A Bayesian Approach to Estimating the Selectivity of Conjunctive Predicates

Max Heimel, Volker Markl, Keshava Murthy
IBM Germany, TU Berlin, IBM San Jose
mheimel@de.ibm.com, marklv@cs.tu-berlin.de, rkeshav@us.ibm.com

Abstract: Cost-based optimizers in relational databases make use of data statistics to estimate intermediate result cardinalities. Those cardinalities are needed to estimate access plan costs in order to choose the cheapest plan for executing a query. Since statistics are usually collected on single attributes only, the optimizer can not directly estimate result cardinalities of conjunctive predicates over multiple attributes. To avoid having to fall back to assuming statistical independence, modern relational database systems offer the possibility to additionally collect joint statistics over multiple attributes. These statistics allow a direct cardinality estimate for conjunctive predicates.

A widely used approach is collecting the number of distinct value combinations as a joint statistic. This can be used for a uniformity based estimate, which assumes each value combination to occur equally often. Although this estimate is likely an improvement, it is still inaccurate, since "real world" data is unlikely to be uniform.

This paper proposes a new approach of estimating the result cardinality of conjunctive predicates over multiple attributes of a relation. The proposed method combines knowledge from single-column histograms using a conditional probability based "uniform correlation" approach. Initial evaluation shows that this method yields better results for estimating predicates on highly correlated attributes than the classic uniformity based approach.

1 Introduction

Modern relational database systems make use of cost-based optimizers to generate the most efficient access plan for a given SQL query. A cost-based optimizer generates a set of possible plans for a given query and estimates the execution cost for each of those plans. The plan with the lowest estimated cost is chosen for execution.

In order to estimate plan costs, the optimizer has to know the cardinalities of intermediate operations contained in the plan. The cost of a relational operation depends on how large the input relations are, therefore these cardinalities are necessary to give an accurate cost estimate. Cost-based optimizers usually use data statistics - e.g. data histograms - to estimate those cardinalities.

Data statistics are usually collected for single attributes only. This allows the optimizer to estimate how many tuples of a relation fulfill a condition for a single attribute. Since the available statistics normally do not contain information about dependency patterns between attributes of a relation, the optimizer has to assume statistical independence between filtered attributes when multiple conjunctive attribute conditions have to be considered. In the case of a filter that queries multiple attributes that share some kind of dependency, the independence assumption can lead to drastically underestimated result cardinalities. Especially in the case of complex SQL queries, those underestimated cardinalities will propagate through the cost estimation - effectively resulting in drastically underestimated plan costs. This error might bias the optimizer into choosing a non-optimal plan for execution.

In order to avoid incorrect independence assumptions during query optimization, modern cost-based optimizers make use of multi-dimensional statistics. Since those statistics are joint over multiple attributes, they inherently include information about potential dependency patterns. Since multi-dimensional histograms are expensive to build and maintain, most database systems use simpler methods. This usually involves collecting the number of distinct combined values of a set of attributes and assuming each of those distinct combinations is equally likely to occur.

Although this simple uniformity assumption leads to better estimations, it also introduces new errors. In reality, the distinct value combinations will almost never occur equally likely. Since joint data statistics are currently not combined with available single-attribute distribution knowledge, the optimizer is basically "giving away" available information on how the combined attributes are distributed. We tackle this problem by proposing a novel approach to achieve better estimates by including available single-predicate knowledge into the cardinality estimate.

2 Related Work

The concept of cost-based optimizers in relational database systems was introduced in system R by IBM. [SAC$^+$79] describes the optimizer of system R and introduces concepts such as selectivity estimation and the independence assumption for multiple attributes. [PSC84] further investigates the selectivity estimation methods presented by system R. Although system R is already 30 years old, the introduced concepts are still valid today. A more modern look at query optimization can be found in [Cha98], which summarizes the major research topics in this field during the last 30 years.

While those three papers give a more high-level overview about the concepts of relational query optimization, there are also several papers dealing with actual methods of performing estimations. [Ioa03] gives an overview over previous and current work in the field of selectivity estimation using single-attribute data histograms. For constructing multi-dimensional histograms, [MPS99] compares multiple algorithms for partitioning a two-dimensional histogram into buckets and shows that finding the optimal partitioning is essentially NP-hard. Due to this, practical multi-dimensional histogram techniques are using either heuristics or query feedback to perform the partitioning. Examples of practical multi-dimensional histograms include [TGIK02] and [SHM$^+$06]. An overview comparing different types of multi-dimensional histograms can be found in [PI97]. [GTK01] presents a slightly different approach for the multi-dimensional estimation, which uses probabilistic models (e.g. Bayesian Networks) to compute a selectivity estimate.

[MHK+07] shows how to improve estimation quality by using the maximum entropy principle to consistently combine estimates from available multi-dimensional statistics. In contrast to this, our approach introduces a novel and improved way of computing a direct cardinality estimate from one statistics set. These improved estimates could then e.g. be used as input for a maximum entropy based estimator, which should effectively improve cardinality estimates even in cases where no fitting set of statistics is available.

3 Problem & Terminology

A conjunctive filter θ consisting of n single predicates is defined as:

$$\theta = \bigwedge_{i=1}^{n} \theta_i \tag{1}$$

For this paper, we assume that each single predicate θ_i is an equality predicate of the form $a_i = v_i$, which can be evaluated as:

$$\theta_i(t) = \begin{cases} true, & \text{if } t(a_i) = v_i \\ false, & else \end{cases} \tag{2}$$

Hereby t is a tuple of relation R, a_i denotes an attribute name of relation R and v_i is a constant value. The expression $t(a_i)$ denotes the value of attribute a_i in tuple t. We will assume that all attributes are from the same relation R and that no attribute is queried twice by θ. This constraint is needed, since a conjunct of two different filters on the same attribute is unsatisfiable and would lead to incorrect estimations.

When no multi-dimensional statistics are available, the optimizer will estimate the cardinality of a conjunctive filter θ by assuming statistical independence between the single attributes. Using this, the cardinality of θ is estimated as:

$$|\sigma_\theta (R)| \approx \frac{1}{|R|^{n-1}} \prod_{i=1}^{n} |\sigma_{\theta_i} (R)| \tag{3}$$

Hereby σ_θ denotes the relational selection, which returns the set of tuples from relation R that fulfill filter θ.

Let us consider a relation describing cars as tuples with attributes *make* and *model*. From data histograms, the optimizer knows that 500 cars are manufactured by "*Opel*" and that 100 cars are of model "*Astra*". Assuming 10.000 tuples within the database, the independence assumption tells us there are 5 cars of model "*Astra*" produced by "*Opel*". This is obviously wrong, since only "*Opel*" produces the "*Astra*": the strong correlation between attributes *make* and *model* leads to an underestimated result cardinality.

One way to avoid underestimated cardinalites due to incorrect independence assumptions is by applying a uniformity based estimation. Hereby, the optimizer assumes that all existing value combinations occur equally likely:

$$|\sigma_\theta (R)| \approx |R| \frac{1}{\vartheta_\theta} \tag{4}$$

In this equation ϑ_θ denotes the number of distinct value combinations in the attribute set queried by filter θ. Using the uniformity based approach, let us reconsider the car example. If we assume there are 125 different value combinations, a uniformity approach would estimate the cardinality for each combination as 80 tuples. This estimate is obviously better than the independence based estimate of 5 tuples.

Although it generally leads to better estimates - especially in case of incorrect independence assumptions - the uniformity approach introduces a new estimation error source. If, for example, we choose to find all cars of model "*F430*" that were produced by "*Ferrari*", the uniformity approach would still result in an estimate of 80 tuples. This is presumably incorrect, since cars by Ferrari are likely to occur much less frequently than cars by Opel. The problem of the uniformity approach is that each combination is treated equally - although they might occur with different frequencies. By including single predicate knowledge from data statistics into the estimate we should be able to improve this. For example: if we know that *Opel* occurs much more often than *Ferrari* within the relation, we should be able to use this knowledge to adjust estimates accordingly.

4 Solution

The relational selection operator can be expressed in form of a probabilistic experiment: We take a random tuple out of relation R and check whether it fulfills filter θ. The probability that the random tuple fulfills the filter condition is:

$$P\left(\theta\left(t\right)\right) = \frac{|\sigma_\theta(R)|}{|R|} \tag{5}$$

Through rearranging equation 5, we are able to express the cardinality of filter θ as:

$$\left|\sigma_\theta\left(R\right)\right| = P\left(\theta\left(t\right)\right)|R| \tag{6}$$

We are thus able to give a cardinality estimate through estimating the filter probability. In the context of relational operations, the "probability" of a filter is called *selectivity*. In short: the selectivity tells us the fraction of tuples in a relation, that fulfil the filter.

Using the probability based approach, we are able to incorporate single-attribute knowledge into the selectivity estimate. We can do this by rewriting the joint probabilty through using *conditional probabilities*. The conditional probability $P(A|B)$ denotes the probability that event A will happen under the condition that event B has already occured. The joint probability is then the product of $P(A|B)$ and $P(B)$.

If we translate the definition of conditional probabilites to estimating the selectivity for a conjunctive pair of predicates θ_1 and θ_2 we get:

$$\begin{aligned} P\left(\theta_1\left(t\right) \wedge \theta_2\left(t\right)\right) &= P\left(\theta_2\left(t\right)|\theta_1\left(t\right)\right)P(\theta_1\left(t\right)) \\ &= P\left(\theta_1\left(t\right)|\theta_2\left(t\right)\right)P(\theta_2\left(t\right)) \end{aligned} \tag{7}$$

From now on, we will refer to the conditional probabilites in equation 7 as *conditional selectivities*. Since the two single predicate selectivities $P(\theta_1\left(t\right))$ and $P(\theta_2\left(t\right))$ can be retrieved from data statistics we are able to estimate the joint selectivity through estimating these conditional selectivities. A very simple way of doing this, is to assume a *uniform correlation* between the two queried attributes. This means we assume the conditional selectivites to be constant over the complete data range:

$$\begin{aligned} P\left(\theta_1\left(t\right) \wedge \theta_2\left(t\right)\right) &\approx \gamma_{\theta_1} P(\theta_1\left(t\right)) \\ &\approx \gamma_{\theta_2} P(\theta_2\left(t\right)) \end{aligned} \tag{8}$$

Equation 7 shows two alternatives for estimating the joint selectivity: one for each part of the filter. The approximation introduced in 8 results in estimation errors of different size for the two alternatives. Since we can't really predict which of the two ways results in the smaller error, we use the average of both alternatives as the final estimate:

$$P\left(\theta_1\left(t\right) \wedge \theta_2\left(t\right)\right) \approx \frac{\gamma_{\theta_1} P(\theta_1\left(t\right)) + \gamma_{\theta_2} P(\theta_2\left(t\right))}{2} \tag{9}$$

The conditional selectivity constants γ_{θ_1} and γ_{θ_2} tell us how many values of the other attribute are - in general - paired with one attribute value. We are able to uniformly estimate those constants as:

$$\gamma_{\theta_i} = \frac{\vartheta_{\theta_i}}{\vartheta_{\theta}} \tag{10}$$

Hereby ϑ_{θ_i} denotes the number of distinct values in the attribute queried by predicate θ_i. By combining equations 6, 9 and 10 we get the final conditional probability based cardinality estimation as:

$$\left|\sigma_{\theta}\left(R\right)\right| \approx \frac{|R|}{2}\left(\frac{\vartheta_{\theta_1}}{\vartheta_{\theta}} P\left(\theta_1\left(t\right)\right) + \frac{\vartheta_{\theta_2}}{\vartheta_{\theta}} P\left(\theta_2\left(t\right)\right)\right) \tag{11}$$

Using the same approach as for predicate pairs, we can generalize formula 11 to conjuncts of n predicates. The resulting estimation formula is:

$$\left|\sigma_{\bigwedge_{i=1}^{n} \theta_i}\left(R\right)\right| \approx \frac{|R|}{n} \sum_{i=1}^{n} \frac{\vartheta_{\theta_i}}{\vartheta_{\theta}} P\left(\theta_i(t)\right) \tag{12}$$

Let us reconsider the car example from chapter 3 using the conditional selectivity based approach. We will assume there are in total 115 distinct models and 25 distinct makes in the database, this allows us to estimate the two conditional selectivity constants for *make* and *model* as:

$$\gamma_{make} = \frac{25}{125} = 0.2 \qquad \gamma_{model} = \frac{115}{125} = 0.92$$

If we assume there are 15 cars of make *Ferrari* and two cars of model *F430*, we are able to use equation 11 for the cardinality estimation. Equation 11 gives us 96 *Astra* cars from *Opel* and 2.42 *F430* cars from *Ferrari*. Those estimates are nearly perfect (we would expect 100 and 2) and show the potential of the conditional selectivity method compared to the uniformity approach.

5 Evaluation

For evaluation, we want to identify the expected general behaviour of the conditional selectivity based estimator compared to the two classic ones (independence and uniformity). The quality of an estimator is dependent on how correct the underlying assumption is for a given relation. For the classic estimators we can give rather accurate predictions:

- The independence assumption based estimator will obviously result in very bad estimations for distributions with high correlation between the attributes.
- In case of the uniformity approach, the estimation error is mainly influenced by the variance of the frequencies within the distribution. A highly varianced frequency distribution should thus result in high errors when using uniformity based estimation.

Evaluating the conditional selectivity method is a bit more complicated, since the used assumption of "uniform correlation" is not as tangible as the other assumptions. However, we can give some general thoughts on the expected behaviour. We expect the conditional selectivity method to work best in case of highly correlated distributions. In these cases, the attributes are very uniformly correlated - in the sense that each value of an attribute has almost exactly one partner in the other attribute(s). For weakly correlated distributions we expect the conditional selectivity method to behave nearly like the classic independence assumption based estimator, since the conditional selectivity estimator approximates the independence estimator in those cases[1].

Figure 1 shows how the estimation error of the three possible estimators (independence, uniformity, conditional selectivity) develops with regard to the correlation factor. For creating the figure, we implemented the conditional selectivity approach into the query optimizer of IBM Informix Dynamic Server and measured the total error for simple select statements on a relation consisting of two attributes. The relation was filled with artificial numerical data with specified variance and correlation, having a skewed frequency distribution to simulate "real-world" data. For figure 1 we created multiple data sets with a

[1] In lowly correlated distributions, the single attributes are nearly statistically independent. This means the number of distinct attribute value combinations approaches the product of the number of distinct attribute values:

$$\vartheta_{\theta} \approx \prod_{j} \vartheta_{\theta_j} \tag{13}$$

If we insert equation 13 into the conditional selectivity estimation formula 12 we get:

$$\left|\sigma_{\bigwedge_{i=1}^{n} \theta_i}(R)\right| \approx \frac{|R|}{n} \sum_{i=1}^{n} P\left(\theta_i(t)\right) \prod_{j \neq i}^{n} \frac{1}{\vartheta_{\theta_j}} \tag{14}$$

We can now apply a simple uniformity based approximation:

$$\left|\sigma_{\bigwedge_{i=1}^{n} \theta_i}(R)\right| \approx \frac{|R|}{n} \sum_{i=1}^{n} P\left(\theta_i(t)\right) \prod_{j \neq i}^{n} P\left(\theta_j(t)\right) \approx |R| \prod_{i=1}^{n} P\left(\theta_i(t)\right) \tag{15}$$

Comparing the right-hand side of equation 15 to the definition of independence assumption based estimation in 3 shows that they are equivalent.

constant variance and increasing correlation factors[2].

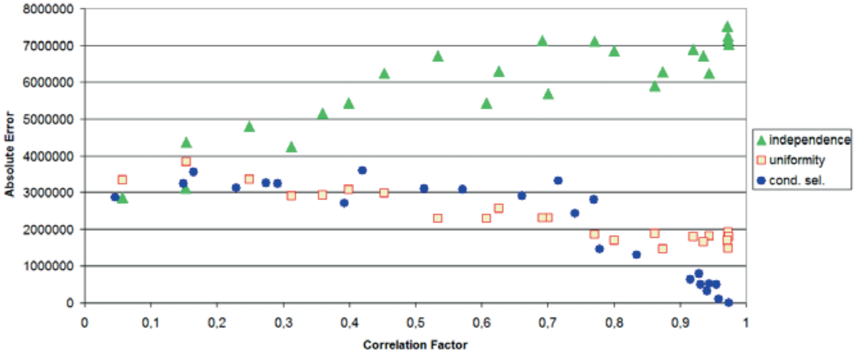


Figure 1: Theoretical comparison of estimators

There are three notable conclusions from figure 1:

1. The independence assumption leads to very high errors in case of high (and medium) correlation factors. The other two estimators lead to drastically better results!

2. The uniformity based estimation leads to a rather constant estimation error, which is not surprising since all data sets we used had roughly the same frequency variance.

3. The conditional selectivity based estimation seems to create roughly the same estimation errors as the uniformity based estimation. However - as we expected - the error decreases drastically for very highly correlated attributes.

6 Experiments

In order to confirm the statements from the evaluation chapter, this section presents experiments conducted using the conditional selectivity method. We compare the cardinality estimation quality of the conditional selectivity and the uniformity approach, using the implementation in IBM Informix Dynamic Server. The experiments were performed on attribute pairs of existing “real-world” customer data. Each attribute pair was queried multiple times using simple conjunctive select queries. To compare the estimation quality, each workload was run twice: once using the classic uniformity estimator and once using the conditional selectivity estimator. Table 1 shows statistical properties of the used groups and whether the estimation improved or deterioated: values larger than one represent an improvment by using the conditional selectivity estimator.

[2] The used correlation factor is the ϕ^2 factor from [IMH$^+$04], which ranges between 0.0 - meaning independence - and 1.0 - meaning complete dependency.

Nr.	#rows	#distinct	variance	ϕ^2	improvement
1	10001	193	264.019	0.992588	- (inf)
2	14260	39	297.967	0.894439	1.753
3	12096	280	105.311	0.631532	1.506
4	12096	40	669.102	0.536271	0.828
5	10001	691	58.3671	0.201792	1.288
6	8872	1832	3.97732	0.00521	2.165

Table 1: Evaluated attribute pairs

To further illustrate the results, figure 2 shows scatter plots presenting the change in estimation errors for all six groups. Each dot of the plot represents one query, the x-coordinate being the estimation error using the uniformity approach, the y-coordinate the error using the conditional selectivity approach. Therefore each dot between the two lines represents a query for which the conditional selectivity approach resulted in an improved estimate.

Looking at the plot for group one, it shows all[3] points lying on the x-axis. This means the conditional selectivity approach resulted in - as expected - perfect estimates for the group with the highest correlation. The plots for groups two and three show most points in between the two lines: For these groups, the conditional selectivity approach results in a reasonable improvement. The same is true for groups five and six. The only group for which the classic uniformity approach worked (slightly) better was group four, which is mid-correlated and has a rather high variance.

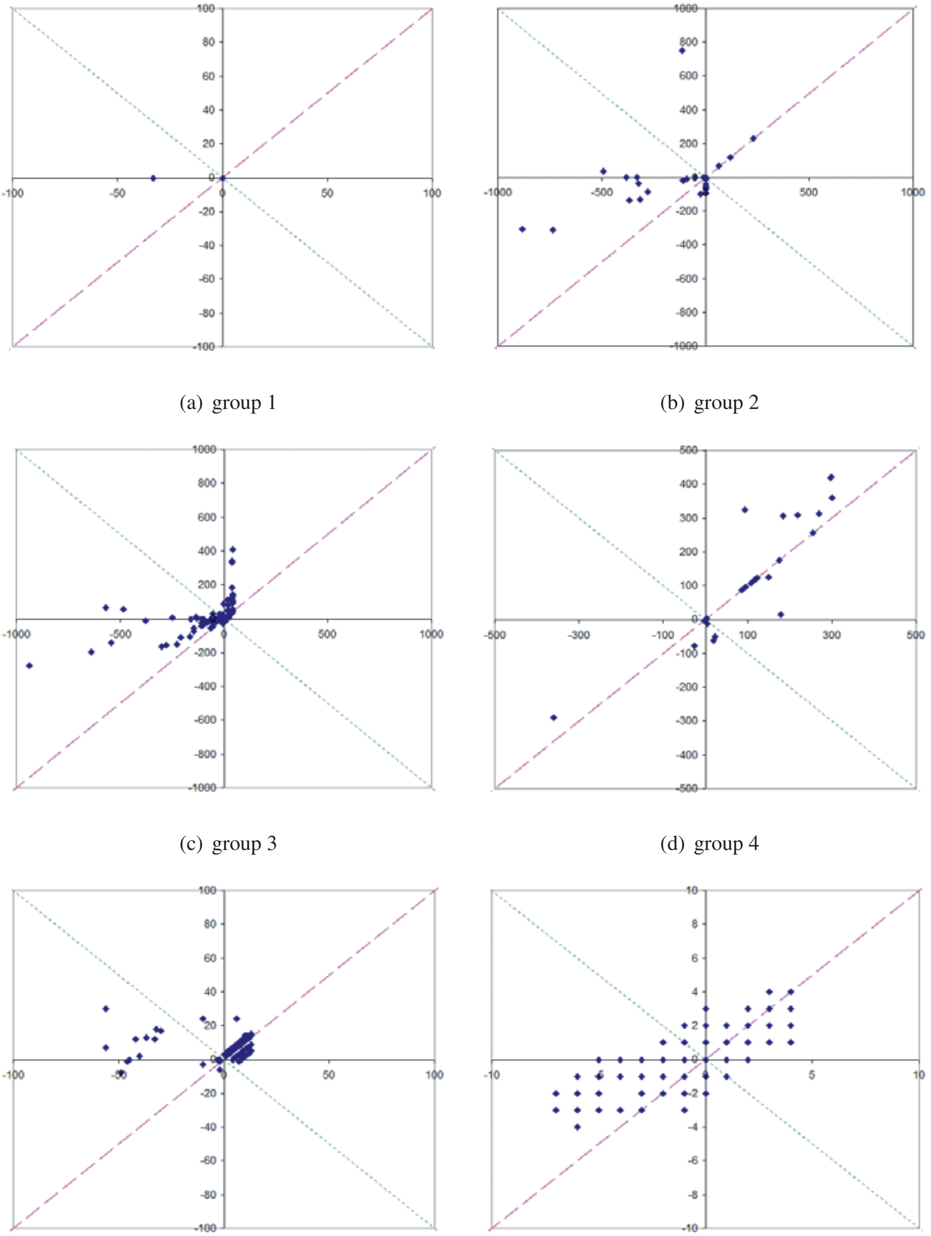
The experimental results are basically consistent with the considerations from the previous chapter: The conditional selectivity based approach seems to result in better estimates for lowly and especially for highly correlated groups. For mid-correlated groups, the two estimators seem to be equally good, with a slight advantage for the uniformity approach.

7 Conclusion & Future Work

Summarizing the results, we can say that the conditional selectivity approach offers nearly perfect estimations for highly correlated attributes - without the need for the expensive construction and maintenance of a multidimensional histogram. Although the estimation quality decreases for lower correlated attributes, the method still seems to at least match the uniformity approach in most cases. Thus, the conditional selectivity based approach provides a clear advantage for highly correlated attributes, while it seemingly doesn't have a serious disadvantage on mildly correlated ones: The method looks very promising and should be studied further in the future.

There are two interesting topics future studies could concentrate on: the first topic is finding heuristics for determining cases in which the uniformity approach works better than the conditional selectivity approach. These heuristics could be used by an "intelligent op-

[3] Since many queries led to the same estimation errors, it appears there are only two points in the plot.



(a) group 1

(b) group 2

(c) group 3

(d) group 4

(e) group 5

(f) group 6

Figure 2: Comparing uniformity and conditional selectivity based estimation

timizer" to automatically determine the best possible estimator for a given relation. The second topic focuses on optimizing the conditional selectivity factors. There is probably a better way of choosing those factors than the proposed approach, which should lead to a better estimation quality for the conditional selectivity approach.

References

[Cha98] Surajit Chaudhuri. An overview of query optimization in relational systems. In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 34–43, New York, NY, USA, 1998. ACM.

[GTK01] Lise Getoor, Benjamin Taskar, and Daphne Koller. Selectivity estimation using probabilistic models. *SIGMOD Rec.*, 30(2):461–472, 2001.

[IMH$^+$04] Ihab F. Ilyas, Volker Markl, Peter Haas, Paul Brown, and Ashraf Aboulnaga. CORDS: automatic discovery of correlations and soft functional dependencies. In *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 647–658, New York, NY, USA, 2004. ACM.

[Ioa03] Yannis Ioannidis. The history of histograms (abridged). In *VLDB '2003: Proceedings of the 29th international conference on Very large data bases*, pages 19–30. VLDB Endowment, 2003.

[MHK$^+$07] V. Markl, P. J. Haas, M. Kutsch, N. Megiddo, U. Srivastava, and T. M. Tran. Consistent selectivity estimation via maximum entropy. *The VLDB Journal*, 16(1):55–76, 2007.

[MPS99] S. Muthukrishnan, Viswanath Poosala, and Torsten Suel. On Rectangular Partitionings in Two Dimensions: Algorithms, Complexity, and Applications. In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pages 236–256, London, UK, 1999. Springer-Verlag.

[PI97] Viswanath Poosala and Yannis E. Ioannidis. Selectivity Estimation Without the Attribute Value Independence Assumption. In *VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases*, pages 486–495, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[PSC84] Gregory Piatetsky-Shapiro and Charles Connell. Accurate estimation of the number of tuples satisfying a condition. In *SIGMOD '84: Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, pages 256–276, New York, NY, USA, 1984. ACM.

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

[SHM$^+$06] U. Srivastava, P. J. Haas, V. Markl, M. Kutsch, and T. M. Tran. ISOMER: Consistent Histogram Construction Using Query Feedback. In *ICDE '06: Proceedings of the 22nd International Conference on Data Engineering*, page 39, Washington, DC, USA, 2006. IEEE Computer Society.

[TGIK02] Nitin Thaper, Sudipto Guha, Piotr Indyk, and Nick Koudas. Dynamic multidimensional histograms. In *SIGMOD '02: Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 428–439, New York, NY, USA, 2002. ACM.