

# Synopses for Query Optimization: A Space-Complexity Perspective

Raghav Kaushik\*  
Microsoft Research  
skaushi@microsoft.com

Raghu Ramakrishnan  
University of  
Wisconsin-Madison  
raghu@cs.wisc.edu

Venkatesan T.  
Chakaravarthy  
University of  
Wisconsin-Madison  
venkat@cs.wisc.edu

## ABSTRACT

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 observes that errors propagate exponentially with the number of joins in a query.

In this paper, 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, 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 pre-computed sample of the database yields probabilistic guarantees that are almost space-optimal, in the sense that in order to provide the same guarantee as sampling, any strategy requires almost the same amount of space. This 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 pre-computed 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 pre-computed samples, and also shows the increasing difference in the effectiveness of samples versus multi-dimensional histograms as the number of joins in the query grows.

## 1. INTRODUCTION

Query optimizers use synopses of the contents of a

---

\*Part of the work done while at the University of Wisconsin

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

PODS 2004 June 14-16, 2004, Paris, France.

Copyright 2004 ACM 1-58113-858-X/04/06 ... \$5.00.

database to decide the most efficient plan of execution, e.g., [31], 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 [18]. Several previous efforts such as [17, 20, 21, 22, 29] have focused on constructing single and multi-dimensional 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 [19]. 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 [2, 3] 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 paper, we study the problem of using synopses for query optimization from a space-complexity perspective. We assume an estimation model that works in two phases — a pre-processing 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 [24]. We focus on three kinds 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).

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 estima-

tion model, for the class of SPJ queries, unless the synopsis essentially contains the whole database, it is impossible to guarantee low — i.e., constant or polylogarithmic — error bounds (Corollary 3). The proof is information-theoretic and holds irrespective of whether the estimation process is deterministic or probabilistic, and covers even the simplest case of single-column selections.

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 absolute 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 [20, 17, 21, 22] in that we characterize the relative optimality of histograms versus all other synopsis structures covered by our estimation model.

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 [19]) 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 pre-computed 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 pre-computing all the joins. We show that by keeping a small sample of the pre-computed 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 7). 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. 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 multi-dimensional 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 [1], 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  $\{R_1, \dots, R_k\}$ , where relation  $R_j$  consists of a fixed ordered list of  $t$  column names  $\{C_{j_1}, \dots, C_{j_t}\}$ . In this paper, we fix  $k$  to be a constant.

An  $(N, t)$ -database instance  $I$  populates each relation  $R_j$  with a multi-set of  $N_j$  tuples, such that the maximum among all  $N_j$  is  $N$ . Each value in a tuple is drawn from

$\{1, 2, \dots, k.N.t\} \cup \{\text{null}\}$ .  $I$  is said to have  $(N, t)$ -rowcols.  $I$  is said to be an  $l(\leq k)$ -relation instance if only  $l$  relations are non-empty.

### 2.2 Queries

An  $(N, t)$ -query  $Q$  is a relational algebra expression involving the operations  $\sigma$ ,  $\pi$  and  $\bowtie$  (in other words, we only consider SPJ queries) over the  $(k, t)$  database schema and constants from  $\{1, 2, \dots, k.N.t\}$ . We assume multi-set 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 three error metrics in this paper:

- *Absolute error*:  $\text{abserr}(x, e) = (x - e, x + e)$  for  $e > 0$ , the interval consisting of all integers between, but not including  $x - e$  and  $x + e$ .
- *Ratio error*: For  $e \geq 1$ ,  $\text{ratioerr}(x, e) = (x/e, e.x)$ .
- *Relative error*: For  $0 < e < 1$ ,  $\text{reterr}(x, e) = (x(1 - e), x(1 + e))$ . Note that a relative error of  $e > 1$  is not interesting since we could always return 0, which attains the bound 1.

### 2.4 Estimation Model

An estimator  $\mathcal{E}(\mathcal{Q})$  for an  $(N, t)$ -workload  $\mathcal{Q}$  consists of a pair of functions  $\langle \mathcal{SF}, \mathcal{EF} \rangle$ , called respectively the *summarizing function* and the *estimator function*. For any  $(N, t)$ -instance  $I$ ,  $\mathcal{SF}(I)$  returns a synopsis  $\mathcal{S}$ . At optimization time, given  $Q \in \mathcal{Q}$ ,  $\mathcal{EF}(Q, \mathcal{S})$ , returns an estimate of  $\text{size}(I, Q)$ .  $\mathcal{EF}$  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  $\mathcal{S}$ . 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  $\mathcal{S}_r$  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  $\mathcal{E} = \langle \mathcal{SF}, \mathcal{EF} \rangle$  is said to *succeed* for query  $Q$  over instance  $I$  if  $\mathcal{EF}(Q, \mathcal{S}) \in \text{err}(\text{size}(I, Q), e)$ , and is said to *fail* otherwise.

- A randomized estimator  $\mathcal{E} = \langle \mathcal{SF}, \mathcal{EF} \rangle$  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  $\mathcal{EF}(Q, \mathcal{S}_r) \in \text{err}(\text{size}(I, Q), e)$ .

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  $\mathcal{E}(\mathcal{Q})$ ,  $\text{Space}(\mathcal{E})$ , is defined to be the maximum space consumed by the synopsis  $\mathcal{S}$  among all  $(N, t)$ -instances.

For error metric  $\text{err}$ , error bound  $e$ , real number  $p, 0 < p < 1$  and positive integer  $f$ , we define  $\text{LARGESPACE}_{\text{err}}(\mathcal{Q}, e, p, f)$  to be smallest  $s$  such that there is an estimator  $\mathcal{E}$  for  $\mathcal{Q}$ : (1) with  $\text{Space}(\mathcal{E}) = s$ , 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  $\text{err}$  and error bound  $e$ . We define  $\text{SPACE}_{\text{err}}(\mathcal{Q}, e, p)$  to be  $\text{LARGESPACE}_{\text{err}}(\mathcal{Q}, e, p, 1)$

## 2.6 Relationship between the error metrics

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

**PROPERTY 2.:** *Assume we have fixed a workload and an instance.*

1. *If an estimator succeeds on a non-empty 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 non-empty 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 paper, 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 which 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  $s < t \cdot \lfloor \frac{N}{c^8 f} \rfloor \cdot (\log_2 \sqrt{5} - 1) - 1$ . Consider the  $(N, t)$ -workload  $\mathcal{Q}$  of single-column equality selection queries. Fix an estimator  $\mathcal{E}$  for  $\mathcal{Q}$ , such that  $\text{Space}(\mathcal{E}) \leq s$ . 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  $n = t \cdot \lfloor \frac{N}{c^8 f} \rfloor$ . Consider the  $(k, t)$ -database schema. Let  $\mathcal{I}(N, t)$  be the family of instances obtained by placing each  $i \in \{1, 2, \dots, n\}$  in the  $(i \bmod t)^{\text{th}}$  column of  $R_1$  and setting its frequency to be one of  $\{f, f \cdot c^2, f \cdot c^4, f \cdot c^6, f \cdot c^8\}$ . 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  $\sigma_{C_{1_a}=i}(R_1), i \in \{1, 2, \dots, n\}$ , 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  $\{1, 2, \dots, n\}$ , by issuing appropriate queries from  $\mathcal{Q}$ . Given a member  $I \in \mathcal{I}(N, t)$ , we can talk about the number of  $i \in \{1, \dots, n\}$ , where the frequency in  $I'$  is within a factor of  $c$  of the frequency in  $I$  (call it  $\text{close}(I, I')$ ). The number of  $I \in \mathcal{I}(N, t)$  such that  $\text{close}(I, I') \geq \lceil \frac{n}{2} \rceil$  is at most  $\binom{n}{\lceil \frac{n}{2} \rceil} \times 5^{\lfloor \frac{n}{2} \rfloor}$ . Now, we have that:

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

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.  $\square$

### 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  $t, N, f \leq \frac{N}{8c}$ , and  $s < t \lfloor \frac{N}{8c} \rfloor \cdot (\log_2 \sqrt{5} - 1) - 1$ . Consider the  $(N, t)$ -workload,  $\mathcal{Q}$  of single-column equality selection queries. Fix an estimator  $\mathcal{E}$  for  $\mathcal{Q}$ , such that  $\text{Space}(\mathcal{E}) \leq s$ . 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  $\{f, f + 2c, f + 4c, f + 6c, f + 8c\}$ .  $\square$

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

1.  $\text{LARGESPACE}_{\text{ratioerr}}(\mathcal{Q}, c, 1/2, f) \in \Omega(N.t/f)$ . In particular,  $\text{SPACE}_{\text{ratioerr}}(\mathcal{Q}, c, 1/2) \in \Omega(N.t)$ .
2.  $\text{LARGESPACE}_{\text{relerr}}(\mathcal{Q}, 1 - 1/c, 1/2, f) \in \Omega(N.t/f)$ . In particular,  $\text{SPACE}_{\text{relerr}}(\mathcal{Q}, 1 - 1/c, 1/2) \in \Omega(N.t)$ .
3.  $\text{LARGESPACE}_{\text{abserr}}(\mathcal{Q}, c, 1/2, f) \in \Omega(N.t/f)$ . In particular,  $\text{SPACE}_{\text{abserr}}(\mathcal{Q}, c, 1/2) \in \Omega(N.t)$ .

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

In particular, if  $e(N) \in \text{polylog}(N)$ , each of these lower bounds is in  $\Omega(t.N/\text{polylog}(N))$ . 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.

## 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  $e : Z \rightarrow Z$  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 [28]. 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.

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  $\Omega(t.N/e(N))$ . Hence, we conclude that histograms are almost space optimal for single-column selections.

## 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 non-empty 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 5.:** *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  $s = t \lfloor \frac{N}{c^4} \rfloor \cdot (\log_2 \sqrt{5} - 1) - 1$ . 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 non-empty queries in  $\mathcal{Q}$ , must satisfy  $\text{Space}(\mathcal{E}) \geq s$ .*

**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.  $\square$

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 6.:** *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  $s = t \lfloor \frac{N}{8\sqrt{c}} \rfloor \cdot (\log_2 \sqrt{5} - 1) - 1$ . 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 non-empty queries in  $\mathcal{Q}$ , must satisfy  $\text{Space}(\mathcal{E}) \geq s$ .*

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.

```

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

```

Figure 1: Sampling estimator for star schema

## 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 pre-computed. For example, if a star schema has fact table *F* and dimension tables  $R_1, \dots, R_l$ , then if we define *FatF* to be the star join  $F \bowtie R_1 \bowtie \dots \bowtie R_l$ , then a star-join query of the form  $\sigma_{p_F}(F) \bowtie \sigma_{p_{i_1}}(R_{i_1}) \bowtie \dots \bowtie \sigma_{p_{i_m}}(R_{i_m})$  is equivalent to  $\sigma_{p_{i_1} \wedge \dots \wedge p_{i_m} \wedge p_F}(FatF)$ . 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 [1], to the problem of query optimization. We refer the reader to [1] 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 non-empty 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 7.:** *Let  $x \in \mathbb{Z}$  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 non-empty, 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 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  $rx/N$ , i.e., 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.

$$\begin{aligned} Pr[Y \geq (1 + \epsilon)rx/N] &\leq e^{-(\epsilon)^2(rx/3N)} \\ \Leftrightarrow Pr[NY/x \geq (1 + \epsilon)r] &\leq e^{-(\epsilon)^2(rx/3N)} \end{aligned}$$

And

$$\begin{aligned} Pr[Y \leq (1 - \epsilon)rx/N] &\leq e^{-(\epsilon)^2(rx/2N)} \\ \Leftrightarrow Pr[NY/x \leq (1 - \epsilon)r] &\leq e^{-(\epsilon)^2(rx/2N)} \end{aligned}$$

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

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

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

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

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

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

This proves the result.  $\square$

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 [5]. 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 8.:** *For any class of selection queries, the fraction of all random samples that succeed (in the sense of Theorem 7) for a fraction of at least  $f, 0 < f < 1$  of the queries is  $\geq (1 - 1/e^2 - 1/e^{16/3} - f)/(1 - f)$ .*

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 [1], by suitably maintaining multiple pre-computed 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*( $\lceil 16 \cdot N/f(N) \rceil$ ) has  $\lceil 16 \cdot N/f(N) \rceil$  rows and *t* columns. In other words, the space consumed is  $t \cdot \lceil 16 \cdot N/f(N) \rceil$  “cells”. The (probabilistic) guarantee yielded is that of a constant ratio error for all  $f(N)$ -large (single-column and multi-column) queries in addition to an absolute error of  $f(N)$  for  $f(N)$ -small queries.

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). Thus, we obtain the remarkable conclusion that sampling is almost space-optimal.

## 5. PERFORMANCE

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, multi-dimensional histograms and wavelets. The goal of this section is to study the performance of the sampling approach against the strategy of using multi-dimensional 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 non key-foreign key joins to future work.

### Analytical Comparison with Multi-Dimensional Histograms

As shown in [25], in order to provide absolute error guarantees of the form  $N/c$  for some constant  $c$ , an equi-depth multi-dimensional histograms needs a number of buckets that is exponential in the number of columns (although several multi-dimensional 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.

### Empirical Comparison

We next focus on an empirical comparison. The most recent multi-dimensional histograms proposed include the ST-holes histogram [6] and the GenHist histogram [14]. In the results reported in [6], 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.

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 multi-dimensional histograms with increasing number of joins by generating a “fattened” table with increasing number of columns.

We use the experimental setup of [6] for this purpose. The data set we use is synthetic and is based on the Gaussian distributions [33] which consist of a predetermined number of overlapping multi-dimensional Gaussian bells. The parameters for these data sets are (1) the number of Gaussian bells

$p$ , (2) the standard deviation of each bell,  $\sigma$ , and (3) a zipfian parameter  $z$  that regulates the total number of tuples contained in each Gaussian bell. We set  $p = 100$ ,  $\sigma = 25$ ,  $z = 1$  by default. The number of data points is fixed at 500000.

The query workload consists of 500 multi-dimensional 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 order to vary the number of joins, we vary the number of dimensions of the data set. 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.

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.

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. 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:

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 an order of magnitude better as the number of joins increases. 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 [14] that sampling is better for higher dimensions.
4. The errors for smaller queries are consistently larger than those for larger queries confirming our analysis

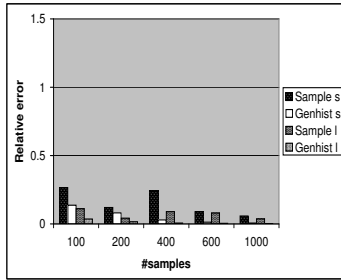


Figure 2: No. of joins = 0

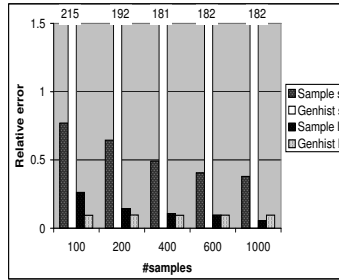


Figure 3: No. of joins = 4

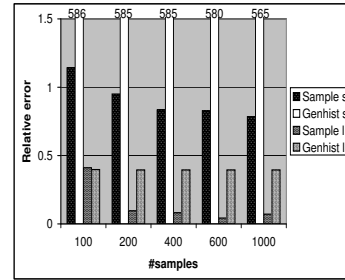


Figure 4: No. of joins = 8

that in limited space, it is difficult to obtain low relative errors.

5. The GenHist histogram is constructed by making several passes over the data, and any commercial implementation would create the histogram over a *sample* of the data. Hence, the difference between the two approaches in a commercial implementation is only likely to increase.
6. Samples on the other hand are very simple to create and algorithms for their incremental maintenance have been proposed in the literature [13].

## 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 [19] and plan space exploration [8, 16]. This paper 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 [23]. This area was introduced by Andrew Yao [34], 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 [4]. 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 [10], and statistics, for instance, in connection with non-parametric density estimation [12]. 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 [26, 24, 15, 9] have been proposed to estimate the result size of queries. The main difference in our approach is the following. First of all, we *pre-compute* 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 [1]. Here, the authors propose storing pre-computed 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 [20, 17, 21, 22, 29, 30, 14, 6, 25], both one-dimensional and multi-dimensional, wavelets [7, 32] and sketches [11, 2] have been proposed. We are not aware of any lower bounds on space complexity for any of these approaches. In addition, for multi-dimensional histograms, while there are measures of optimality such as V-optimality [29], the only properties of the form shown for sampling in this paper (in Theorem 7), that we are aware of are the ones observed in [28]. We are not aware of any such properties of wavelets. Wavelets, in addition, have the limitation that they are only applicable 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 [2]. 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 paper, 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. 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 multi-dimensional 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 multi-dimensional histograms) that provides absolute error guarantees for multi-column selections in limited space.
- While the results for sampling in this paper have been proved for key-foreign key joins, it would be interesting to extend them 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.
- In any commercial implementation of sampling, it would not in general be feasible to maintain (pre-computed) 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 paper, 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.

We hope to address these questions in future work.

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

## 8. REFERENCES

- [1] S. Acharya, P. Gibbons, V. Poosala, and S. Ramaswamy. Join synopses for approximate query answering. In *SIGMOD*, 1999.
- [2] N. Alon, P. Gibbons, Y. Matias, and M. Szegedy. Tracking join and self-join sizes in limited storage. In *PODS*, 1999.
- [3] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In *STOC*, 1996.
- [4] Z. Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, Department of Computer Science, University of California-Berkeley, 2002.
- [5] N. Bruno and S. Chaudhuri. Statistics on query expressions. In *SIGMOD*, 2002.
- [6] N. Bruno, S. Chaudhuri, and L. Gravano. A multi-dimensional workload-aware histogram. In *SIGMOD*, 2001.
- [7] K. Chakrabarti, M. Garofalakis, R. Rastogi, and K. Shim. Approximate query answering using wavelets. In *VLDB*, 2000.
- [8] S. Chatterji, S. S. K. Evani, S. Ganguly, and M. D. Yemmanuru. On the complexity of approximate query optimization. In *PODS*, 2002.
- [9] S. Chaudhuri, R. Motwani, and V. Narasayya. On random sampling over joins. In *SIGMOD*, 1999.
- [10] S. D. Conte and C. de Boor. *Elementary Numerical Analysis: An Algorithmic Approach*. McGraw Hill Publishing Company, 1972.
- [11] A. Dobra, M. Garofalakis, J. E. Gehrke, and R. Rastogi. Processing complex aggregate queries over data streams. In *SIGMOD*, 2002.
- [12] T. Gasser, J. Engel, and B. Seifert. Non-parametric density estimation. In *Ann. Stat.*, 1985.
- [13] P. Gibbons, Y. Matias, and V. Poosala. Fast incremental maintenance of approximate histograms. In *VLDB*, 1997.
- [14] D. Gunopulos, G. Kollios, V. Tsotras, and C. Domeniconi. Approximating multi-dimensional aggregate range queries over real attributes. In *SIGMOD*, 2000.
- [15] P. Haas and A. Swami. Sequential sampling procedures for query size estimation. In *SIGMOD*, 1992.
- [16] T. Ibaraki and T. Kameda. On the optimal nesting order of computing n-relational joins. In *TODS*, 1984.
- [17] Y. E. Ioannidis. Universality of serial histograms. In *VLDB*, 1993.
- [18] Y. E. Ioannidis. The history of histograms. In *VLDB*, 2003.
- [19] Y. E. Ioannidis and S. Christodoulakis. On the propagation of errors in the size of join results. In *SIGMOD*, 1991.
- [20] Y. E. Ioannidis and S. Christodoulakis. Optimal histograms for limiting worst-case error propagation in the size of join results. In *TODS*, 1993.
- [21] H. Jagadish, V. Poosala, N. Koudas, K. Sevcik, S. Muthukrishnan, and T. Suel. Optimal histograms with quality guarantees. In *VLDB*, 1998.
- [22] N. Koudas, S. Muthukrishnan, and D. Srivastava. Optimal histograms for hierarchical range queries. In

*PODS*, 2000.

- [23] E. Kushilevitz and N. Nisan. *Communication Complexity*. Cambridge University Press, 1997.
- [24] R. Lipton, J. Naughton, and D. Schneider. Practical selectivity estimation through adaptive sampling. In *SIGMOD*, 1990.
- [25] M. Muralikrishna and D. DeWitt. Equi-depth histograms for estimating selectivity factors for multi-dimensional queries. In *SIGMOD*, 1988.
- [26] F. Olken and D. Rotem. Simple random sampling from relational databases. In *VLDB*, 1986.
- [27] C. Papadimitriou. *Computational Complexity*. Addison Wesley, 1994.
- [28] G. Piatetsky-Shapiro and C. Connell. Accurate estimation of the number of tuples satisfying a condition. In *SIGMOD*, 1984.
- [29] V. Poosala and Y. E. Ioannidis. Balancing histogram optimality and practicality for query result size estimation. In *SIGMOD*, 1995.
- [30] V. Poosala, Y. E. Ioannidis, P. J. Haas, and E. J. Shekita. Improved histograms for selectivity estimation of range predicates. In *SIGMOD*, 1996.
- [31] P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. Access path selection in a relational DBMS. In *SIGMOD*, 1979.
- [32] J. Vitter and M. Wang. Approximate computation of multidimensional aggregates of sparse data using wavelets. In *SIGMOD*, 1999.
- [33] S. William, H. Press, B. Flannery, and W. Vetterling. *Numerical Recipes in C: The art of scientific computing*. Cambridge University Press, 1993.
- [34] A. Yao. Some complexity questions related to distributive computing. In *STOC*, 1979.