison
and
VENKATESAN T. CHAKRAVARTHY
IBM India Research Lab

Database systems use precomputed synopses of data to estimate the cost of alternative plans during query optimization. A number of alternative synopsis structures have been proposed, but histograms are by far the most commonly used. While histograms have proved to be very effective in (cost estimation for) single-table selections, queries with joins have long been seen as a challenge; under a model where histograms are maintained for individual tables, a celebrated result of Ioannidis and Christodoulakis [1991] observes that errors propagate exponentially with the number of joins in a query.

In this article, we make two main contributions. First, we study the space complexity of using synopses for query optimization from a novel information-theoretic perspective. In particular, we offer evidence in support of histograms for single-table selections, including an analysis over data distributions known to be common in practice, and illustrate their limitations for join queries. Second, for a broad class of common queries involving joins (specifically, all queries involving only key-foreign key joins) we show that the strategy of storing a small precomputed sample of the database yields probabilistic guarantees that are almost space-optimal, which is an important property if these samples are to be used as database statistics. This is the first such optimality result, to our knowledge, and suggests that precomputed samples might be an effective way to circumvent the error propagation problem for queries with key-foreign key joins. We support this result empirically through an experimental study that demonstrates the effectiveness of precomputed samples, and also shows the increasing difference in the effectiveness of samples versus multidimensional histograms as the number of joins in the query grows.

Categories and Subject Descriptors: E.4 [**Coding and Information Theory**]: *Data Compaction and Compression*; H.2.4 [**Database Management**]: Systems—*Query processing*

Authors' addresses: R. Kaushik, Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399; email: skaushi@microsoft.com; J. F. Naughton and R. Ramakrishnan, University of Wisconsin–Madison, Department of Computer Science, 1210 West Dayton Street, Madison, WI 53705; email: {naughton;raghu}@cs.wisc.edu; V. T. Chakravarthy, IBM India Research Lab, New Delhi, India 110016; email: vechakra@in.ibm.com.

General Terms: Theory, Performance

Additional Key Words and Phrases: Cardinality estimation, histograms, sampling

1. INTRODUCTION

Query optimizers use synopses of the contents of a database to decide the most efficient plan of execution (e.g., Selinger et al. [1979]), and synopsis-based cost estimation is widely recognized as one of the central challenges in query optimization. Using histograms as a synopsis method has been extensively studied [Ioannidis 2003]. Several previous efforts such as Ioannidis [1993], Ioannidis and Christodoulakis [1993], Jagadish et al. [1998], Koudas et al. [2000], and Poosala and Ioannidis [1995] have focused on constructing single and multidimensional histograms that are optimal.

On the other hand, to the best of our knowledge, the only study of the *hardness* of the problem of using synopses for cost estimation is Ioannidis and Christodoulakis [1991]. In this widely cited paper, the error in the estimate of the join result size is shown to grow exponentially as the number of joins increases. The model of estimation used for joins is that the data distribution in the join columns for each individual relation is independently approximated, say through a histogram, and that the join result is estimated by joining these approximated distributions. This model is consistent with what most database systems implement in practice. However, the conclusions of this study do not apply to synopsis techniques that follow a different estimation model. For example, techniques such as the recently proposed sketches [Alon et al. 1999, 1996] are not covered by this model since they summarize the data distribution in the join column of a relation into a single number without storing, either accurately or approximately, the frequency of any individual join element.

In this article, we study the problem of using synopses for query optimization from a space-complexity prespective. We assume an estimation model that works in two phases—a preprocessing phase that processes the database and computes synopses, and a run-time phase that, given an input Select-Project-Join (SPJ) query, uses these synopses to provide an estimate of the result size. This model covers all techniques that do not examine the data during optimization time, which includes most proposed techniques including histograms and sketches. An example of a technique *not* covered by our model is adaptive sampling [Lipton et al. 1990]. We consider the following measures of error—absolute error, defined as the (absolute value of) the difference between the correct result and the estimated result; ratio error, defined as the ratio between the estimated result and the correct result; and relative error, defined as the ratio of the absolute error to the correct result (the ratio and relative error measurements assume that the result is non-empty). In addition to these standard measures of error, we consider a fourth metric of error. Often, query optmizers choose between alternative plans of execution depending on whether the result size of a query is large or small [Chaudhuri 1998; Selinger et al. 1979]. In such cases, an accurate estimate of the result size is not as crucial as deciding whether the result size is above or below a given threshold. We define this requirement to be the threshold error requirement.

Our first main contribution is a series of results that shed light on the use of synopses for cost estimation, and in particular, the strengths and weaknesses of histograms on individual tables. We begin by showing that under our estimation model, for the class of SPJ queries, unless the synopsis essentially contains the whole database, it is impossible to guarantee low—that is, constant or polylogarithmic—error bounds for the absolute, ratio and relative error metrics (Corollary 3). This applies even for the simplest case of single-column equality selections. If we consider the threshold error requirement with the values of the threshold linear in the table size which we believe to be the most interesting case, we can address equality selections efficiently, by simply storing the high-frequency elements in the synopsis. However, for the class of range slections, we cannot provide any non-trivial guarantees unless the synopsis essentially contains the whole database (Corollary 7). This result serves to distinguish equality selection queries from range selection queries, and also to distinguish the threshold error metric from the standard metrics. The above results are information-theoretic and hold irrespective of whether the estimation process is deterministic or probabilistic.

This negative result suggests that we must be willing to tolerate looser error guarantees, which are known to be provided by histograms for single-column selections. This is in keeping with traditional wisdom that histograms are sufficient to handle single-column selections in practice. Indeed, we show that for single-column selections, histograms are almost optimal, in that for a given error requirement, they provide the best possible space complexity, even considering probabilistic alternatives. This is our second result (Section 3.2), and it differs from prior work on constructing optimal histograms such as Ioannidis and Christodoulakis [1993], Ioannidis [1993], Jagadish et al. [1998], and Koudas et al. [2000] in that we characterize the relative optimality of histograms versus all other synopsis structures covered by our estimation model.

Empirical results with histograms [Ioannidis 1993; Ioannidis and Christodoulakis 1993; Jagadish et al. 1998; Koudas et al. 2000; Poosala and Ioannidis 1995] typically yield errors that are much better than indicated by our worst case analysis. Note that our results on histograms do not assume anything about the data distribution. We then consider example data distributions that broadly exhibit the property that a few values occur with a very high frequency (Section 3.3). Such distributions are known to be very common [Christodoulakis 1984; Faloutsos and Jagadish 1992]. For these distributions, we show that we can suitably construct histograms that yield guarantees that are much better than the worst-case guarantees (Property 6 and Lemma 12).

Next, we consider queries with joins, and show (Section 4.1) that the space needed to provide a similar absolute error guarantee is higher, adding further evidence (complementing the error-propagation result of Ioannidis and Christodoulakis [1991]) that maintaining histograms on individual tables is not a promising approach to estimating the cost of join queries.

We observe that this negative result does not hold for the special case of key-foreign key joins, and this leads to our second main contribution, which is to show that precomputed samples are an effective approach to estimating the cost of queries with (arbitrarily many) key-foreign key joins. We model the special case of queries with (only) key-foreign key joins as a distributed selection over a single "fattened" table defined by precomputing all the joins. We show that by keeping a small sample of the precomputed join and estimating the query cost by running the query on the sample and scaling up results, we obtain an estimate that is provably good with high probability (Theorem 15). Our estimate has the property that when the query result size is high, there is a probabilistic bound on the ratio error, whereas when the query result size is small, there is a probabilistic bound on the absolute error. This fits nicely with the observation that ratio error matters more for larger results and absolute error matters more for smaller results. This result also translates into a looser guarantee for the threshold error metric. We also show that the amount of space consumed by this solution for a given guarantee is almost optimal (Section 4.3). Our empirical results show the effectiveness of this solution versus multidimensional histograms as the number of joins grows. We note that this strategy is an extension of *join synopses*, which have been proposed for approximate query processing [Acharya et al. 1999], to the problem of query optimization, where the space and error requirements are very different.

2. PRELIMINARIES

2.1 Data Model

A (k, t) database schema $\mathcal{D}$ consists of a fixed set of relation names $\{R_1, \ldots, R_k\}$, where relation R_j consists of a fixed ordered list of t column names $\{C_{j_1}, \ldots, C_{j_t}\}$. In this article, we fix k to be a constant.

An (N, t)-database instance I populates each relation R_j with a multiset of N_j tuples, such that the maximum among all N_j is N. Each value in a tuple is drawn from $\{1, 2, \ldots, k.N.t\} \cup \{\text{null}\}$. I is said to have (N, t)-*rowcols*. I is said to be an $l(\leq k)$-relation instance if only l relations are nonempty.

2.2 Queries

An (N, t)-query Q is a relational algebra expression involving the operations σ, π and $\bowtie$ (in other words, we only consider SPJ queries) over the (k, t) database schema and constants from $\{1, 2, \ldots, k.N.t\}$. We assume multiset semantics for these operations.

We refer to a finite class of (N, t)-queries as an (N, t)-workload. For a workload $\mathcal{Q}$ consisting only of SPJ queries, let *size* be a function that given an (N, t)-database instance I and query $Q \in \mathcal{Q}$, returns its result size measured as the number of rows returned. For instance I and integer f, the subset of $\mathcal{Q}$ with result size $< f$ is defined as the set of f-*small* queries, and its complement is defined as the set of f-*large* queries.

2.3 Error Metrics

One central goal of maintaining statistics over a database is to estimate *size* approximately. An error metric is a function *err* that takes as input a number x to be approximated, an error bound e, and returns an interval on the real line.

Statistics computed over a database typically target a specific error metric. We consider the following error metrics in this article:

—*Absolute Error*: $abserr(x, e) = (x - e, x + e)$ for $e > 0$, the interval consisting of all integers between, but not including $x - e$ and $x + e$.

—*Ratio Error*: For $e \geq 1$, $ratioerr(x, e) = (x/e, e.x)$.

—*Relative Error*: For $0 < e < 1$, $relerr(x, e) = (x(1 - e), x(1 + e))$. Note that a relative error of $e > 1$ is not interesting since we could always return 0, which attains the bound 1.

—*Threshold Error*: $thresherr(x, e) = [0, e)$ if $x < e$ and $[e, \infty)$ if $x \geq e$. This metric is motivated by the fact that often, query optimizers do not want tight guarantees on the result size, rather they only require to know whether the result size is above or below a certain threshold, typically a large threshold. A classic example of this is when given a single-table selection, the optimizer has to decide between an unclustered index seek and a table scan. This decision depends on whether the number of records returned by the selection is above or below a threshold, typically a fixed fraction of the table size.

2.4 Estimation Model

An estimator $\mathcal{E}(Q)$ for an (N, t)-workload $\mathcal{Q}$ consists of a pair of functions $<\mathcal{S}F, \mathcal{E}F>$, called respectively the *summarizing function* and the *estimator function*. For any (N, t)-instance I, $\mathcal{S}F(I)$ returns a synopsis $\mathcal{S}$. At optimization time, given $Q \in \mathcal{Q}$, $\mathcal{E}F(Q, \mathcal{S})$, returns an estimate of *size*(I, Q). $\mathcal{EF}$ is only allowed to access the summary $\mathcal{S}$, and not I itself. We do not restrict the computational power of either function.

Since the computational power of the estimator function is not restricted, we require it to be deterministic. On the other hand, we do allow the summarizing function to be randomized. A randomized summarizing function $\mathcal{SF}(I, r)$ takes an additional input r, a random string r chosen uniformly from a (finite) domain *Rand*, and produces a summary S. We do not place any restriction on the size of *Rand*, so long as it is finite. For the same I and $\mathcal{Q}$, we obtain different summaries $\mathcal{S}_r$ depending on the random string r. Without loss of generality, we assume that all summaries S_r consume the same amount of space. An estimator is said to be *deterministic* if the summarizing function is deterministic, and *randomized* if the summarizing function is randomized.

For an error metric *err* and an error bound e:

—A deterministic estimator $\mathcal{E} =<\mathcal{S}F, EF>$ is said to *succeed* for query Q over instance I if $\mathcal{E}F(Q, \mathcal{S}) \in err(size(I, Q), e)$, and is said to *fail* otherwise.

—A randomized estimator $\mathcal{E} =<\mathcal{S}F, EF>$ is said to *succeed* with probability p for query Q over instance I if a fraction of at least p of all the random strings r yield summaries $\mathcal{S}_r$ such that $\mathcal{E}F(Q, \mathcal{S}_r) \in err(size(I, Q), e)$.

Fix a workload $\mathcal{Q}$ and an estimator $\mathcal{E}$ for $\mathcal{Q}$. Let I be an instance and $\mathcal{Q}_I \subseteq \mathcal{Q}$. $\mathcal{E}$ is said to have *p-success* for I over $\mathcal{Q}_I$, if:

(1) $\mathcal{E}$ is deterministic and it succeeds on a fraction of at least p queries in $\mathcal{Q}_I$.

(2) $\mathcal{E}$ is randomized and for each query in $\mathcal{Q}_I$, it succeeds with probability at least p.

This model of estimation covers all proposed techniques for statistics estimation that do not examine the data at optimization time, such as histograms and sketches. We relate deterministic and randomized estimators through the following property, obtained by a simple averaging argument, which is useful in later sections.

Property 1. If there is a randomized estimator for a workload $\mathcal{Q}$ that has p-success over instance I, then there is a deterministic estimator for $\mathcal{Q}$ that consumes the same amount of space and also has p-success over I.

2.5 Space Complexity

Clearly, we are interested in estimators where the summary does not consume too much space. Fix an (N, t)-workload $\mathcal{Q}$.

The space consumed by an estimator $\mathcal{E}(Q)$, *Space*($\mathcal{E}$), is defined to be the maximum space consumed by the synopsis $\mathcal{S}$ among all (N, t)-instances. We call an estimator $\mathcal{E}$ *s-bounded* if *Space*($\mathcal{E}$) $\leq s$.

For error metric *err*, error bound e, real number p, $0 < p < 1$ and positive integer f, we define $\text{LARGESPACE}_{err}(\mathcal{Q}, e, p, f)$ to be smallest s such that there is an estimator $\mathcal{E}$ for $\mathcal{Q}$: (1) with *Space*($\mathcal{E}$) $= s$, and (2) which for each (N, t)-instance, has p success over the subset of f-large queries; here success is defined with respect to *err* and error bound e. We define $\text{SPACE}_{err}(\mathcal{Q}, e, p)$ to be $\text{LARGESPACE}_{err}(\mathcal{Q}, e, p, 1)$

2.6 Relationship between the Error Metrics

Finally, before moving on to the rest of the article, we relate our error metrics through the following property.

Property 2. Assume we have fixed a workload and an instance.

(1) If an estimator succeeds on a nonempty query Q with respect to the relative error metric, where the error bound is $1 - 1/e$ ($e \geq 1$), then it also succeeds with respect to the ratio error metric with error bound e.

(2) Suppose estimator $\mathcal{E}$ succeeds on a nonempty query Q with respect to the absolute error metric with error bound $e - 1 (e \geq 1)$. Then, the estimator $\mathcal{E}'$ which behaves exactly like $\mathcal{E}$, except that it returns an estimate of 1 when $\mathcal{E}$ returns 0, succeeds with respect to the ratio error metric with bound e.

3. SINGLE-COLUMN SELECTIONS

We begin with a study of the simplest case of single-column selections, which is deemed largely solved in practice. Nonetheless, it is of interest because any lower bounds that we obtain for this case carry over trivially to more complex queries. Further, since we are interested in space complexity of estimators, the results we show in this section for histograms complete the picture.

We first show that unless we essentially store the whole database, it is impossible to even probabilistically guarantee small ratio errors for single-column selections. If we allow queries that are empty, then this result becomes trivial since any strategy that yields *any* bound on the ratio error must return 0 if the correct result size is 0, and hence can be used to identify the values present in the database. Hence, in order to make our study meaningful, for the rest of the article, we only consider queries where the result is not empty. In particular, we assume that the database is not empty.

3.1 Lower Bounds

We first prove the following general theorems that lead to the results we show in this section.

Ratio and Relative Errors

THEOREM 1. *Fix a real number $c \geq 1$. Pick an error metric between:*

(1) *the ratio error metric with bound c,*

(2) *the relative error metric with bound $1 - 1/c$*

Fix positive integers t, N, $f \leq \frac{N}{c^8}$ and $s < t.\lfloor \frac{N}{c^8.f} \rfloor.(\log_2 \sqrt{5} - 1) - 1$. Consider the (N, t)-workload $\mathcal{Q}$ of single-column equality selection queries. Fix an estimator $\mathcal{E}$ for $\mathcal{Q}$, such that $Space(\mathcal{E}) \leq s$. Then, there exists a family of single-relation (N, t)-instances such that for a majority of these instances, $\mathcal{E}$ has less than $1/2$ success over the subset of f-large queries.

PROOF. By Properties 1 and 2, it is sufficient to show this result for deterministic estimators and the ratio error metric.

Let $n = t.\lfloor \frac{N}{c^8.f} \rfloor$. Consider the (k, t)-database schema. Let $\mathcal{I}(N, t)$ be the family of instances obtained by placing each $i \in \{1, 2, \ldots, n\}$ in the $(i \bmod t)$th column of R_1 and setting its frequency to be one of $\{f, f.c^2, f.c^4, f.c^6, f.c^8\}$. In order to set the number of rows to N, we add nulls that fill up the relation appropriately. All other relations in the schema are empty. The number of instances in $\mathcal{I}(N, t)$ is 5^n.

The subset of f-large queries of $\mathcal{Q}$ includes all queries of the form $\sigma_{C_{1_a}=i}(R_1), i \in \{1, 2, \ldots, n\}$, where $a = i \bmod t$.

Every synopsis $\mathcal{S}$ produced by $\mathcal{E}$ yields a unique instance I' obtained by successively finding an estimate of the frequency of each element in $\{1, 2, \ldots, n\}$, by issuing appropriate queries from $\mathcal{Q}$. Given a member $I \in \mathcal{I}(N, t)$, we can talk about the number of $i \in \{1, \ldots, n\}$, where the frequency in I' is within a factor of c of the frequency in I (call it $close(I, I')$). The number of $I \in \mathcal{I}(N, t)$ such that $close(I, I') \geq \lceil \frac{n}{2} \rceil$ is at most $\binom{n}{\lceil \frac{n}{2} \rceil} \times 5^{\lfloor \frac{n}{2} \rfloor}$. Now, we have that:

$$\begin{aligned} 2^s \binom{n}{\lceil \frac{n}{2} \rceil} 5^{\lfloor \frac{n}{2} \rfloor} &\leq 2^s 2^n 5^{\frac{n}{2}} \\ &< \frac{1}{2}(5^n) \\ \text{since } 2^s &< \frac{1}{2}(\frac{\sqrt{5}}{2})^n \end{aligned}$$

Hence, at least half the relation instances in $\mathcal{I}(N, t)$ differ from *each* possible instance I' that the estimator function could output, on at least half the values by a factor of $\geq c$. Thus, the estimator must fail on at least $\lceil \frac{n}{2} \rceil$ values on each of these instances. □

Absolute Error

An absolute error requirement is stronger than a ratio error requirement. As the following theorem shows, the lower bounds for absolute error are stronger.

THEOREM 2. *Fix a real number $c \geq 1$. Fix the error metric to be the absolute error metric with bound c. Fix positive integers t, N, $f \leq \frac{N}{8.c}$, and $s < t.\lfloor \frac{N}{8.c.f} \rfloor.(\log_2 \sqrt{5} - 1) - 1$. Consider the (N, t)-workload, $\mathcal{Q}$ of single-column equality selection queries. Fix an estimator $\mathcal{E}$ for $\mathcal{Q}$, such that Space($\mathcal{E}$) $\leq s$. Then, there exists a family of single-relation (N, t)-instances such that for a majority of these instances, $\mathcal{E}$ has less than $1/2$ success over the subset of f-large queries.*

PROOF. The proof mimics the argument of Theorem 1 for absolute errors. The only difference is that the frequencies are set to be one of $\{f, f + 2.c, f + 4.c, f + 6.c, f + 8.c\}$. □

COROLLARY 3. *Consider the (N, t)-workload, $\mathcal{Q}$ of single-column equality selection queries. Fix constant $c \geq 1$.*

(1) *$LARGESPACE_{ratioerr}(\mathcal{Q}, c, 1/2, f) \in \Omega(N.t/f)$. In particular, $SPACE_{ratioerr}(\mathcal{Q}, c, 1/2) \in \Omega(N.t)$.*
(2) *$LARGESPACE_{relerr}(\mathcal{Q}, 1 - 1/c, 1/2, f) \in \Omega(N.t/f)$. In particular, $SPACE_{relerr}(\mathcal{Q}, 1 - 1/c, 1/2) \in \Omega(N.t)$.*
(3) *$LARGESPACE_{abserr}(\mathcal{Q}, c, 1/2, f) \in \Omega(N.t/f)$. In particular, $SPACE_{abserr}(\mathcal{Q}, c, 1/2) \in \Omega(N.t)$.*

Suppose we are allowed errors that are "small" functions (e.g., polylogarithmic in N). We have:

COROLLARY 4. *Consider the (N, t)-workload, $\mathcal{Q}$ of single-column equality selection queries. Fix an error function $e(N) \geq 1$.*

(1) *$SPACE_{ratioerr}(\mathcal{Q}, e(N), 1/2) \in \Omega(N.t/e^8(N))$.*
(2) *$SPACE_{relerr}(\mathcal{Q}, 1 - 1/e(N), 1/2) \in \Omega(N.t/e^8(N))$.*
(3) *$SPACE_{abserr}(\mathcal{Q}, c, 1/2) \in \Omega(N.t/e(N))$.*

In particular, if $e(N) \in \mathsf{polylog}(N)$, each of these lower bounds is in $\Omega(t.N/\mathsf{polylog}(N))$. The above results show that in general, no strategy for statistics estimation, whether deterministic or randomized, that is covered by our estimation model—including histograms, sketches and wavelets—can guarantee small errors for even the simplest case of single-column selections, unless essentially the whole database is stored as a synopsis.

Threshold Error

Recall that the threshold error metric is motivated by the fact that query optimizers often do not require tight bounds on the error. Rather, it suffices to determine whether the query result size is above or below a certain threshold, typically linear in the table size.

Notice that for large values of the threshold (e in the definition), this problem can be solved efficiently for equality selection queries by simply storing the elements with high frequencies.

Property 3. Consider a relation R with (N, t)-rowcols. Let $e = N/f$. There exists an α such that in space $s = \alpha ft$, we can satisfy the threshold error requirement with threshold e for equality selection queries. We achieve this by storing the f points with highest frequency for each column.

This property serves to differentiate the threshold error metric from the other error metrics we consider. The question arises what happens with range queries. Notice that since the threshold error metric treats the estimated result as effectively Boolean (larger or smaller than a threshold), we can proceed as follows: The summarizing function can simply toss a coin and store a single bit in the synopsis based on the outcome of the coin toss. If the bit stored is 1 (correspondingly, 0), the estimator function, for any given query, always returns a value higher (correspondingly, lower) than the threshold.

Property 4. This randomized estimation scheme has 1/2-success over any workload of range queries. By Property 1, this yields a deterministic scheme that works for at least 50% of the queries in the workload.

Unfortunately, as we show next, no matter what the threshold value, it is impossible to do better without using space $\Omega(Nt)$. In order to prove this, we consider synopsis schemes for the set membership problem, which we define next.

Definition 5. Fix an integer M and let $U_M = \{1, 2, \ldots, M\}$. A set-membership scheme at length M is a pair of procedures (SG, E) such that, given any subset S of U_M, SG generates a synopsis s and given the synopsis s and an element $x \in U_M$, E outputs whether $x \in S$ or not (the output need not be correct). We say that the scheme has p-success if for any subset S of U_M, the output of the estimator is correct for at least p fraction of elements in U_M.

Using the efficient list-decodable codes from Hastad et al. [2001], we can trivially prove the following lemma, which we regard to be of independent interest.

LEMMA 5. *Fix any $\epsilon > 0$. Then there exist constants α and β, and an infinite sequence of integers $M_1, M_2, M_3, \ldots$ with the following properties.*

—*For each $i \geq 1$, any set-membership scheme at length M_i with success probability $1/2 + \epsilon$ must use synopses of size at least αM_i.*

—*For each $i \geq 1$, $M_{i+1} \leq \beta M_i$.*

We reduce the set membership problem to the statistics estimation problem constrained by the threshold error metric to obtain the following result.

THEOREM 6. *Consider the threshold error metric with threshold T. Fix $0 < \epsilon < 1/2$. Let α, β be the constants in Lemma 5. Fix positive integers t, $N > T$.*

There exists a workload $\mathcal{W}$ consisting of single-column range selection queries and a family of instances $\mathcal{I}(N, t)$, each with (N, t) rowcols, such that any s-bounded estimator for $\mathcal{W}$ with more than $(1/2 + \epsilon)$ success over this workload for each instance must satisfy: $s > \frac{\alpha}{\beta}(N - T)t$.

Proof. Let $N' = N - T$. Let M be the largest among the M_i in Lemma 5 that are $\leq N'.t$. Consider the set membership problem at length M.

We create a relation instance for each subset of the universe $U = \{1, 2, \ldots, M\}$. Given a subset $S \subseteq U$, we proceed as follows. We map $j \in U$ to the $\lceil j/t \rceil$th column. If $j \in S$, we store the element $2j$, else we store the element $2j + 1$. For each column p, there is a largest $l_p \in U$ that could get mapped to p. We add the elements $2l_p + 2, 2l_p + 4, \ldots 2l_p + 2T$. In each column, we add dummy elements with a very large value to set the number of rows to be N.

Note that the range $[2j, 2(j + T - 1) + 1]$ always has exactly T elements. And the interval $[2j + 1, 2(j + T - 1) + 1]$ has T elements if and only if $j \notin S$. For each $j \in U$, we include the query $[2j + 1, 2(j + T - 1) + 1]$ in the workload $\mathcal{W}$. Clearly, an estimation scheme with more than $(1/2 + \epsilon)$-success over this workload for each relation instances yields a set-membership scheme with success probability $(1/2 + \epsilon)$ for length M. Hence, the space consumed by the estimation scheme must be at least $\alpha M \geq \frac{\alpha}{\beta}(N - T)t$. □

Corollary 7. *Consider the (N, t)-workload $\mathcal{W}$ of single-column range selection queries in the above theorem. $SPACE_{thresholderr}(\mathcal{W}, T, 1/2 + \epsilon) \in \Omega(N.t)$*

Fix threshold T. Fix numbers $l < T < u$. Since the threshold requirement as stated is impossible to achieve unless we use space linear in the size of the database, we relax the metric by only requiring the estimator to return either of the intervals $[0, u)$ and (l, ∞) that contains the query result size. We refer to this as the (l, T, u)-threshold error requirement. As for the other error metrics, we define $\text{SPACE}_{weakthresherr}(\mathcal{Q}, (l, T, u), p)$ to be the smallest s with an estimator $\mathcal{E}$ for $\mathcal{Q}$ such that: (1) $Space(\mathcal{E}) = s$, and (2) for each (N, t)-instance, it has p success; here success is defined with respect to the (l, T, u)-threshold error metric.

Special cases of interest are when $l = T/c, u = T.c$, and when $l = T - c, u = T + c$, for some c. We have the following space lower bounds for the relaxed threshold error requirement.

Theorem 8. *Consider the $(T/c, T, T.c)$-threshold error metric. Fix $0 < \epsilon < 1/2$. Let α, β be the constants in Lemma 5. Fix positive integers t, N. There exists a workload $\mathcal{W}$ consisting of single-column range selection queries and a family of instances $\mathcal{I}(N, t)$, each with (N, t) rowcols, such that any s-bounded estimator for $\mathcal{W}$ with more than $(1/2+\epsilon)$ success over this workload for each instance must satisfy: $s > \frac{\alpha}{\beta}t\lfloor\frac{N}{Tc}\rfloor$.*

Proof. Let $N' = \lfloor\frac{N}{T.c}\rfloor$. Let M be the largest among the M_i in Lemma 5 that are $\leq N'.t$. Consider the set membership problem at length M.

We create a relation instance for each subset of the universe $U = \{1, 2, \ldots, M\}$. Given a subset $S \subseteq U$, we proceed as follows. We map $j \in U$ to the $\lceil j/t \rceil$th column. If $j \in S$, we store the element j with frequency T/c, else we store it with frequency $T.c$. In each column, we add dummy elements with a very large value to set the number of rows to be N.

Now the threshold error requirement for the range query $[j, j]$ can be satisfied only by *one* of the intervals $[0, T.c)$, $(T/c, \infty)$, depending on whether $j \in S$ or not. Consider the workload of queries $[j, j]$, $j \in U$. We can see that an estimation scheme for this workload translates to a set-membership scheme at length M with the same success. The result follows. □

COROLLARY 9. *For the (N, t)-workload, $\mathcal{W}$ in the theorem above,*

$$SPACE_{weakthresherr}(\mathcal{W}, (T/c, T, T.c), 1/2 + \epsilon) \in \Omega(t.N/(T.c)).$$

For the $(T - c, T, T + c)$-threshold error requirement, we have the following result.

THEOREM 10. *Consider the $(T - c, T, T + c)$-threshold error metric. Fix $0 < \epsilon < 1/2$. Let α, β be the constants in Lemma 5. Fix positive integers t, N. There exists a workload $\mathcal{W}$ consisting of single-column range selection queries and a family of instances $\mathcal{I}(N, t)$, each with (N, t) rowcols, such that any s-bounded estimator for $\mathcal{W}$ with more than $(1/2 + \epsilon)$ success over this workload for each instance must satisfy: $s > \frac{\alpha}{\beta} t(\lfloor \frac{N}{2c} \rfloor - \frac{T}{2c})$.*

PROOF. Let $\gamma > 1$ be such that $T/\gamma.c$ is a (positive) integer. Define $T' = T/\gamma.c$, and $N' = \lfloor N/\gamma.c \rfloor$.

We proceed with N' and T' as the N and T in the proof of Theorem 6, except the frequency of each element is set at $\gamma.c$ and use the fact that $\gamma \leq 2$. □

COROLLARY 11. *For the (N, t)-workload, $\mathcal{W}$ in the theorem above, $SPACE_{weakthresherr}(\mathcal{W}, (T - c, T, T + c), 1/2 + \epsilon) \in \Omega(t.N/c)$.*

3.2 Histograms are Almost Space-Optimal

The above lower bounds hold for the simplest case of single column selections, which is contrary to traditional wisdom that histograms are sufficient for this case. The reason is that, at least for single column selections, higher errors can be tolerated in practice.

Consider single relation instances with N rows and t columns. Let $e \colon Z \to Z$ denote an error function. It is known that an equi-depth histogram on each column with $\lceil \frac{N}{2.e(N)} \rceil$ buckets yields an absolute error of at most $e(N)$ for single-column equality and range selection queries [Piatetsky-Shapiro and Connell 1984]. For instance, an equi-depth histogram with 20 buckets yields an absolute error of at most 10% of the number of rows in the table. Notice that the same equi-depth histogram satisfies the $[T - e(N), T, T + e(N)]$ threshold error requirement.

Now, from Corollary 4, we know that *any* deterministic or randomized estimator that yields an absolute error of $e(N)$ for single-column equality selection queries must use space $\Omega(t.N/e(N))$. Similarly, from Corollary 11, any estimation scheme that satisfies the weaker $[T - e(N), T, T + e(N)]$-threshold error requirement must use space $\Omega(t.N/e(N))$. Hence, we conclude that histograms are almost space optimal for single-column selections.

3.3 Behavior of Histograms on Special-Case Distributions

Histograms are deemed to have largely solved the problem of single-column selections. The analysis above deals with the worst-case behavior of histograms, in the absence of any assumptions on the distribution of the data. Empirical results in previous work [Ioannidis 1993; Ioannidis and Christodoulakis 1993; Jagadish et al. 1998; Koudas et al. 2000; Poosala and Ioannidis 1995] show that the errors obtained are much smaller than indicated by our analysis of the previous section. The question arises whether there are specific data distributions where errors are smaller.

In this section, we consider example data distributions that broadly exhibit the property that a few values occur with a very high frequency. Such distributions are known to be very common [Christodoulakis 1984; Faloutsos and Jagadish 1992]. For such distributions, we show that we can suitably construct histograms that yield guarantees that are much better than the worst-case guarantees.

We begin with the zipfian distribution. We consider the form of the zipfian distribution given by the formula:

$$f = \frac{T}{r}$$

where f is the frequency of an element f is inversely related to its rank in the descending order of frequency, and T is of the form αN. Here, N is the number of rows in the relation. Note that if the column has D distinct values, we have:

$$1 + \frac{1}{2} + \cdots + \frac{1}{D} = \frac{1}{\alpha}$$

Since $\Sigma(1/n)$ diverges, the number of distinct values is at most a constant.

Property 5. For the zipfian distribution as defined above, storing the frequencies of the distinct values explicitly is an accurate histogram with zero error.

While real data distributions are based on generalizations of the zipfian distribution as defined above, Property 5 captures the intuition behind why end-biased histograms [Poosala and Ioannidis 1995] perform really well for the zipfian distribution.

In general, histograms perform much better than the worst-case guarantee when the number of distinct values in a table is much smaller than the number of rows, for example $O(\sqrt{N})$, when we can explicitly represent the full relation using a histogram.

We next consider the exponential distribution, where we have that the fraction of elements that are $\leq x$ is $1 - \exp^{(-\lambda x)}$. We proceed by processing the elements in descending order of their frequency, storing each explicitly till the remaining elements together constitute, say $< N/100$ (N is the number of rows). For these remaining elements, we use an equi-depth histogram $\sqrt{N}$ buckets. We recall that this is referred to as the *Compressed* histogram in prior work [Poosala et al. 1996]. This yields an absolute error of at most $\sqrt{N}/50$.

LEMMA 12. *For the exponential distribution, ths compressed histogram described above uses $\sqrt{N} + c$ buckets for some constant c and yields an absolute error of $\sqrt{N}/50$.*

This is to be contrasted with the generic equi-depth histogram described in the previous section that yields an absolute error of $\sqrt{N}$. Again, the underlying theme here is that an end-biased histogram that stores the highest frequency elements explicitly yields a much lower error, owing to the vagaries of the data distribution.

Observe that among all single-column selections, histograms first of all focus on the subclass of equality and range selection queries. For arbitrary selections, the errors yielded by histograms can be much higher than what we have analyzed in Section 3.2. In addition, the above analysis shows that for commonly occuring data distributions where only a few values occur very often, the errors yielded by histograms are even smaller than what is indicated by our worst-case analysis. This explains the empirical results obtained in previous work.

4. JOINS

We begin with a discussion of arbitrary joins, and then move on to key-foreign-key joins.

4.1 Arbitrary Joins

We now consider instances with (in general) more than one nonempty relation. Performing a 2-way self-join over a relation squares the frequencies of its elements. Based on this observation, we obtain the following results.

THEOREM 13. *Fix a real number $c \geq 1$. Pick an error metric between (1) the ratio error metric with bound c, (2) the relative error metric with bound $1 - 1/c$. Fix positive integers t, N and $s = t.\lfloor \frac{N}{c^4} \rfloor.(\log_2 \sqrt{5} - 1) - 1$. There exists an (N, t)-workload $\mathcal{Q}$ of queries with 2-way equijoins and single-column equality selection, such that any estimator $\mathcal{E}$ for $\mathcal{Q}$ which, for each (N, t)-instance, has $1/2$ success over the subset of nonempty queries in $\mathcal{Q}$, must satisfy: Space$(\mathcal{E}) \geq s$.*

PROOF. This follows by reducing the single-column equality selection problem to a problem involving equi-joins. Consider the family of relation instances used in the proof of Theorem 1, setting $f = 1$. Consider the following strategy to estimate the frequency of any element i in I. We estimate the size of the query $\sigma_{C_{1_a}=i}(R_1 \bowtie_{C_{1_a}} R_1)$ (here, $a = i \bmod t$), and use its square root as the estimate for the frequency of i. If the self-join estimator has ratio error c, then this estimate for the frequency of i has ratio error at most $\sqrt{c}$. Hence, the result follows from Theorem 1. □

We note again that an absolute error requirement is stronger than a ratio error requirement and hence, we obtain a tighter bound for absolute errors, by essentially the same proof as above.

THEOREM 14. *Fix a real number* $c \geq 1$. *Set the error metric to be the absolute error metric with bound c. Fix positive integers t, N and* $s = t.\lfloor \frac{N}{8\sqrt{c}} \rfloor.(\log_2 \sqrt{5} - 1) - 1$. *There exists an* (N, t)*-workload* $\mathcal{Q}$ *of queries with* 2*-way equijoins and single-column equality selection, such that any estimator* $\mathcal{E}$ *for* $\mathcal{Q}$ *that, for each* (N, t)*-instance, has* $1/2$ *success over the subset of nonempty queries in* $\mathcal{Q}$*, must satisfy:* $Space(\mathcal{E}) \geq s$.

Hence, in particular, if we want to estimate two-way join sizes with an absolute error of say, $\sqrt{N}$, then we need $\Omega(t.N^{3/4})$ space. In particular, building a histogram on each relation of $\sqrt{N}$ buckets is not sufficient.

We obtain a corollary analogous to Corollary 3 which shows that for all SPJ queries, no strategy for statistics estimation can guarantee small errors unless essentially the whole database is stored as a synopsis.

Since the threshold error requirement is a boolean requirement, the result of Corollary 7 is directly applicable.

4.2 Key-Foreign Key Joins

Consider the problem of estimating the result sizes of SPJ queries where we focus only on key-foreign key joins, which is the most common case in practice. For a large class of schemas, such as star schemas, an SPJ query with only key-foreign key joins is a selection query over the "fattened" fact table where all joins are precomputed. For example, if a star schema has fact table F and dimension tables $R_1, \ldots, R_l$, then if we define *FatF* to be the star join $F \bowtie R_1 \bowtie \cdots \bowtie R_l$, then a star-join query of the form $\sigma_{p_F}(F) \bowtie \sigma_{p_{i_1}}(R_{i_1}) \bowtie \cdots \bowtie \sigma_{p_{i_m}}(R_{i_m})$ is equivalent to $\sigma_{p_{i_1} \wedge \cdots \wedge p_{i_m} \wedge p_F}(FatF)$. Based on this observation, we model the problem of SPJ query estimation to be one of estimating selections over a single table. This strategy is an extension of *join synopses*, which have been proposed for approximate query processing [Acharya et al. 1999], to the problem of query optimization. We refer the reader to Acharya et al. [1999] for a detailed analysis of the class of queries where this strategy is applicable.

In our setting, we model this by restricting ourselves to selection queries over single-relation (N, t)-instances. Without loss of generality, we assume that the only nonempty relation (in the (k, t)-database schema) is R_1. Consider an estimator where the summarizing function takes a uniform sample of x rows from R_1, and the estimator function, given a selection query, simply evaluates it on the sample and scales up the result size. The procedure for a star schema is shown in Figure 1. Call this estimator *Sample*(x).

THEOREM 15. *Let* $x \in Z$ *with* $x > 0$. *Consider the* (N, t)*-workload of selection queries over* R_1. *The randomized estimator Sample*(x) *has the following property. For any* (N, t)*-instance, it succeeds with probability* $\geq 1 - (1/e^2 + 1/e^{16/3})$ (1) *for each* $(16N/x)$*-small query with respect to the absolute error metric with bound* $16N/x$*, and* (2) *for each* $(16N/x)$*-large query with respect to the ratio error metric with bound* 2.

PROOF. Since we only consider single-relation (N, t)-instances, where only R_1 is nonempty, we know that the number of rows in R_1 is N. Consider a query Q that selects r rows out of N. Then, since the sampling process is uniform, with

```
procedure Sample($Sch$, $x$)
//$Sch$ is a star schema with fact table $F$
//and dimension tables $D_1, \ldots, D_l$
//This procedure creates a sample of size $x$
begin
    Compute $FatF$ = the star join
    $F \bowtie D_1 \bowtie \ldots \bowtie D_l$
    for ($i$ from 1 to $x$)
        draw a random row from $FatF$
    These $x$ rows form the synopsis $S$
end
procedure UseSample(Sample $S$, Query $Q$)
//$Q$ is a star-join of the form
//$\sigma_{p_F}(F) \bowtie \sigma_{p_{i_1}}(D_{i_1}) \bowtie \ldots \sigma_{p_{i_m}}(D_{i_m})$
//$i_j \in \{1, \ldots, l\}$, each $p_{i_j}$ and $p_F$ is a predicate
begin
    Let $Q' = \sigma_{p_{i_1} \wedge \ldots \wedge p_{i_m} \wedge p_F}(S)$
    Run $Q'$ to obtain $r$ rows
    Return $r.N/x$, where $N$ is the number of rows in $FatF$
end
```

Fig. 1. Sampling estimator for star schema .

respect to Q, it can be viewed as a Bernoulli trial where the success probability is $\frac{r}{N}$.

Suppose running Q on the sample yields Y rows. Then, the estimate for this query is $(N/x)Y$. The expected value of Y is rx/N, that is, we expect the result size to be scaled down in the sample. Hence, we expect the estimator to return the correct value r. In order to assess how far it deviates from the expected value, we use Chernoff bounds. We consider the case when Y is above its expected value.

$$Pr[Y \geq (1+\epsilon)rx/N] \leq \exp(-(\epsilon)^2(rx/3N))$$
$$\Leftrightarrow Pr[NY/x \geq (1+\epsilon)r] \leq \exp(-(\epsilon)^2(rx/3N))$$

and

$$Pr[Y \leq (1-\epsilon)rx/N] \leq \exp(-(\epsilon)^2(rx/2N))$$
$$\Leftrightarrow Pr[NY/x \leq (1-\epsilon)r] \leq \exp(-(\epsilon)^2(rx/2N))$$

Setting $\epsilon = 16N/rx$, we get

$$\begin{aligned} Pr\left[NY/x \geq r + \frac{16N}{x}\right] &\leq \exp(-(16N/rx)^2(rx/3N)) \\ &= \exp(-256.N/(3.rx)) \\ &\leq \exp(-16/3) \text{ if } r \leq 16N/x \end{aligned}$$

Similarly, setting $\epsilon = 1$, we get:

$$\begin{aligned} Pr[NY/x \geq 2r] &\leq \exp(-rx/(3N)) \\ &\leq \exp(-16/3) \text{ if } r \geq 16N/x \end{aligned}$$

Finally, setting $\epsilon = 1/2$, we get:

$$\begin{aligned} Pr[NY/x \leq r/2] &\leq \exp(-rx/(8N)) \\ &\leq \exp(-2) \text{ if } r \geq 16N/x \end{aligned}$$

This proves the result. □

COROLLARY 16. *The randomized estimator Sample(x) also satisfies the* $(16N/x, 32N/x, 64N/x)$*-threshold error requirement with probability* $\geq 1 - (1/\exp 2 + 1/\exp(16/3))$.

What is interesting about this solution is that this guarantee holds *irrespective* of the data distribution. This is in contrast with the attribute value independence assumption made by commercial optimizers that is known to lead to large estimation errors [Bruno and Chaudhuri 2002]. Note also that the above guarantees do not assume anything about the nature of the selection predicate. Hence, this result holds for equality, range and even disjunctive selections.

By using a simple averaging argument, we can show that:

COROLLARY 17. *For any class of selection queries, the fraction of all random samples that succeed* (*in the sense of Theorem* 15) *for a fraction of at least* f, $0 < f < 1$ *of the queries is* $\geq (1 - 1/\exp 2 - 1/\exp(16/3) - f)/(1 - f)$.

Setting $f = 0.6$, we find that about 65% of all random samples succeed for at least 60% of all queries. Hence, a majority of the random samples have a high success ratio.

As discussed in the work on join synopses [Acharya et al. 1999], by suitably maintaining multiple precomputed samples, we can extend the above properties to arbitrary SPJ queries over snow-flake schema. Indeed, for any fixed join template, irrespective of whether the joins are key-foreign key, or even equijoins, as long as the join size is linear in the size of the base tables, we can easily extend the sampling strategy to handle this case.

4.3 Sampling is Almost Space-Optimal

Consider a function f: $Z \rightarrow Z$. If *FatF* has N rows and t columns, *Sample* $(\lceil 16.N/f(N) \rceil)$ has $\lceil 16.N/f(N) \rceil$ rows and t columns. In other words, the space consumed is $t.\lceil 16.N/f(N) \rceil$ "cells". The (probabilistic) guarantee yielded is that of a constant ratio error for all $f(N)$-large (single-column and multicolumn) queries in addition to an absolute error of $f(N)$ for $f(N)$-small queries. Note that by Corollary 16, the same sample satisfies the $(f(N), 2.f(N), 4.f(N))$ threshold error requirement with high probability.

Recall that Corollary 3 shows that *any* deterministic or randomized estimator that provides a constant ratio error for $f(N)$-large single-column equality selection queries must use space $\Omega(t.N/f(N))$ (measured in bits). Also, by Corollary 9, any estimation scheme that satisfies the $(f(N), 2.f(N), 4.f(N))$-threshold error requirement must use space $\Omega(t.N/f(N))$. Thus, we obtain the remarkable conclusion that sampling is almost space-optimal.

5. EXPERIMENTAL STUDY

As observed in Section 4.2, a select-project-key-foreign key join query is a select-project query over a "fattened" table corresponding to the join without any selections. Hence, in addition to the sampling approach, it is also possible to use techniques such as sketches, multidimensional histograms and wavelets. The goal of this section is to study the effectiveness of the sampling approach against the strategy of using multidimensional histograms to estimate the result size of select-project-join queries, especially as the number of joins in the query increases. We defer an empirical comparison with sketches and wavelets to future work, noting that wavelets have a limitation in that they are only applicable for numeric attributes. We only consider key-foreign key joins and defer an analysis of nonkey-foreign key joins to future work.

5.1 Analytical Comparison with Multidimensional Histograms

As shown in Muralikrishna and DeWitt [1988], in order to provide absolute error guarantees of the form N/c for some constant c, an equi-depth multidimensional histograms needs a number of buckets that is exponential in the number of columns (although several multidimensional histograms have been proposed later, to the best of our knowledge, none of these comes with better provable guarantees). On the other hand, the sampling estimator consumes space *linear* in the number of columns to yield probabilistic guarantees.

5.2 Empirical Comparison

We next focus on an empirical comparison.

5.2.1 *Modeling Joins.* Observe that in a star schema, as we increase the number of joins by joining the fact table with more dimension tables, the number of attributes over the "fattened" table over which the equivalent selection query is expressed increases. Indeed, if we assume that each table participating in the join has exactly two columns, one which is the joining column and another which contributes one column to a selection, then the number of joins is the same as the number of dimensions. Hence, we study the behavior of sampling and multidimensional histograms with increasing number of joins by generating a "fattened" table with increasing number of columns. In order to vary the number of joins, we vary the number of columns of the data set.

5.2.2 *Data.* We use the experimental setup of Bruno et al. [2001] for this purpose. The data set we use is synthetic and is based on the Gaussian distributions [William et al. 1993] which consist of a predetermined number of overlapping multidimensional Gaussian bells. The parameters for these data sets are (1) the number of Gaussian bells p, (2) the standard deviation of each bell, σ, and (3) a zipfian parameter z that regulates the total number of tuples contained in each Gaussian bell. For all the experiments, the number of data points is fixed at 500000. The default setting for parameters is $p = 100$, $\sigma = 25$, $z = 1$.

We also report results over a real database, the Census3d database [Blake and Merz 1998], a 3-dimensional projection of a fragment of the US Census Bureau data. It contains 210138 tuples.

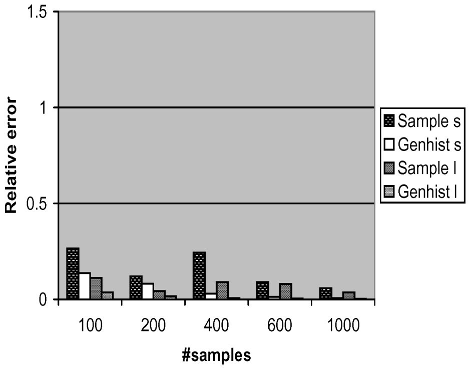

Fig. 2. No. of joins = 0.

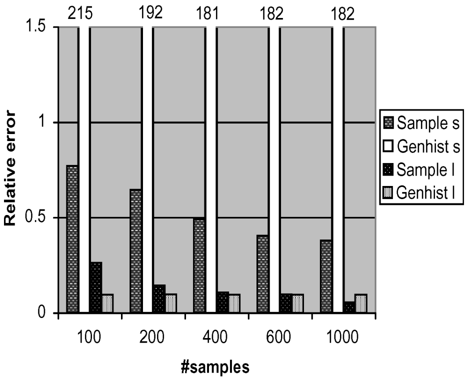

Fig. 3. No. of joins = 4.

5.2.3 *Workload.* The query workload consists of multidimensional range queries generated by creating a query center uniformly at random and expanding the query boundary to obtain a hyper-rectangle that occupies 20% of the total volume of the data domain. We classify these queries as *large* if their result size is more than 10% of the data size, and *small* otherwise. We note here that sampling is not restricted to work for this class of queries alone and that the analysis in Section 4.2 holds for arbitrary selection queries. In all our experiments, the number of queries in a workload is set to 500.

5.2.4 *Histograms.* The most recent multidimensional histograms proposed include the ST-holes histogram Bruno et al. [2001] and the GenHist histogram [Gunopulos et al. 2000]. In the results reported in [Bruno et al. 2001], it is found that the GenHist and ST-holes methods are superior to the rest and also to the solution implemented in practice based on the attribute value independence assumption. Hence, we focus on the ST-holes and GenHist histograms for a comparison against sampling. Across all of our experiments over this data set, the ST-Holes histogram which refines buckets based on a query workload performs comparably to the GenHist histogram. Hence, we report only the numbers for the GenHist histogram.

Given a sample size, we fix the number of buckets of the histograms appropriately so that the total space consumed by the data structures is the same. Our results are reported in terms of the number of samples.

5.2.5 *Varying the Number of Joins.* Figures 2, 3 and 4 show the results for 0, 4 and 8 joins (0 joins refers to a single table selection), intended to represent respectively the case of low, medium and large number of joins. Figure 5 shows the results on the Census3d database. The X-axis shows the sample sizes we use and the Y-axis shows the average relative error. In order to deal with empty queries, for the purposes of error measurement, we treat them as having a single result. For each sample size, we separately report the average relative errors for small and large queries (we use the short hand s for small and l for large). We observe the following:

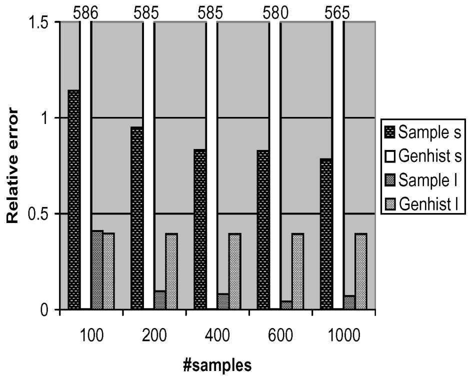

Fig. 4. No. of joins = 8.

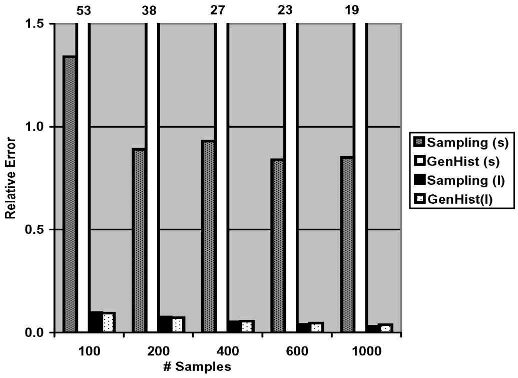

Fig. 5. Census data: 3 dimensions.

(1) For the case of 0 joins, which is a single table selection, histograms perform better than sampling, which is only to be expected. However, the errors obtained through sampling are within the bounds of what is required for query optimization. In particular, for large queries, the relative error is within 10% for all sample sizes more than 100.

(2) For small queries, sampling does significantly better. One potential reason could be that for this data set and this workload of queries, histograms over-estimate the result for small queries, yielding high relative errors. On the other hand, sampling under-estimates the result sizes for small queries yielding lower relative errors. For example, in the extreme case of an empty query, sampling always produces an empty result, whereas a histogram could produce a really high result depending on the query. We examine the absolute errors to test this hypothesis. While sampling still performs much better (factors of 2 to 4 times), the difference is not orders of magnitude. This is consistent with the above hypothesis.

(3) For large queries, sampling is always competitive with GenHist and does significantly better as the number of joins increases, especially for sample sizes of 400 and above. This is consistent with the conjecture made in Gunopulos et al. [2000] that sampling is better for higher dimensions.

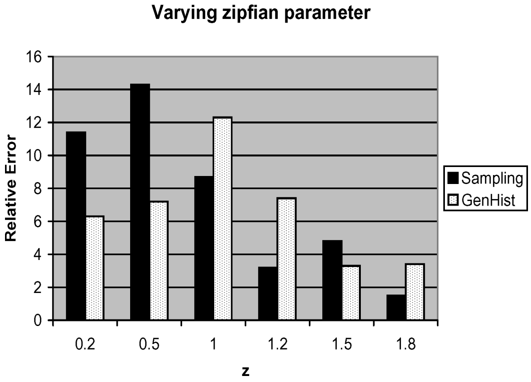

Fig. 6. Varying z.

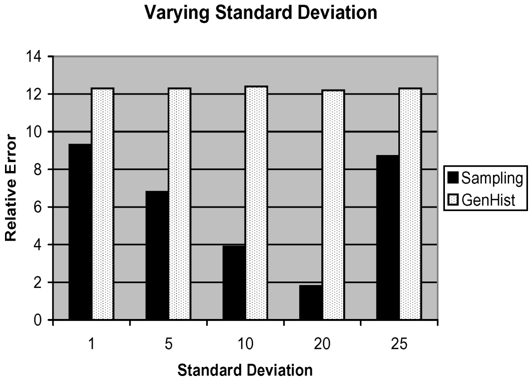

Fig. 7. Varying σ.

(4) The errors for smaller queries are consistently larger than those for larger queries confirming our analysis that in limited space, it is difficult to obtain low relative errors.

5.2.6 *Varying Data Characteristics.* We next address the comparison between GenHist and Sampling when the underlying data distribution varies, roughly from uniform to more skewed. We fix the number of joins at 4. We vary both the zipfian parameter z and the standard deviation σ. For each value of the parameter, we compare the average relative error over large queries yielded by GenHist against the one yielded by Sampling. The trends are very similar for the smaller queries and we do not report them. Figures 6 and 7 show the results of these experiments. The X-axis shows the varying parameter (respectively, z and σ). The Y-axis plots the average relative error of GenHist and Sampling. We can observe that both GenHist and Sampling tend to be more effective when the data is skewed. This is not surprising since when the data is dominated by fewer values, both these techniques can identify these dominant values and yield lower errors for larger queries. But, the figures indicate that sampling tends to be even more effective than GenHist when the skew increases. Thus,

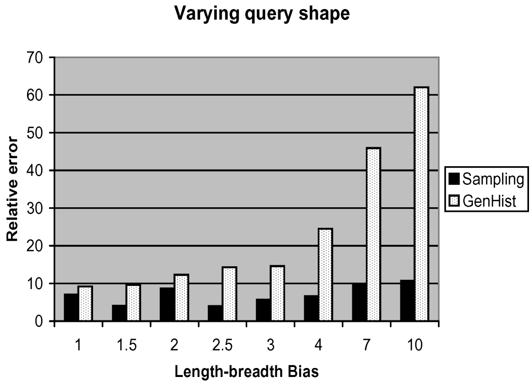

Fig. 8. Varying query shape.

for example, when z is 0.2, GenHist yields a lower error. On the other hand, when z is 1.8, Sampling yields a lower error.

5.2.7 *Varying the Query Shape.* We next study the impact of varying the query *shape*. Recall that the multidimensional queries are generated by creating a query center uniformly at random and expanding the query boundary to obtain a hyper-rectangle. Clearly, a robust statistics estimation strategy must be robust as the shape of this hyper-rectangle varies from square to "thin" rectangles where some dimensions are much wider than others. We generate our queries with a length-breadth *bias* parameter that determines the shape of the resulting query. The parameter is used as follows. A query is initially expanded to a hypercube and then half the dimensions are expanded by factor *bias*, shrinking the other half by the same factor, thus keeping the volume of the hyper-rectangle constant. Notice that when *bias* is 1, we generate hypercubes. Figure 8 plots the relative error (for larger queries) yielded by GenHist and Sampling as *bias* varies. This experiment is over the synthetic data where the number of joins is fixed at 4. We notice an interesting trend—the error yielded by the multidimensional histogram GenHist increases significantly as *bias* increases. One explanation for this is the error yielded by a histogram over a query is proportional to the number of buckets that partially intersect the query. A "thin" hyper-retangle is likely to partially intersect more buckets and hence is likely to yield a higher error. Sampling on the other hand yields a steady error, independent of the query shape. This shows that sampling is more robust as a statistics estimation strategy.

5.2.8 *Summary.* We now summarize our empirical results.

(1) Sampling out-performs multidimensional histograms when (1) the number of joins increases, (2) when the data becomes more skewed (even though the effectiveness of both tends to increase when the data is skewed).
(2) Sampling is much more robust as the *shape* of the multidimensional query changes. On the other hand, GenHist incurs significantly higher errors as the hyper-rectangle becomes less square.

(3) The GenHist histogram is constructed by making several passes over the data. Since commercial implementations typically create the histogram over a *sample* of the data [Chaudhuri et al. 1998], the difference between the two approaches in a commercial implementation is only likely to increase.

(4) Samples on the other hand are very simple to create and algorithms for their incremental maintenance have been proposed in the literature [Gibbons et al. 1997].

6. RELATED WORK

There are several sources of error in query optimization such as the statistics used, the plan space explored and the cost model that computes the effectiveness of a plan. Previous work has addressed complexity issues in both statistics [Ioannidis and Christodoulakis 1991] and plan space exploration [Chatterji et al. 2002; Ibaraki and Kameda 1984]. This article focuses on the statistics aspect.

6.1 Space Complexity

As mentioned in Section 1, the only results on hardness of gathering statistics that we are aware of is the work by Ioannidis and Christodoulakis on error propagation. In particular, we are not aware of any results on space complexity of gathering statistics.

Analyzing the complexity of gathering statistics is reminiscent of the field of communication complexity [Kushilevitz and Nisan 1997]. This area was introduced by Yao [1979], with the goal of providing a framework to analyze distributed computations. In the most widely studied two-party model, this problem deals with how many bits Alice and Bob have to exchange in order to compute a function when the input is split between them. Communication complexity is a powerful abstraction used to prove several lower bounds, including some recent lower bounds for computation on streaming data [Bar-Yossef 2002]. Our setting has fundamental differences. We are trying to compute a synopsis that can be constructed by making *multiple* passes over the data. Thus, in order to compute any specific function whose result is small (e.g., frequency moments), we are allowed to pre-compute its result as part of the synopsis. It is possible to model the synopsis we talk about as an approximation to the data distribution. However, common notions of measuring the “distance” between two data distributions such as mean square error and the L1 norm are “global”—they do not allow for the possibility of the approximation being close to the actual distribution on a large fraction of the values and being arbitrarily erroneous elsewhere. However, for query optimization, strategies where errors are low for a large fraction of queries—as opposed to *all* queries—are acceptable. We are not aware of any published work based on communication complexity that deals with such a notion of approximation.

6.2 Sampling

The problem of approximating a given data distribution has been studied in several scientific communities including numerical analysis, in the context of approximating a function in a piecewise fashion by a class of simple functions such as polynomials [Conte and de Boor 1972], and statistics, for instance, in connection with non-parametric density estimation [Gasser et al. 1985]. The effort in these areas has been focused on minimizing error without taking space constraints into account.

In the database community, approaches based on sampling such as Olken and Rotem [1986], Lipton et al. [1990], Haas and Swami [1992], and Chaudhuri et al. [1999] have been proposed to estimate the result size of queries. The main difference in our approach is the following. First of all, we *precompute* a set of samples for a given star or snow-flake schema. More importantly, in contrast with earlier work, we do not ask what is the minimum number of samples for a given error bound. Instead, we fix the sample size and analyze the guarantee it provides—in particular, the error metric we use depends on the query. We also show that the space consumed by sampling for such a guarantee is essentially optimal. The work that comes closest to our solution is the technique of computing *join synopses* for approximate query answering [Acharya et al. 1999]. Here, the authors propose storing precomputed samples of the results of relevant key-foreign key joins in a given star or snow-flake schema. They introduce an algorithm that finds the minimum *set* of samples to be maintained for a given schema so as to be able to answer all join queries. They also discuss the number of samples to be maintained for a given error bound and algorithms to update the samples as the underlying data changes. Most of the work in this paper is complementary to the sampling solution we propose. The main difference in our setting is that there is a strong space constraint, which is less stringent in the context of approximate query processing.

In addition, several techniques based on histograms [Ioannidis and Christodoulakis 1993; Ioannidis 1993; Jagadish et al. 1998; Koudas et al. 2000; Poosala and Ioannidis 1995; Poosala et al. 1996; Gunopulos et al. 2000; Bruno et al. 2001; Muralikrishna and DeWitt 1988], both one-dimensional and multidimensional, wavelets [Chakrabarti et al. 2000; Vitter and Wang 1999] and sketches [Dobra et al. 2002; Alon et al. 1999] have been proposed. We are not aware of any lower bounds on space complexity for any of these approaches. In addition, for multidimensional histograms, while there are measures of optimality such as V-optimality [Poosala and Ioannidis 1995], the only properties of the form shown for sampling in this article (in Theorem 15), that we are aware of are the ones observed in Piatetsky-Shapiro and Connell [1984]. We are not aware of any such properties of wavelets. Wavelets, in addition, have the limitation that they have been mainly studied for numeric attributes. Sketches do have probabilistic guarantees associated with them. However, even for a two-way join, the only upper bound proven on the variance of the sketch estimate is directly proportional to the self-join size of each relation and inversely proportional to the square of the query result size [Alon et al. 1999]. Hence, the variance is likely to be very high when the query result size is small, unless we store a large number of sketches.

7. CONCLUSIONS

In this article, we studied the problem of synopses for query optimization from a space-complexity perspective. Our information-theoretical analysis showed the intuitive result that obtaining synopses with very low error bounds in limited space is impossible, even if we are willing to settle for probabilistic bounds.

We then considered looser error bounds and showed that histograms are essentially optimal for single-dimension selection queries, in that any technique that offers the same error guarantee they provide requires almost the amount of space they consume. We also showed that for special classes of data distributions that are believed to be common in real-life data, the guarantees yielded by histograms are much better than what our worst-case analysis indicates. For the case of selections with joins, for the large class of key-foreign key joins, we showed that taking a small sample of the selection-free join provides an effective space-bounded synopsis. We also showed that this is essentially optimal, again in the sense that the guarantee provided by sampling requires almost the amount of space consumed by sampling. Finally, we presented experimental results that supported our theoretical results by comparing the benefits of sampling versus multidimensional histograms.

We now list potential future directions for research.

—An interesting open problem is whether there exists any data structure (in particular, some variant of multidimensional histograms) that provides absolute error guarantees for multicolumn selections in limited space.

—While the results for sampling in this paper have been proved for key-foreign key joins, an interesting question would be whether they extend to special cases of non key-foreign key joins. Even in the absence of theoretical bounds, it is likely that sampling will continue to be effective for a range of join selectivities, and it is important to be able to characterize this range.

—We were able to show that histograms are able to successfully exploit properties of data distributions believed to be common in real life to yield really small error bounds. The question arises whether along the same lines, there are classes of multidimensional data distributions where we can design synopsis structures that exploit the properties of these distributions to yield tighter guarantees than what our worst-case analysis indicates.

—In any commercial implementation of sampling, it would not in general be feasible to maintain (precomputed) samples of all possible key-foreign key templates. Strategies need to be discovered for computing statistics in the presence of partial samples. A natural question here would be if we can do any better than using independence assumptions.

—One crucial advantage of histograms over sampling is that they store the number of *distinct* values per bucket. In this article, we did not focus on distinct value estimation, which is a crucial component of statistics estimation. A natural extension of our work involves exploring the distinct value estimation problem.

ACKNOWLEDGMENTS

We thank Jin-Yi Cai, Surajit Chaudhuri and Rajasekar Krishnamurthy for useful discussions. We also thank Nicolas Bruno for sharing the code for the ST-Holes histogram and letting us use his experimental setup.

REFERENCES

Acharya, S., Gibbons, P., Poosala, V., and Ramaswamy, S. 1999. Join synopses for approximate query answering. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Alon, N., Gibbons, P., Matias, Y., and Szegedy, M. 1999. Tracking join and self-join sizes in limited storage. In *Proceedings of the 11th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'99)*. ACM, New York.

Alon, N., Matias, Y., and Szegedy, M. 1996. The space complexity of approximating the frequency moments. In *Proceedings of the Symposium on Theory of Computing*, ACM, New York.

Bar-Yossef, Z. 2002. The complexity of massive data set computations. Ph.D. thesis, Department of Computer Science, University of California-Berkeley.

Blake, C. and Merz, C. 1998. UCI repository of machine learning databases.

Bruno, N. and Chaudhuri, S. 2002. Statistics on query expressions. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Bruno, N., Chaudhuri, S., and Gravano, L. 2001. A multi-dimensional workload-aware histogram. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Chakrabarti, K., Garofalakis, M., Rastogi, R., and Shim, K. 2000. Approximate query answering using wavelets. In *Proceedings of the International Conference on Very Large Databases*.

Chatterji, S., Evani, S. S. K., Ganguly, S., and Yemmanuru, M. D. 2002. On the complexity of approximate query optimization. In *Proceedings of the 11th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'99)*. ACM, New York.

Chaudhuri, S. 1998. An overview of query optimization in relational systems. In *Proceedings of the 11th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'99)*. ACM, New York.

Chaudhuri, S., Motwani, R., and Narasayya, V. 1999. On random sampling over joins. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Chaudhuri, S., Motwani, R., and Narasayya, V. R. 1998. Random sampling for histogram construction: How much is enough? In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Christodoulakis, S. 1984. Implications of certain assumptions in databaser performance evaluation. *Trans. Datab. Syst.*

Conte, S. D. and de Boor, C. 1972. *Elementary Numerical Analysis: An Algorithmic Approach*. McGraw Hill, New York.

Dobra, A., Garofalakis, M., Gehrke, J. E., and Rastogi, R. 2002. Processing complex aggregate queries over data streams. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Faloutsos, C. and Jagadish, H. 1992. On B-tree indices for skewed distributions. In *Proceedings of the International Conference on Very Large Databases*.

Gasser, T., Engel, J., and Seifert, B. 1985. Non-parametric density estimation. *Ann. Stat.*

Gibbons, P., Matias, Y., and Poosala, V. 1997. Fast incremental maintenance of approximate histograms. In *Proceedings of the International Conference on Very Large Databases*.

Gunopulos, D., Kollios, G., Tsotras, V., and Domeniconi, C. 2000. Approximating multidimensional aggregate range queries over real attributes. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Haas, P. and Swami, A. 1992. Sequential sampling procedures for query size estimation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Hastad, J., Guruswami, V., Sudan, M., and Zuckerman, D. 2001. Combinatorial bounds for list decoding. In *IEEE Trans. Inf. Theory*.

Ibaraki, T. and Kameda, T. 1984. On the optimal nesting order of computing n-relational joins. *Trans. Datab. Syst.*

Ioannidis, Y. E. 1993. Universality of serial histograms. In *Proceedings of the International Conference on Very Large Databases*.

Ioannidis, Y. E. 2003. The history of histograms. In *Proceedings of the International Conference on Very Large Databases*.

Ioannidis, Y. E. and Christodoulakis, S. 1991. On the propagation of errors in the size of join results. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Ioannidis, Y. E. and Christodoulakis, S. 1993. Optimal histograms for limiting worst-case error propagation in the size of join results. *Trans. Datab. Syst.*

Jagadish, H., Poosala, V., Koudas, N., Sevcik, K., Muthukrishnan, S., and Suel, T. 1998. Optimal histograms with quality guarantees. In *Proceedings of the International Conference on Very Large Databases*.

Koudas, N., Muthukrishnan, S., and Srivastava, D. 2000. Optimal histograms for hierarchical range queries. In *Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM, New York.

Kushilevitz, E. and Nisan, N. 1997. *Communication Complexity*. Cambridge University Press.

Lipton, R., Naughton, J., and Schneider, D. 1990. Practical selectivity estimation through adaptive sampling. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Muralikrishna, M. and DeWitt, D. 1988. Equi-depth histograms for estimating selectivity factors for multi-dimensional queries. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Olken, F. and Rotem, D. 1986. Simple random sampling from relational databases. In *Proceedings of the International Conference on Very Large Databases*.

Piatetsky-Shapiro, G. and Connell, C. 1984. Accurate estimation of the number of tuples satisfying a condition. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Poosala, V. and Ioannidis, Y. E. 1995. Balancing histogram optimality and practicality for query result size estimation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Poosala, V., Ioannidis, Y. E., Haas, P. J., and Shekita, E. J. 1996. Improved histograms for selectivity estimation of range predicates. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Selinger, P. G., Astrahan, M. M., Chamberlin, D. D., Lorie, R. A., and Price, T. G. 1979. Access path selection in a relational DBMS. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

Vitter, J. and Wang, M. 1999. Approximate computation of multidimensional aggregates of sparse data using wavelets. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, New York.

William, S., Press, H., Flannery, B., and Vetterling, W. 1993. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press.

Yao, A. 1979. Some complexity questions related to distributive computing. In *Proceedings of the ACM Symposium on Theory of Computing*, ACM, New York.

Received October 2004; revised July 2005; accepted August 2005