A Causal Model of DBMS Suboptimality

The query optimization phase within a DBMS ostensibly finds the fastest query execution plan from a potentially large set of enumerated plans, all of which correctly compute the specified query. Infrequently the optimizer selects the wrong plan, for a variety of reasons. *Suboptimality* is indicated by the existence of a query plan that performs more efficiently than the DBMS' chosen plan, for the same query. From the engineering perspective, it is of critical importance to understand the prevalence of suboptimality and its causal factors. We study these aspects across DBMSes to identify the underlying causes. We propose a novel structural causal model to explicate the relationship between various factors in query optimization and suboptimality. Our model associates suboptimality with the factors of complexity of the query and optimizer and concomitant unanticipated interactions among the components of the optimizer. This model induces a number of specific hypothesis that were subsequently tested on multiple DBMSes. We observe that suboptimality prevalence correlates positively with the number of operators available in the DBMS (one proxy for optimizer complexity) and with the number of plans generated by the optimizer (another proxy for optimizer complexity) and with the number of correlation names (a measure of query complexity), providing empirical support for this model as well as implications for fundamental improvements of these optimizers. The results of this study also uncovered evidence of a potential fundamental limitation of cost-based optimization. **TODO:** *be more specific!* This paper thus provides a new methodology to study mature query optimizers, that of empirical generalization, proposes a novel causal model for query suboptimality, and tests several hypotheses deriving from this causal model.

ACM Reference Format:
A Causal Model of DBMS Suboptimality *ACM Trans. Datab. Syst.* 1, 1, Article 1 (February 2016), 37 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

DBMSes underlie all information systems and hence optimizing their performance is of critical importance. The DBMS's query optimizer plays an important role. But what if the optimizer *doesn't*: what if it selects the wrong plan?

This paper provides a thorough investigation into DBMS *suboptimality*, when the DBMS chooses a slower plan over a faster plan for a given query. We systematically examine the factors influencing the number of suboptimal queries. There could be multiple causes of the suboptimality. One possible cause could be some peculiarity within the tens of thousands of lines of code of that query optimizer. Another possible cause could be the query's complexity. A third possible reason could be some fundamental limitation *within the general architecture of cost-based optimization* that will always render a good number of queries suboptimal. Prior research in other domains shows that increasing the complexity negatively influences performance [3; 21].

To better understand the impact of different factors on suboptimality of query performance and the interaction between operators, especially in a dynamic environment, an experimental approach is needed. Our research introduces a novel approach to better understand the factors influencing query performance in DBMSes. Based on existing research and general knowledge of DBMSes, we propose an innovative predictive model to understand the presence and influence of suboptimality in query evaluation. We use an experimental methodology with the DBMS as a subject to test our hypotheses with respect to factors influencing suboptimality, applying to our experiment data collected over about a cumulative 30 months of query executions (over 21,000 hours to run almost a million query executions). **TODO:** *Young: Table I says something different: over 16,000 hours ro run almost two million query exec*
Our research falls within creative development of new evaluation methods and metrics for artifacts, which were identified as important design-science contributions [10].

The key contributions of this paper are as follows.

— We use an innovative *methodology* that treats DBMSes as experimental subjects.
— We find that for a surprisingly large portion of queries, the plan chosen by the query optimizer is not the best plan, for some cardinality, of the underlying tables.
— We propose and test a *predictive model* for DBMSes to better understand the factors causing suboptimality.
— **TODO:** *Rick:Mention generational result*
— The predictive model and these experimental results suggest several specific engineering directions.

This paper takes a scientifically rigorous approach to an area previously dominated by the engineering perspective, that of database query optimization. Our goal is to understand cost-based query optimizers as a *general* class of computational artifacts and to come up with insights and ultimately with predictive theories about how such optimizers, again, as a general class, behave. These theories can be used to further improve DBMSes through engineering efforts that benefit from the fundamental understanding that the scientific perspective can provide.

One might ask, shouldn't the task of optimizing queries be left to DBAs? In databases, and especially in data warehouses, the number of users writing and running queries has been growing exponentially. This growth is aided, in part, by the drag and drop query tools provided by the different systems. For example, subject areas allow business users to write queries without knowing anything about the underlying database structure. This, coupled with constantly growing data presents new challenges for tuning. For example, a large data warehouse we are familiar with runs about 30,000 queries on a daily basis. Also, tuning a subject area or tables for one group of queries can negatively impact the performance of other queries. Query optimization experts often take hours to tune and test existing canned queries; the amount expended on this one system for manual query optimization approaches $100K/year. Therefore, we argue that it is important to understand how existing query optimizers can be further improved, and indeed, fundamental limitations inherent in these optimizers.

We focus here on the effectiveness of query optimization. The query optimization phase within a DBMS ostensibly finds the fastest query execution plan from a potentially large set of enumerated plans, all of which correctly compute the specified query. Occasionally the "optimizer" selects a suboptimal plan, for a variety of reasons. We define *suboptimality* as the existence of a query plan that performs more efficiently than the plan chosen by the query optimizer. From the engineering perspective, it is of critical importance to understand the phenomenon of suboptimality.

We study these aspects across DBMSes to identify the underlying causes. We have developed an initial predictive causal model that identifies four factors that may play a role in suboptimality. The ultimate goal is to *understand* a component within a DBMS, its cost-based optimizer, through the articulation and empirical testing of a general scientific theory.

In Section 3 we briefly summarize the vast amount of related work in query optimization to establish the technical basis for our study. The following section introduces the methodology we will follow, that of empirical generalization. We then present a predictive, causal model of suboptimality and empirically test eight specific hypotheses derived from that model on three DBMSes. These are the first scientific results that we are aware of that apply *across* DBMSes, rather than on a single, specific DBMS or on a specific algorithm. This model has implications for research in engineering more efficient DBMSes.

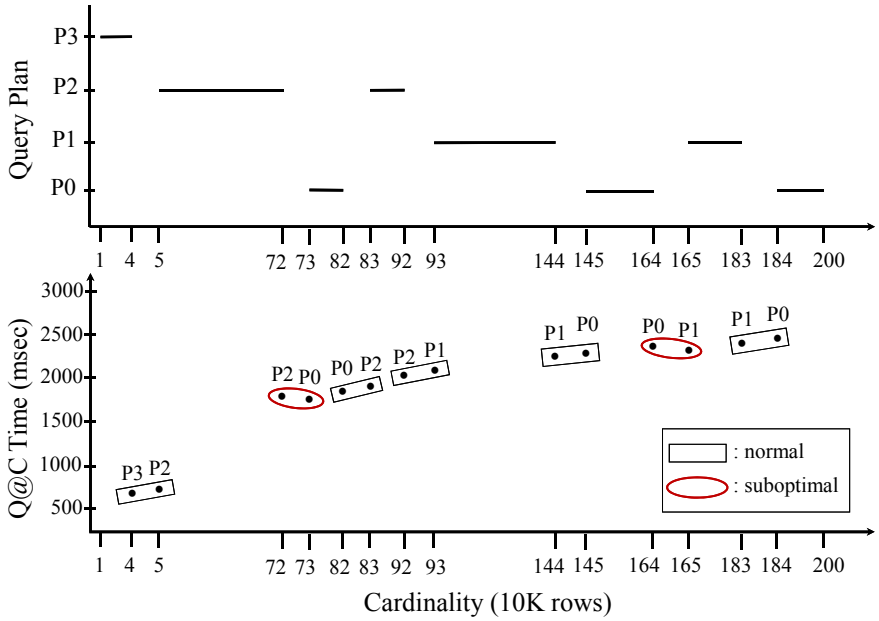

Fig. 1. An Example of Suboptimality and Fluttering

2. MOTIVATION

Consider a simple select-project-join (SPJ) query, with a few attributes in the SELECT clause, a few tables referenced in the FROM clause, and a few equality predicates in the WHERE clause. This query might be an excerpt from a more complex query, with the tables being intermediate results.

```
SELECT t0.id1, t0.id2, t2.id4, t1.id1
FROM ft_HT3 t2, ft_HT2 t1, ft_HT1 t0
WHERE (t2.id4=t1.id1 AND t2.id1=t0.id1)
```

The optimizer generates different plans for this query as the cardinality of the `ft_HT1` table varies, an experiment that we will elaborate later in depth.

The upper graph in Figure 1 represents the plans chosen by a common DBMS as the cardinality of `FT_HT1` decreases from 2M tuples to 10K tuples in units of 10K tuples. The x-axis depicts the estimated cardinality and the y-axis a plan chosen for an interval of cardinalities. So Plan P0 was chosen for 2M tuples, switching to Plan P1 at 1,830,000 tuples, back to Plan P0 at 1,640,000 tuples, and so on, through the sequence P0, P1, P0, P1, P2, P0, P2, and finally P3 at 40,000 tuples.

The lower graph in Figure 1 indicates the query times executed at adjacent cardinalities when the plan changed, termed the "query-at-cardinality" (Q@C) time. For some transitions, the Q@C time at the larger cardinality was also larger, as expected. But for other transitions, emphasized in red ovals, the Q@C time at the larger cardinality was smaller, such as the transition from plan P1 at 1,650,000 to P0 at 1,640,000 tuples. Such pairs identify suboptimal plans. For the pair at 720,000 tuples, P0 required 2.35sec whereas P1 at a larger cardinality required only 2.41sec. This query exhibits seven plan change points, two of which are suboptimal.

This query also illustrates an interesting phenomenon, in which the optimizer returns to an *earlier* plan. Sometimes the optimizer starts oscillating between two plans,

sometimes even switching back and forth when the cardinality estimate changes by a small percentage. The example query showed returning to P0 twice and to P1 and to P2 each once. We call this phenomenon, in which the query optimizer returns to a previous plan, "query optimizer flutter," or simply "flutter."

We have found through our experiments that flutter and suboptimality are all around us: *every* DBMS that we have examined, including several open source DBMSes and several proprietary DBMSes, covering much of the installed base worldwide, exhibit these phenomena, even when optimizing very simple queries. In the Confirmatory Experiment described in detail in Section 6.3, we started with 6,966 query instances (a query run on a specific DBMS) after our extensive query measurement protocol . While about 20%, 1,491, of these query instances contained only one plan, a few of the other query instances switched plans at almost every change in cardinality (we varied the cardinality in increments of 10K tuples, a total of 200 cardinalities): see Figure 2. Slightly over half, 4,093, exhibited suboptimality somewhere along those 200 cardinalities; a few had many changes to a plan that was in fact suboptimal, as indicated in Figure 3, across all four DBMSes considered.

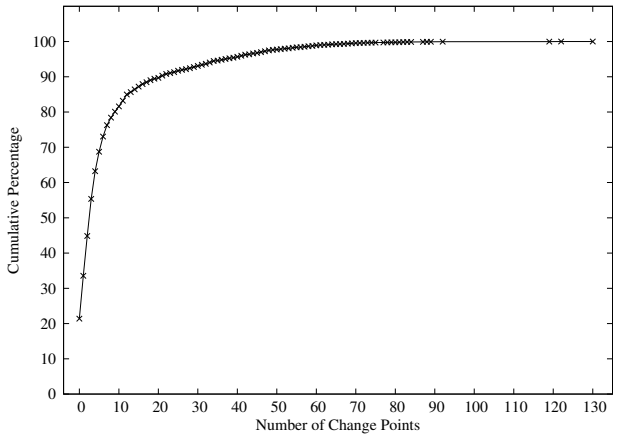

Fig. 2. Cumulative percentage of queries exhibiting the indicated number of plan changes

One oft-stated observation is that the role of query optimization is not to get the *best* plan, but rather to get a plan that is acceptably good. Figure 4 shows the cumulative distribution of the relative amount of suboptimality (where an x-value of 100 denotes that the query ran 100% slower than the optimal query, that is, twice as long). The good news is that 2,738 query instances, or 67% of the suboptimal queries, exhibited only a small degree of suboptimality: less than 30%. The challenge is that over fifth of all queries (1,355) exhibited a significant amount of suboptimality ($>= 30\%$); the relative slowdown can be quite large for some queries.

We started with 7,640 query instances (a particular query running on a particular DBMS, cf. Experiment 7 of Table I in Section 6.3. After our protocol, we were left with 6,966 query instances, of which 5,475 had at least one change point, so those are the ones we consider further. Of those, 1,382 had *no* suboptimality, so 4,093 (75%) had some suboptimality.

1,951 (36% of the query instances with a change point) have the higher cardinality running at least 20% faster than the lower cardinality. That means that 2,142 query instances (slightly over half of the suboptimal queries) were barely suboptimal (<20%

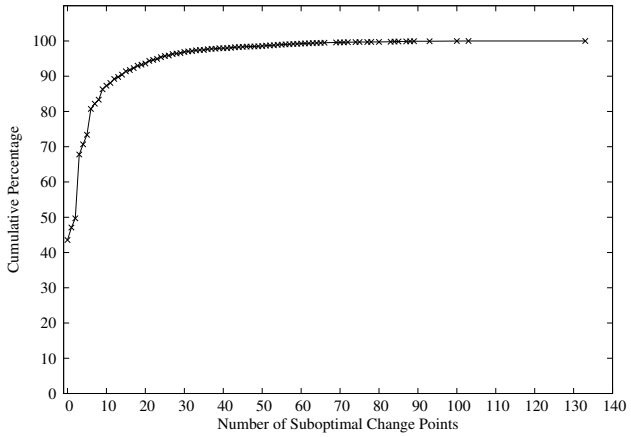

Fig. 3. Cumulative percentage of queries exhibiting the indicated number of changes to a *sub-optimal* plan

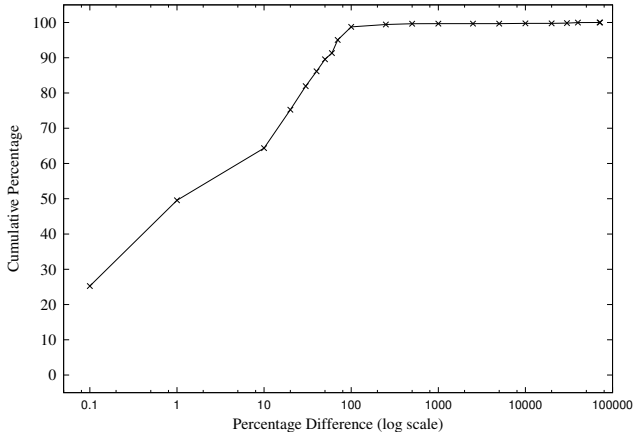

Fig. 4. Cumulative percentage of queries exhibiting the indicated percentage relative suboptimality

slower) and about a third (1,355) were considerably suboptimal (where the higher cardinality ran at least 30% faster than the lower cardinality).

We emphasize three important points. First, we used a sophisticated query measurement methodology that reduces the measurement variance, so that the query plans we identify as suboptimal definitely are so. Second, these results are over four DBMSes, and thus, such phenomena are not dependent on a particular implementation of cost-based optimization. Rather, they seem to be common to *any* cost-based optimizer, independent of the specific cardinality estimation or plan costing or plan enumeration algorithm or implementation. Third, suboptimal plans are *not* the result of poor coding or of inadequate algorithms. We view query optimization in modern DBMSes as an engineering marvel, especially given the complexity of the SQL language and the requirements and expectations of DBMS users, who often demand that important queries simply not get slower with a new release of the DBMS. Rather, the prevalence of suboptimality observed here is a reflection of the complexity of the task of query optimization.

This is where the shift to a new methodology comes in. We want to understand cost-based optimization deeply, in this case to determine if a fundamental limitation

exists, which means that we need to study multiple instances of that optimization architecture, to achieve generalizable results. To do so, we will articulate a predictive causal model for how and in what circumstances suboptimality arises and will provide compelling evidence via hypothesis testing that this model characterizes the behavior of query optimizers in general. This is what is meant by "empirical generalization" and why it is needed to answer such questions. In the long history of research in database query optimization, or even of databases in general, our model and its hypothesis tests are the first predictive results that we are aware of that apply *across* DBMSes, rather than on a single, specific DBMS or on a specific algorithm. This is due to the adoption of a new research methodology, that of empirical generalization. Such a causal model can provide guidance to the community about where fundamental research is needed and to the DBMS engineers about where to focus their efforts.

3. RELATED WORK

SQL has emerged as the *de facto* and *de jure* standard relational database query and update language. It emerged as an ANSI and ISO standard in 1986–7, with a major extension and refinement as SQL-92 [19] and an even larger extension as SQL:1999 [18] and refinement as SQL:2008 [14] and SQL:2011.

There has been extensive work in query optimization over the last 40 years [12; 15]. Query optimization and evaluation proceeds in several general steps [22]. First, the SQL query is translated into alternative query evaluation plans based on the relational algebra via *query enumeration*. The cost of each plan is then estimated and the plan with the lowest estimated cost is chosen. These steps comprise query optimization, specifically *cost-based query optimization* [24]. The selected query plan is then evaluated by the query execution engine which implements a set of physical operators, often several for each logical operator such as join [7].

An influential survey [4] does a superb job of capturing the major themes that have pursued in the hundreds of articles published on this general topic. Chaudhuri reviews the many techniques and approaches that have been developed to represent the query plans, to enumerate equivalent query plans, to handle some of the more complex lexical constructs of SQL, to statistically summarize the base data, and to compute the cost of evaluation plans. He also mentions some of the theoretical work (which is much less prevalent) to understand the computational complexity of these algorithms. Most of this research may be classified as adopting an engineering perspective: how can we architect a query optimizer "where (1) the search space includes plans that have *low cost* (2) the costing technique is *accurate* (3) the enumeration algorithm is *efficient*. Each of these three tasks is nontrivial and that is why building a good optimizer is an enormous undertaking." [4, page 35]

To determine the best query access plan, the cost model estimates the execution time of each plan. There is a vibrant literature on this subject [13; 17], including proposals for histograms, sampling, and parametric methods. Again, most of these papers are engineering studies, providing new techniques that improve on the state-of-the-art through increased accuracy or performance. There have also been a few mathematical results, such as "the task of estimating distinct values is *provably* error prone, i.e., for any estimation scheme, there exists a database where the error is significant" [4].

An optimizer for a language like SQL must contend with a huge search space of complex queries. Its first objective must be *correctness*: that the resulting query evaluation plan produce the correct result for the query. A secondary but clearly very important objective is *efficiency*; after all, that is the raison d'etre for this phase. As is well known, the name for this phase is an exaggeration, as existing optimizers do not produce provably optimal plans. That said, the query optimizers of prominent DBMSes generally do a superb job of producing the best query evaluation plan for most queries. This per-

formance is the result of a fruitful collaboration between the research community and developers.

Early investigation of plan suboptimality resulted in approaches such as dynamic query-reoptimization [1; 16], which exploit more accurate runtime statistics that appear while a query is being executed, to steer in-flight plan reoptimization. The very presence of such a radical change to the normal optimize-execute sequence indicates that plan suboptimality was of interest to some researchers.

However, even with great effort over decades, optimizers as a general class are still poorly understood. As has been observed, "query optimization has acquired the dubious reputation of being something of a black art" [2]. DeWitt has gone farther, stating that "query optimizers [do] a terrible job of producing reliable, good plans [for complex queries] without a lot of hand tuning" [25, page 59]. And as we will see, suboptimality is possible even when considering only simple queries.

While this paper does not provide direct solutions to address suboptimality, we envision that by adopting our proposed predictive model for suboptimality, engineering practice, such as dynamic reoptimization just mentioned may benefit. We elaborate on this subject in Sections 6.8 and 8, where we discuss the engineering implications of our causal model.

4. A MODEL OF SUBOPTIMALITY

The purpose of query optimization is to generate optimal plans. So why would suboptimality occur in the first place? Query optimizers are highly complex, comprised of tens or hundreds of thousands of lines of code. There are several reasons for this complexity. First, an optimizer must contend with the richness of the SQL language, whose definition requires about 2000 pages [14], with a multiple of linguistic features. Second, an optimizer must contend with the richness of the physical operators available to it. DBMSes have a range of algorithms available to evaluate each of many algebraic operators. Third, the optimizer must contend with an exponential number of query evaluation plans. Kabra and DeWitt [16] identify several other sources of complexity: inaccurate statistics on the underlying tables and insufficient information about the runtime system: "amount of available resources (especially memory), the load on the system, and the values of host language variables." (page 106). They also mention user-defined data types, methods, and operators allowed by the newer object-relational systems [18]. Thus, the task of optimization is very complex, with the result that the optimizers themselves consist of a collection of "components," that is, the rules or heuristics that it uses during optimization, with each of these components being itself complex.

We wish to understand the causal factors of suboptimality, through a predictive model that explicitly states the interactions between these causal factors. We test this model through experiments over tens of thousand of queries and hundreds of thousands of query executions, showing that there is strong support for this model. We then extract engineering implications from the model, suggestions for the most productive places to look to reduce suboptimality and thus to improve existing query optimizers.

Here we examine the hypothesized influence that each independent variable will have on the one dependent variable, query suboptimality (with some of the the influences mediated by one of the constructs). In the next section we will operationalize these variables, explaining how each is controlled or measured.

The model concerns five general constructs that we hypothesize will play a role in suboptimality: optimizer complexity, schema complexity, query complexity, data complexity, and plan space complexity. Two constructs include several specific variables that contribute to that construct as a whole. Our model distills many of the widely-

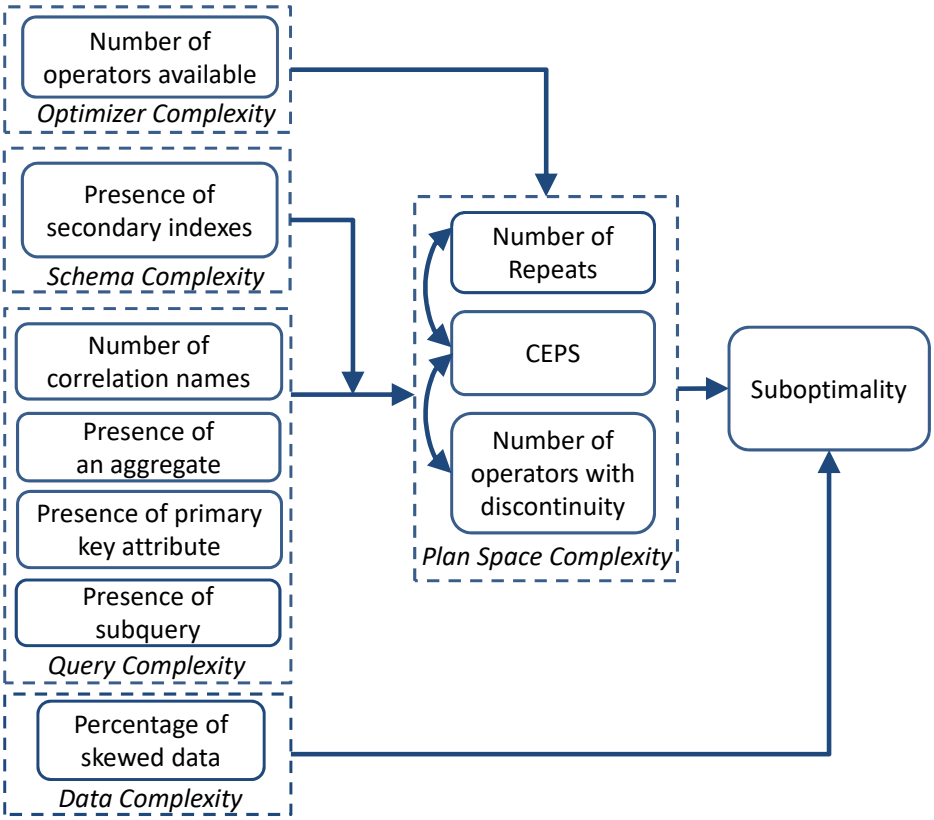

Fig. 5. Predictive Model of Suboptimality

held assumptions about query optimization. Our contribution is the specific structure of the model and the specific operationalization of the factors included in the model.

Our theory, which we will investigate in some detail and which is the basis for our predictive model, is that suboptimality is due in part to the complexity of the optimizer and concomitant unanticipated interactions between the different components used within the optimizer with various sources of complexity in producing a good (hopefully optimal) plan. We argue that with the proliferation of concerns and variability that an optimizer must contend with, it is extremely difficult to ensure that for an arbitrary query that the optimizer will *always* pick the right plan. There are many sources of complexity that may challenge one or more components of the optimizer; our theory implicates specific sources as contributing to observed suboptimal plans.

The model depicted in Figure 5 suggests specific factors that may impact the prevalence of suboptimality. This model has one dependent variable, *suboptimality*, on the far right. We observe this dependent variable in our experiments by determining whether each particular query, run on a particular DBMS and using a particular schema, is anywhere suboptimal. In Section 5, we delve into the details of how we operationalize this and the other variables we now examine.

The model has several independent variables which influence suboptimality.

For the construct of *optimizer complexity* we have one independent variable, "Number of operators available (in the DBMS)." We can manipulate this variable by our choice of DBMS: each DBMS has a set of operators available to its query evaluator and available to the query optimizer to use with in query plans.

We have one variable for the construct of *schema complexity*: "Presence of secondary indexes." If true, that means that every non-primary-key attribute, for the variable table as well as all the fixed tables, has a secondary index associated with it. The rationale is the the presence of secondary indexes expands the number of possible plans: each predicate in the query can often be mapped to one or more operators utilizing that index.

For the construct of *query complexity* we have identified four variables: "Number of correlation names", "Presence of an aggregate", "Presence of primary key attribute", and "Presence of subqueries". For each independent variable, we have interventional control in our experiments, in that we can manipulate the values of these variables through the construction of the actual query to be optimized. The first is the number of correlation names defined in the FROM clauses. The second is whether a single aggregate operator (that is, either `Max` or `Avg`) appears in the SELECT clause. The third is whether a primary key attribute appears in at least one predicate in the WHERE clause. The fourth is the number of subqueries that appear in the WHERE clause (these subqueries evaluate to a single value that is equality-compared with an attribute, in the where clause). The rationale is that each may expand the number of plans or otherwise renders the search for the optimal plan more complex.

We have one variable for the construct of *data complexity*: that of "Percentage of skewed data". Skew is generally defined as how values are distributed with a column of a table. We refine this to "how many duplicate values are present," with zero skew meaning that there are no duplicate values present and 100% skew implies that there is but one value in the entire column. The presence of skew complicates query time estimation, which in turn complicates the search for the optimal plan.

In the middle of the figure is the construct of *plan space complexity*. Given a particular query, its measurable complexity will impact the total number of candidate plans considered by the optimizer. However, this latter factor is not directly observable, again, especially for proprietary systems. However, we *can* measure the number of plans actually generated by the optimizer when presented with different cardinalities of the underlying tables. We term this set of plans the "effective plan space" and the number of such plans, the cardinality of the effective plan space, or "*CEPS*". Similarly, we cannot directly observe the cost model utilized by the optimizer for the operators it considers, but can classify the cost model of each operator as continuous (smoothly varying with cardinality of its input(s) and without jumps) or discontinuous (sometimes producing observable jumps in the predicted cost), and can thus count "the number of operators with discontinuity." (Such operators have also been termed "nonlinear" [11], though we focus on a more specific aspect of discontinuity.) Finally, the variable "Number of Repeats" is the number of times a plan is reused across the cardinality of the variable table. For example, if the sequence of plans for a query as the cardinality increases is $A\ B\ B\ B\ C\ C\ A\ B\ B\ B\ C\ B\ C\ C\ C\ B$, then the "Number of Repeats" will be six: removing sequential duplicates results in $A\ B\ C\ A\ B\ C\ B\ C\ B$, with the last six distinct plans being duplicates of the first three. Interestingly, CEPS plus the number of repeats gives you the number of plans in the sequence, again, after removing sequential duplicates, in this case, nine.

Plan space complexity is an *intervening* variable, in that it is dependent on some of the variables on its left but is observable and some exerts influence on the dependent variable on its right. We can observe these variables within experiments and indirectly influence their value but cannot directly intervene. For example, we can influence the

CEPS through manipulating the values for the independent variables of optimizer and query complexity, but cannot directly specify a value for CEPS within an experiment.

Given these six constructs and eleven specific variables depicted in Figure 5, let's now examine the causal relationships between these variables. Such relationships are specific *interactions* between the the constructs (that is, their associated variables), as hypothesized by this predictive causal model.

One causal factor of this model is the optimizer complexity. We hypothesize that optimizer complexity has influence over suboptimality indirectly, via plan space complexity. We hypothesize that an optimizer with a larger number of available operators will generate more plans and hence increase plan space complexity.

Hypothesis 1: Number of operators available will be positively correlated with, (a) Number of Repeats, (b) CEPS, and (c) Number of operators with discontinuity.

We now turn to query complexity, a construct associated with four independent variables. As with the optimizer complexity construct, we include in our model an indirect interaction through plan space complexity.

A high value of each of these specific variables implies a more complex query. We expect a strong relationship between number of correlation names to CEPS (because it is well-known that the number of potential join combinations is exponential to number of correlation names), to number of repeats (number of repeats could partially track CEPS), and number of operators with discontinuity (for the same reason).

TODO: *Rick: Please review wording.*

Hypothesis 2: Number of correlation names will be strongly correlated with (a) Number of Repeats, (b) CEPS, and (c) Number of operators with discontinuity. This holds similar with Presence of an aggregate (correlations d–f), Presence of a primary key attribute (correlation f–i), and Presence of subqueries (correlations j–l).

We don't feel that skewed data will impact plan space complexity, but rather that it will negatively impact accuracy of plan cost estimation, and thus increase suboptimality.

Hypothesis 3: Percentage of skewed data will be negatively correlated with Suboptimality.

Let's now turn to plan space complexity. We hypothesize a positive correlation between the two variables associated with this construct. As the number of plans considered by the optimizer increases, so should CEPS, which could increase the number of operators with discontinuity that is observed.

The model also has two interactions within plan space complexity, between Number of repeats and CEPS and between CEPS and Number of operators with discontinuity variables.

Hypothesis 4: Plan space complexity (CEPS) and Number of operators with discontinuity will be positively correlated.

We hypothesize that greater plan space complexity should make it more difficult to optimize the query. It is important to note that the occurrences of sub-optimality is not necessarily dependent on the presence of a discontinuous plan operator. However, we predict that when discontinuous plan operators appear in candidate plans, they may introduce complexity to query optimization, especially at the cardinality where discontinuity is possible. This is because the performance of such operators is sensitive

to the input size in a rather complex way. Any inaccuracy in the estimation of plan statistics may lead the optimizer to select a suboptimal plan.

Hypothesis 5: Presence of secondary indexes keys will strengthen the correlations between Number of operators available and the Plan Space Complexity, specifically, Number of repeats, CEPS, and Number of operators with discontinuity.

A plan chosen by the optimizer once may be reconsidered next time, because of the past experience of selecting and using that plan among many other plans. Even if many different plans are already used, for the same reason it may be easier for the optimizer to revisit a pool of the previously used plans and choose one of them than to explore other new plans, given the complexity of plan cost estimation. Then the larger CEPS, the more times the optimizer repeats using the same plans. We thus hypothesize that the number of repeats has a positive correlation with CEPS.

The schema complexity construct consists of the variable "presence of secondary indexes." Such indexes provide opportunities for the optimizer to consider more candidate plans that use these indexes. Those additional plans enable the optimizer to possibly do a better job, while also adding complexity to the optimization process. We hypothesize that this factor has a more complex role in the model, serving as a *moderator* of two interactions previously introduced. We hypothesize that the overall effect of the secondary indexes is to reduce the strength of the interaction between optimizer and query complexity to plan space complexity.

Hypothesis 6: The Number of Repeats will positively correlate with CEPS.

Hypothesis 7: Suboptimality will be positively correlated with Plan space complexity, that is, (a) Number of Repeats, (b) CEPS, and (c) Number of operators with discontinuity.

The Schema Complexity construct consists of the variable "presence of secondary indexes." Such indexes provide opportunities for the optimizer to consider more candidate plans that use these indexes. Those additional plans enable the optimizer to possibly do a better job, while also adding complexity to the optimization process. We hypothesize that this factor has a more complex role in the model, serving as a *moderator* of a mediating interaction previously introduced. We hypothesize that the overall effect of Presence of secondary indexes is to reduce the strength of the interaction between Query Complexity and Plan Space Complexity.

We have just *described* how suboptimality might arise, through a theory and its elaborated causal model, which implies seven specific hypotheses. We now need to move to *prediction*. How might we test such a model?

The first step to test this model is to *operationalize* each variable. In the next section we describe explicitly how each variable is defined. For the independent variables, we must be able to intervene, that is, set their values before the experimental test commences. For the latent and dependent variables, we need to be able to measure their values during each experiment.

5. VARIABLE OPERATIONALIZATION

In this section we specify how we operationalized each of the seven variables in the model. Recall that each independent variable is a property of a DBMS (number of operators available), of the schema (presence of secondary indexes), of a query (number of correlation names, presence of an aggregate, presence of primary key attribute, and presence of subquery), of the data (percentage of skewed data), of the plan space (number of repeats, CEPS, and number of discontinuous operators). There is one dependent variable of our model (suboptimality).

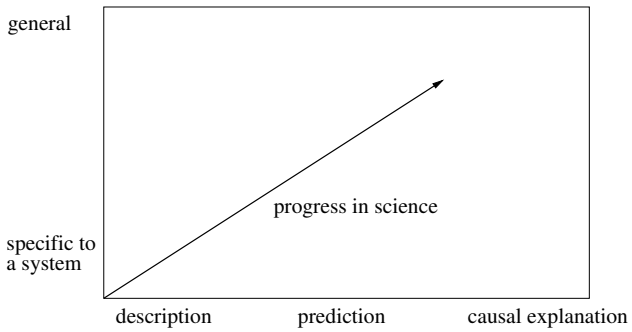

Fig. 6. Empirical Generalization

It is important to note that all manipulation must be done *outside* the DBMS. For proprietary systems we do not have access to the internal code. We do not know (*cannot* know) all the plans that were considered, nor the details of how the plans are selected. But such access is not needed; indeed, to be able to study a phenomenon across many DBMSes, such access is not practical. But by designing the experiment to examine the plans that each DBMS actually produces, and thus to examine phenomena that can be externally visible, we can obtain valuable insights into general classes of computational artifacts. Our ultimate goal is to make statements that hold across cost-based optimizers in general, thereby moving up the y-axis of Figure 6.

5.1. Optimizer Complexity

By "number of operators available in the DBMS" we mean the number of operators available for selection, projection, join and aggregate functions. Our experiments intervene on this variable by selecting a particular DBMS on which to evaluate each query. Across the available DBMSes, as the number of operators available increases, the complexity of the optimizer increases because it has to choose between more operators.

The `EXPLAIN PLAN` facility specifies the operators(s) employed in that plan. For each DBMS, we collect all the unique operators from all the plans and count the number of these distinct operators, each used by for least one query at at least one cardinality by that DBMS. The number of operators used by a DBMS ranged from 8 to 53 across all the queries and data sets used in the Exhaustive, Exhaustive with Keys, and Exploratory Experiments, to be discussed in Section 6.3.

5.2. Schema Complexity

The presence of secondary indexes is easy to operationalize. We generate two databases, one without any secondary indexes and one with a key specified for each table on the non-key (that is, other than the first) attribute.

5.3. Query Complexity

Query complexity is also relatively easy to operationalize. We randomly generate queries, such as the example presented in Section 2. Each query is a select-project-join-aggregate query, with a few attributes in the SELECT clause, a few tables referenced in the FROM clause, a few equality predicates in the WHERE clause, and zero or one aggregate functions in the SELECT clause. As such, some of the complexities mentioned by Kabra and DeWitt [16], such as user-defined data types, methods, and operators, are not considered. Concerning presence of primary key attribute, if this

independent variable was set to 1, we ensured that there was at least one primary key attribute in one of the comparisons in the WHERE condition.

The queries were generated by a simple algorithm. The SELECT clause will contain from one to four attributes, hence, an average of 2.5 attributes. The number of correlation names in the FROM clause varied from one to four, with duplication of tables allowed (duplicate table names within a FROM clause implies a self-join). In the queries that were generated, from one to four unique tables were mentioned in the FROM clause. Somewhat fewer tables were mentioned than the number of correlation names, as the presence of self-joins reduced the number of unique tables referenced by the queries.

For the query in Section 2, the number of correlation names is 3, the presence of an aggregate is false (0), the presence of primary key attribute is false (0), and the presence of subqueries is false (0).

We ensure that Cartesian products are eliminated We do this by connecting all the correlation names that appear in the FROM statement via equi-joins on random attributes. The comparisons are all equality operators. To ensure that the queries are as simple as possible, we do not include any additional predicates in the WHERE clause. This is realized by setting the attributes `maxIsAbsolute` to `true` and `complexUsePercentage` to `100`. Basically, "complex" predicates eliminate the Cartesian product, and by setting complex predicates as "absolute," no additional predicates are included except for those which are necessary for eliminating Cartesian product. Also for simplicity, we include neither disjunctions nor negations.

The presence of subquery was 0 or 1, with 0 indicating no subquery. For each query needing a subquery, we picked a separate generated query and rendered it as a subquery to replace an attribute in the where clause of the original generated query. As an example when this variable was one, we started with query qs1-2, shown here.

```
SELECT t0.id2, SUM(t1.id3)
FROM ft_HT2 t0, ft_HT2 t1
WHERE (t0.id2=t1.id1)
GROUP BY t0.id2
```

We then generated another simple select-project (SP) query at random concerning the variable table and simply replaced one side (in this case, the right side) with the generated entire query as a subquery, to produce this final query.

```
SELECT t0.id2, SUM(t1.id3)
FROM ft_HT2 t0, ft_HT2 t1
WHERE (t0.id2 IN (SELECT t2.id3 FROM ft_HT1 t2))
GROUP BY t0.id2
```

We thus have a maximum of only one level of nesting.

5.4. Data Complexity

This independent construct has one variable: percentage of skewed data.

Skew has a very specific definition in statistics, involving elongating the left or right tail of a distribution, thereby moving the mean left or right of the median (in a symmetric distribution the mean = median = mode).

But we start with a distribution without a tail: the uniform distribution: the values from 1 to 2 million (2M). We consider this to be a skew of 0 (no skew). At the other end of the spectrum is one in which all the values are identical, or a skew of 1.0.

We define the skew as the reciprocal of the number of distinct values. For 2M distinct values, the skew would be 1/2M, which is practically 0. For 1 distinct value, the skew

would be 1.0. For 2 distinct values, the skew would be 0.5. For 10 distinct values, the skew would be 0.1.

We can generate the table of 2M rows by generating values sequentially from 1 to the number of distinct values. This creates a "span of values." We repeat this for the second span if necessary, and on and on, until we have 2M values in all.

When varying the cardinality, we remove 10K values from the variable table and then copy those tuples to a new table to ensure that every page is as full as possible (that is, 100% load factor). This gives us a table of 1.99K tuples. (We then get a query plan for this table.) We repeat this removal process until the final cardinality reaches 10K.

The way we effect the 10K removal is as follows. The key idea is to remove individual spans until we've deleted 10K values. Since we don't touch the remaining spans, the number of distinct values does not change, and so the skew remains constant.

Let's first illustrate with a skew factor of 1.0. This translates to exactly 1 value for the entire table, or 2M spans. To shorten, we remove 10K spans, equal to 10K values. This removal keeps the skew at 1.0, due to the unique values in the remaining spans. At the end, 10K spans, each with a single value, remain.

What about a skew factor of 0.1? This skew factor translates to 10 distinct values. So we generate 200K spans, each with 10 values. To shorten, we delete 1,000 spans, which is equivalent to removing 10K values. As before, this does not change the skew. At the end, the table will have 1,000 spans with 10K values, retaining a skew factor of 0.1.

Let's see an example at the other end of the spectrum, of a skew factor of 1/2M. This skew factor, which is almost close to 0, gets translated to 2M distinct values. So we generate a single span of 2M values. To shorten the table, it makes no sense to remove this single span in its entirety. But we can remove 10K values from this single span. Note that this changes the skew to 1/1.99M, then eventually to 1/10K, which is still very close to zero (the skew has changed from .0000005 to .00001).

We use five skewness values: 0 (approximately), 0.001, 0.1, 0.5, and 1.0, for this independent variable.

5.5. Plan Space Complexity

This explanatory construct includes three variables.

As discussed in Section 4, the "cardinality of the effective plan space" (CEPS) is the number of plans selected as optimal for that query being evaluated on one of the 200 cardinalities for the variable table. It is "effective" because it was chosen, as opposed to all of the plans that were considered but not chosen. (Recall that for proprietary DBMSes, we do not have access to such plans.) Note that we count only distinct plans. As we saw in Figure 1, fluttering queries return to a previous plan. This particular query has a CEPS of 4.

The number of repeats is just the number of times a plan associated with a smaller cardinality is repeated, after removing sequential duplicates.

We also wanted to evaluate the contribution of cost model non-linearity to suboptimality. We found that it is possible to assemble, from outside the DBMS, an approximation of the cost formula for that operator, and thus directly observe whether it is non-linear. Specifically, the result of SQL's `EXPLAIN PLAN` (a result that is particular to each DBMS) includes the *estimated cost* of each stage of the plan. That cost (estimate) is a dependent variable, one that can be observed.

To operationalize the "number of operators with discontinuity" we classify each DBMS operator as either continuous or discontinuous, to be defined shortly. We then examine the queries to identify the plan change points, which provide the *effective plan space*, a set of plans chosen for one or more cardinalities of the variable table. For

each distinct query plan in the effective plan space, we count the number of discontinuous plan operators that appear in that plan. Some of these plans may contribute no such operators, some may contribute one such operator, and some may contribute several such operators. We then sum the counts of the distinct plans in the effective plan space, yielding an integer as the operationalization of "number of discontinuous operators" for each query. This count per query varied from 0 to 85 for the queries we tested in the confirmatory analysis.

To classify each operator as continuous or discontinuous, we use the data from the Exhaustive Experiment. We used a small subset of queries to be used later in testing the model, to attempt to observe the same operators as encountered in that study. We collected all query plans at all possible cardinalities (200 in all) and looked for "jumps," that is, when the cost model for an operator in the plan exhibited discontinuity. Jumps are determined from the estimated cost extracted from each plan operator. We provide an example of a jump below. If a jump is observed, we classify that particular plan operator as discontinuous.

A *jump* is a pair of close cardinalities (in our case, separated by 10K tuples) in which the query optimizer's cost model is *discontinuous*, that is, does not smoothly increase from the lower cardinality to the upper cardinality. To identify jumps in the first step, we examine the slopes between each pair of consecutive cardinalities (again, separated by 10K tuples). We expect that the slope (that is, the first derivative) is well-behaved and thus does not change much for consecutive cardinalities. A jump thus indicates a large deviation in the second derivative of the cost model across the two cardinalities.

The Exhaustive experiment gathers the cost model for each operator in each plan for each cardinality (in our case, for cardinalities ranging from 10K to 2M in steps of 10K, or 200 points). Figure 7 presents for a single query, the *cost* of each of the hash-join operators utilized in each plan at a particular cardinality. Each distinct point type depicts an individual hash-join operator. Note that when two identical query plans appear at two different cardinalities, we consider all the plan operators found at the same position within these two plans to be identical. This figure shows only a small portion of the 200 cardinalities, and is typical (we generally observed jumps at low cardinalities for these particular queries and relation sizes). As shown by this figure, the hash-join operator represented by '+' has two discontinuous jumps, both appearing below cardinality 150K. The other hash-join operator represented by $\times$ has one discontinuous jump. Between the jumps, the behavior is linear. It is our guess that such a jump is caused by the transition between one pass of a disk-based hashing technique to two passes (and indeed for one of the operators, perhaps the one lower in the operator tree, the transition to three passes, hence, the two jumps). That said, all that matters for our predictive model is that operators that experience such discontinuity in their cost models present opportunities for suboptimality.

To identify such jumps, we examine the second derivative (the change in the first derivative, that is, the change in the slope). A jump will be indicated by a larger than normal second derivative. By identifying a sudden change in the second derivative, we can effectively spot the cardinality at which an operator in a plan becomes discontinuous.

Formally, we compute the slope (S) of the *estimated* cost (of the cost model, formalized as C_{card}, where *card* is the input cardinality) between each pair of consecutive cardinalities as $S_{card} = (C_{card+10K} - C_{card})/10K$. This computes a series of slope values. We then compute the standard deviation of all the slope values. For example, examining Figure 7, the slopes are small except at three places, one for the green operator and two for the red operator. By identifying the slope values that are greater than one standard deviation over the average value, the discontinuous operators can

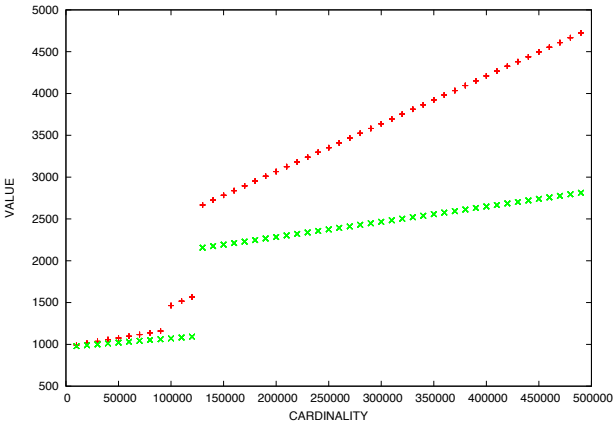

Fig. 7. An Example of Discontinuous Plan Operators (Hash-Join)

be identified: A "discontinuous operator" is one for which a jump is observed in that operator in one of the plans for at least one of the input queries.

5.6. Suboptimality

We now consider the one dependent variable. How might suboptimality be observed? We have developed a system, DBLAB, that allows us to perform experiments to study this phenomenon of suboptimality. DBLAB submits queries to the DBMS, while varying the cardinality of one of the tables, requesting in each case the evaluation plan chosen by the DBMS. This is done using the `EXPLAIN` SQL facility available in modern DBMSes. (The Picasso system also used this facility to visualize the plan space chosen by a DBMS optimizer [8; 9].) We can then compare the performance (execution time) of various plans for the query, to identify those situations when a suboptimal plan was chosen, when in fact there was a different plan that was semantically equivalent to the chosen plan (that is, yielded the identical result) but which ran faster.

We modify the cardinality to produce multiple execution plans for a given query. For one of the DBMSes, we can modify the stored table statistics directly. For the other DBMSes, we had to do so indirectly, by varying the size of the table and running the optimizer on tables of different size. As we vary the cardinality, we collects all the plans that the optimizer felt were appropriate for that query.

Our definition of suboptimality assumes that actual execution time for any query plan is *monotonic non-decreasing*, that is, unchanged or increasing as the cardinality increases. The intuitive justification is that at the higher cardinality, the plan *has* to do more work, in terms of CPU time and/or I/O time, to process the greater number of tuples. (For an operator having a discontinuous cost model, as the anticipated input cardinality grows, there will be *jumps*, in which the predicted cost is temporarily much greater. Section 5.5 explains further how we specifically operationalize this property. Also note that we don't consider SQL operators such as `EXCEPT` that are not monotonic.) We formalize this property as follows.

Definition: *Strict Monotonicity*: given a query Q and an actual cardinality a,
$$\forall p \in \mathit{plans}(p)\ \forall c > a\ (\mathit{time}(p, c) \geq \mathit{time}(p, a))$$

where *plans*(p) is the set of plans generated by the optimizer and *time*(p, c) is the execution time of plan p on the data set with the varying table at cardinality c. ■

Note that the comparison is with the same plan p, occurring at higher cardinalities.

To test this assumption, we ran an experiment that we term "Monotonicity." This experiment considered 60 queries, chosen from the pool of queries generated for testing suboptimality, and timed them for cardinalities from 10K to 2M tuples, in steps of 10K tuples (hence, we used 200 cardinalities), for each DBMS (for one DBMS that was very slow, we started with 30K tuples, as discussed in Section 6.3). We varied the cardinality of the variable table by starting with the maximum size, running the queries, then deleting 10K tuples and repeating. We performed an "`ANALYZE TABLE`" function to force the DBMS to update the table's statistics to be accurate before actually evaluating the query. **TODO:** *Young: we did this for all experiments, not just monotonicity, right?* We then examined every case where the same plan for a query was run at different cardinalities; that is, we compared two query times measured by `currentTimeMillis()` in JAVA per a pair of different cardinalities at which the same plan is observed during the study. We expected that as the cardinality decreased, runtime would monotonically decrease. However, due to the variance in query time measurement observed even when the cardinality was identical, we encountered spurior violations.

Assuming a normal distribution for our time measurements, 95% of the distribution falls between $-2 * \sigma$ and $+2 * \sigma$. Therefore, to statistically infer with a 95% confidence interval that a violation occurred, we relaxed our definition of monotonicity to the following.

Definition: *Non-Strict Monotonicity*: given a query Q, an actual cardinality C, and the standard deviation of the query executions for cardinality C as σ, Q is non-strict monotonic if $\forall p \in \textit{plans}(p)\ \forall c' > C$,

$$\textit{time}(p, c') \times (1 + \sigma_{c'}) \geq \textit{time}(p, C) \times (1 - \sigma_C)\ . \qquad \blacksquare$$

As we will see in Section 6.3, we observed only 3,347 violations (0.73%) of non-strict monotonicity, for the largest experiment (Experiment 7: Confirmatory) across all the DBMSes we studied, justifying our conclusion that the DBMSes under study are indeed monotonic.

We can now turn to suboptimality. Recall that the monotonicity test examines two adjacent Q@Cs for which the *same plan* is observed. To detect suboptimality, we look for adjacent Q@Cs with *different plans*, termed a "change point," mentioned earlier. We look for such change points where the computed query time at the *upper* cardinality is *smaller* than the computed query time at the *lower* cardinality. Say the lower cardinality used Plan *A* and the upper cardinality exhibited Plan *B*. Had the DBMS query optimizer selected Plan *B* for the lower cardinality, the query time would have been smaller than that for Plan *A*, which follows directly from the monotonicity assumption. The conclusion is that for the lower cardinality, the optimizer picked the less efficient plan, and thus, this query exhibits suboptimality. Note that since this approach cannot consider plans that were never chosen, it very likely misses some suboptimal plans (for which there was a better plan not seen), and thus produces a conservative estimate of suboptimality.

Our definition of suboptimality compares the computed runtimes and standard deviations at the cardinality just before the change point (designated as $n-1$) and at the change point (that is, n).

The query is said to be suboptimal if $\textit{time}_{n-1} - 0.5 \cdot \textit{stddev}_{n-1} \geq \textit{time}_n + 0.5 \cdot \textit{stddev}_n$. Suboptimality is coded as four levels (0–3), based on the distance in half standard deviations, up to 1.5 standard deviations. We then sum this value over all pairs of cardinalities with different plans (the change points) for the query.

TODO: *Sabah: this paragraph talks about a suboptimality of up to 78, but the above papra says there are j*

Note that while DBLAB examines the plan at every cardinality, it only has to actually execute the query at pairs of cardinality that are change points. Since many fewer Q@Cs were involved, we could try many more queries than the Exhaustive Experiment. In the Exploratory Experiment to be described in Section 6.3, the value of suboptimality ranged from 0 (no suboptimality) to 78, with the majority between 0 and 9. A full 51% of the queries were suboptimal. Because the occurrence of large values of this measure was so rare, we did a log transformation: $\log_{10}(1 + \textit{subopt})$ in the confirmatory analysis.

6. TESTING THE CAUSAL MODEL

6.1. Experimental Setup

The measurements were collected using Tucson Timing Protocol Version 1 (TTPv1) [5] and Version 2 (TTPv2) [6] on a suite of five machines, each an Intel Core i7-870 Lynnfield 2.93GHz quad-core processor on a LGA 1156 95W motherboard with 4GB of DDR3 1333 dual-channel memory and Western Digital Caviar Black 1TB 7200rpm SATA hard drive, running Red Hat Enterprise Linux Server release 5.8 (Tikanga) for TTPv1 and release 6.4 (Santiago) for TTPv2, with a kernel of 2.6.32-358.18.1. The protocol provided calculated query evaluation time, including computation and I/O time, in msec. Both protocols were utilized exactly as specified.

No runs violated the experiment-wide sanity checks. For the largest experiment, the Confirmatory experiment 7 discussed below, approximately 10.5% of the query exectuions and 5.1% of the queryies-at-cardinality (Q@Cs) were dropped due to query execution and Q@C sanity checks. As a result, excessive variation in calculated query time was observed in only 0.003% of the Q@Cs. A total of 0.35% of Q@C adjacent pairs violated relaxed monotonicity and 0.68% of the Q@Cs violated strict monotonicity, which is acceptable.

We thank the developers of this protocol and of the software that enabled the running of many, many queries for providing us this software.

In these experiments, we installed each disk directly on the machine that also runs the DBMS, ensured a cold cache (disk drive, disk controller, O/S, and DBMS buffer), and discarded any sequence of query executions that appear to be the result of query result caching. We also ensured within-run plan repeatability.

In the following, we describe in detail the data used by our experiments and the experimental scenarios we defined.

6.2. Data sets

We generate our experiment data set randomly in the experiments. However, we use seeds to control the random data generator so that it can produce repeatable data as required. (Randomly generated data will probably be easier for the optimizer to deal with than skewed data.)

Our experiment data set consists of relational tables. There are two types of tables. The first is a "fixed table" that, once created and populated, will never be modified in the future. In contrast, the second type is a "variable table." We alter the cardinality, physically, of such tables as the experiments are being performed. We organized three configurations for the data sets. The first configuration is a small data set with four tables, each with four integer-typed attributes. Each table is populated with one million rows. The size of each table is roughly 16Mbyte (=4 (bytes) $\times 4(columns) \times$1,000,000 (rows)). We also produced

Table I. Experiments 1–7: Detailed Run Statistics

	Experiment	*Protocol*	*Cumulative Hours*	*Number of Query Instances*	*Number of Q@Cs*	*Number of QEs*	*Number of Retained QEs*
1	Monotonicity	—	38	60	12,000	12,000	12,000
2	Initial Exhaustive	TTPv1	1,672	160	32,000	320,000	244,787
3	Exhaustive	TTPv2	1,544	160	32,000	320,000	319,980
4	Exhaustive with Keys	—	28	200	40,000	—	—
5	Initial Exploratory	TTPv1	560	780	8,842	88,420	68,891
6	Exploratory	TTPv2	1,663	1,200	12,560	125,600	114,377
7	Confirmatory	TTPv2	11,375	7,640	99,558	995,580	890,631
	Total		16,880	10,200	236,960	1,861,600	1,650,666

a version of the data set with primary keys (of the first attribute) for all tables. **TODO:** *Young: is the following true" I don't think so...: Second, to perform queries with more relations, we cre*

6.3. The Experiments

We are interested in predicting the behavior of DBMSes through our model. We selected four relational DBMSes, some open source and some proprietary, that are representative of the relational DBMS market. Each was used in its stock configuration.

In each experiment, we varied the cardinality from 2M (maximum) to 10K (minimum), in increments of 10K. For the one DBMS that was slower than the others and was timing out for the majority of the queries when run between 10K and 2M, we reduced the size of the tables and varied the cardinality from 60K (maximum) to 300 (minimum), in increments of 300.

Utilizing JDBC to manipulate independent variables from outside the DBMS allows us to empirically generalize by moving up the y axis of Figure 6, from one system to several systems and then to a general theory. We don't reveal the identity of the DBMS we studied, for two reasons. First, commercial DBMSes include in their user agreements requirements not to release performance data. This is detrimental to science, but we have no choice but to live with that restriction. However, in some sense the specific DBMS doesn't matter, as we are studying phenomena about cost-based optimizers *in general*, and so are interested in making statements that apply to all the experimental subjects in our study.

Table I exhibits the statistics of running a number of queries in our experiments. We performed six separate experiments, each looking at a different aspect. It is important to emphasize that while the *queries* all came from the same query pool and while the data sets were also shared by the experiments, the *query executions* for the five experiments are disjoint. As a side comment, we mention that for all four DBMSes, the query plans generated by a DBMS for a particular Q@C of a particular query varied between the experiments, but not between the query executions (QEs) of that query instance at that cardinality, by virtue of the way we designed the measurement protocol.

The first experiment, termed "Monotonicity," was described in Section 5.6. This experiment ran quickly, as it only involved 12,000 QEs. That experiment helped us realize that we needed to be much more sophisticated in our approach to timing queries.

When TTPv1 became available, we performed our second experiment, termed "Initial Exhaustive," which more accurately tested the monotonicity assumption, as also described in Section 5.6. This experiment involved 160 query instances, 32,000 Q@Cs, and thus 320,000 QEs, requiring 1,672 cumulative hours. The (very small) percentage of strict monotonicity violations observed was consistent with the remaining variance of the query time measurement, concluding that none of the DBMSes violated monotonicity. This also provides a validation of our definition of suboptimality, which requires monotonicity.

We used the plans generated from Initial Exhaustive to classify operators as continuous or discontinuous, as discussed in detail in Section 5.5. Again we note that this experiment used a subset of the set of *queries* (but *not* query executions) from the query pool, to ensure that we see the same operators as encountered in that study.

We later reran this experiment using TTPv2 (we term this simply "Exhaustive"). Note that the number of retained QEs went up a lot (that is the main benefit of TTPv2), while the number of hourse required actually went down (another benefit).

We also ran an experiment (number 3 in the table) on the Exhaustive query set but using the data set with primary keys, termed "Exhaustive with Keys." However, in this case we did not actually execute the queries, but rather just examined the plans that were returned from the DBMS to identify additional discontinuous operators.

The fourth experiment was for initial exploratory analysis of a prior version of the causal model. (This fourth experiment used the first version of the Tucson Timing Protocol (TTPv1) [5].) The earlier model had fewer independent variables yet also had more complex interactions. Specifically, the model did not include the independent query complexity variables of presence of secondary indexes, presence of subquery, nor the schema complexity independent variable of presence of secondary indexes, nor the plan space complexity independent variable of number of repeats. The prior model also had presence of primary keys in the schema complexity construct rather than the query complexity construct. In reformulating that independent variable, we were able to remove a complex mediating moderator between the plan space complexity and suboptimality constructs.

The fifth experiment was an "Initial Exploratory" analysis of the causal model, run on a representative sample of queries and data sets: (a) 600 queries of one DBMS, (b) 120 queries of another DBMS, (c) 10 queries from each other two DBMSes, all without primary keys defined, and (d) 10 queries from each of the four DBMSes on the primary key data set, for exploration across all combinations, a total of 780 query instances. These query We ran the DBMSes on the *change points*: consecutive cardinalities (10K apart; 300 for one DBMS) with different plans. The protocol retained about 78% of the QEs.

We later reran the exploratory analysis using TTPv2, this time on a larger sample of queries and data sets: (a) 200 queries of four DBMSes without primary keys defined and (b) 100 queries from each of the four DBMSes on the primary key dataset, for exploration across all combinations, a total of 1,200 query instances (= 300 $\times$ 4 (DBMSes)). We termed this sixth experiment "Exploratory." We ran the DBMSes on the *change points*: consecutive cardinalities (10K apart; 300 for one DBMS) with different plans. This experiment required 1,663 hours. The experimental methodology retaining 12,100 Q@Cs (3.7% were dropped), covering 1,123 query instances (6.4% were dropped). Only 443 strict monotonicity violations (0.78%) and 301 relaxed ones (0.54%) were observed. This exploratory analysis allowed us to refine the operationalizations.

The seventh and final experiment was used for confirmatory analysis of the model and thus was called "Confirmatory." This was the most time-consuming of the experiments. Here we used (a) 800 queries on the data set without primary keys defined and (b) 510 of those queries that had joins on the primary key attributes, on the data set with primary keys, (c) 100 queries on the data set without skew and primary keys, (d) 100 queries with a subquery on the data set without primary keys, (e) 100 queries on the data set with primary keys and secondary indexes defined, (f) 100 queries with a subquery on the data set with primary keys defined, (e) 100 queries with a subquery on the data set with primary keys and secondary indexes defined, and (g) 100 queries on the data set with primary keys and secondary indexes defined, for a total of 1,910 queries and a total of 7,640 query instances (over the four DBMSes). Again, we ran the DBMSes at the 99,558 Q@Cs that were observed, roughly 13

per query. **TODO:** *Young, where did this number come from?:* The protocol accurately timed 53,547 Q@C pairs as change points.

In this Confirmatory Experiment, we observed 3,347 strict monotonicity violations (0.74%), and 1,966 relaxed monotonic violations (0.43%) (out of a total of 452,684 pairs), which provides further confidence that monotonicity also applies to operators found in queries over data in various contexts (as described above) and that our operationalization of suboptimality is a valid one.

In the remainder of this section, we focus using the measured independent, latent, and dependent variables in the Confirmatory experiment to test predictions that arise out of our causal model in Figure 5.

6.4. Descriptive Statistics

Several initial conclusions can be drawn from this confirmatory experiment, which was the result of several years of programming effort and about 30 months of experimental runs. First, *every* DBMS exhibits suboptimality. Thus, this phenomenon is likely to be a fundamental aspect of either the algorithm (cost-based optimization) or the creator of the algorithm (human information processing). Our model includes both effects. Perhaps surprisingly, more than half (3,933 queries out of the 6,983 query instances that emerged from the measurement protocol) exhibited suboptimality somewhere in the range of cardinality of the varying table. Most instances of those queries (3505) had suboptimality at least at level 3. CEPS ranged from 1 to 24, plans across the cardinality range. The number of discontinuous operators observed in plans for a query averaged 9.4, with 85 being the maximum. Number of Repeats ranged from 0 to 108, with the mean being 4.99.

6.5. Correlational Analysis

We tested Hypotheses 1–7 using the strength and significance of correlations of variables involved. These hypotheses predicts eleven positive relationships. Table II lists the hypotheses followed by the correlation observed when testing each hypothesis. ("NS" denotes not significant at the 0.05 level, the accepted standard for significance. "—" denotes no prediction arising from the model.) As can be seen, most of the predictions (nine) arising from the causal model are supported and significant. The two exceptions are Hypotheses 1 and 2. Hypothesis 1 was not supported because the correlation between number of operators available and suboptimality was negative, while we predicted positive correlation. Hypothesis 2 was not significant.

6.6. Regression Analysis

TODO: *Sabah: revisit this entire section, as we changed our model* We ran a regression over the independent variables of the model that predict suboptimality over the data from the Confirmatory experiment. Our model explained 34% of the variance of the suboptimality dependent variable.

We also did regressions on the causal variables for CEPS, number of repeats and for number of operators with discontinuity. Our model explained 31% of the variance for CEPS, % of the variance for number of repeats, and 35% of the variance for discontinuity.

Hypothesis 8 predicts that the presence of secondary indexes will be a moderator, affecting the interactions between query complexity with plan space complexity. It thus provides predictions that the strength of such interactions will decrease in the presence of secondary keys. When secondary indexes were not specified, our model explains % of the variance of CEPS, and % of the variance of discontinuity, with all of the independent variables. When secondary indexes were defined on the underlying

Table II. Testing Hypotheses 1–7: Correlations on the Confiratory Study

Variable	*Suboptimality*	*Repeats*	*CEPS*	*Discontinuity*
Operators in DBMS	—	**H1a:**−0.16	**H1b:**	**H1c:**
Correlation names		**H2a:** 0.54	**H2b:**	**H2c:** 0.49
Skew	**H3:**	—	—	—
Presence of an aggregate		**H:** 0.02	**H:** 0.10	**H:** 0.33
Repeats	**H7a:**	—	**H6:**	—
CEPS	**H7b:** 0.48	—	—	**H4:**
Discontinuity	**H7c:** 0.31	—	—	—

Table III. Testing Hypothesis 8: Interaction Strength on the Confirmatory Study

Variable	*Suboptimality*		*CEPS*		*Discontinuity*	
	Not PK	PK	Not PK	PK	Not PK	PK
Correlation names	0.44	0.41	0.56	0.52	0.53	0.41
Presence of an aggregate	0.01	0.02	0.07	0.15	0.32	0.35
CEPS	0.54	0.51	—	—	—	—
Discontinuity	0.41	0.40	—	—	—	—

tables, our model explains % of the variance of suboptimality, 28% of the variance of CEPS, and % of the variance of discontinuity. These findings are all consistent with our hypothesis predicting a positive moderation effect of primary keys.

TODO: *Sabah: revisit. Not sure about the purpose of this paragraph* We also examined the individual contributions to each of these three variables (e.g., suboptimality) from their causal variables (in the case of suboptimality, from the four variables of correlation names, presence of an aggregate, CEPS, and discontinuity). Table III shows that the strength of all interactions goes down as predicted, except for ones involving aggregates. Recall that the predictive model indicates weaker effects generally for aggregates, as they are evaluated late in the query, and the primary key attributes are not necessarily included in the grouping attributes.

TODO: *Sabah: redo hypotheses in these tables, for example, H1 doesn't go to subopt, then add the confirma*

6.7. Summary of Model Testing

In our experimental design, we started with a structural causal model that encapsulates our theory for how cost-based query optimizers might select a suboptimal plan for a query at a cardinality. This model implies the eight specific hypotheses listed in Section 4. We then performed four experiments to refine our operationalizations (such as discontinuity, via the Exhaustive Experiment and number of operators available with the additional Exhaustive with Keys Experiment), to test fundamental assumptions (such as monotonicity, via the Monotonicity Experiment), and to test and make minor refinements to our model (via the Exploratory Experiment). While exploring the data, one must be cognizant of the possibility of Type 1 errors: false positives that lead one to believe a relationship exists when it doesn't.

To control for Type 1 errors, we then performed the Confirmatory Experiment on a completely different data set consisting of many more query instances, 7,640 in all, running on four DBMSes that each utilize cost-based query optimization, to test our refined model. Statistical inference is only possible in confirmatory analysis, where the model and hypotheses are selected a priori.

Correlational analysis provides significant support for the model, except for Hypotheses 1 and 2, with the implication that the influence of the number of operators available remains largely unexplored. Via regression analysis, the model explains 34% of the variance of suboptimality (overall: with primary keys a little lower and without primary keys a little higher due to that moderating effect), 31% of CEPS, and 35% of discontinuity. In all but one case, the causal influence of independent and latent vari-

ables is significant ($p < 0.05$). The direction of the regression coefficients, again in all but one case as predicted, also provides strong support for the model.

Now that the causal model has been found to be supported by the confirmatory analysis, we turn to possible implications of this model.

6.8. Identifying Root Causes of Suboptimality

Our goal in this paper has been to understand cost-based query optimizers as a *general* class of computational artifacts and to articulate and test a predictive model characterizing how such optimizers, again, as a general class, behave. This model can be used to further improve DBMSes through engineering efforts that benefit from the fundamental understanding that the scientific perspective can provide.

Our model includes three causal factors of suboptimality: DBMS optimizer complexity, query complexity, and plan space complexity. Of these factors, the regression coefficient that was highest was for CEPS (cf. Table III: 0.51–0.54, normalized) The next highest regression factor was number of tables involved, which is highly correlated with CEPS (cf. Table II: 0.54). These two observations imply that the number of plans being considered is a major determinate of suboptimality, indicating that choosing among many plans is hard, despite many decades of research and development.

The next most influential factor is the number of discontinuous operators (cf. Table III: 0.40–0.41, normalized), implicating the cost model.

The one following that is the number of operators provided by the DBMS. This factor is totally correlated with the DBMS, and so is confounded with other specificities of the coding of each DBMS. That said, it is still the case that as the number of operators goes up, across all the DBMSes studied, the amount of suboptimality increases. This factor also strongly impacts CEPS, and so indirectly impacts suboptimality through that path.

These factors implicate two broad root causes of suboptimality across DBMSes: (i) the cost model and (ii) the plan search process.

7. DIMINISHING RETURNS?

In order to increase query performance, DBMSes are extended over time with new relational query operators. Also over time, DBMSes are extended with new storage and indexing structures, which themselves elicit new query operators. Each subsequent generation of the DBMS thus supports an ever-expanding collection of operators, thus the current version is simply the most recent within a series of DBMS *generations*.

7.1. A Gedanken Experiment

Consider a Gedanken experiment that analyzes the plan for each query at each cardinality, that is, for each Q@C, for each DBMS generation. In early generations, there will be fewer plan changes as the cardinality increases for a given query, just because there are fewer operators available. For later generations, some of the Q@Cs will be associated with different plans, enabled by the new operators added by subsequent generations. As our causal model in Figure 5 predicts, the effective plan space for a query may grow in subsequent generations, due to the increased number of operators available.

In many (hopefully most) cases, the new plan will be more efficient than staying with the previous plan selected in a previous generation. After all, that is the very reason the new operator(s) were added to that generation. However, in the presence of suboptimality, sometimes the new plan at that Q@C is *not* preferable, as that DBMS generation's query optimizer selected the wrong plan. Indeed, the query optimizer also evolves and improves with each new generation, in part to minimize the chance of selecting the wrong plan.

Our predictive model suggests that suboptimality is causally impacted by the independent variable of CEPS (cardinality of effective plan space **TODO:** *Rick:Connect with number of operators available and with Section 6.8* . A possible implication is that as the generations of the DBMS add more and more operators, suboptimality will also increase. **TODO:** *Rick:No: causality is with CEPS, not number of operators!*

TODO: *Rick:Make point that our causal analysis is across DBMSes with different numbers of operators. He*

The underlying question then becomes, does an additional operator made available in a subsequent generation of the DBMS actually help or hurt?

7.2. Simulating DBMS Generations

While we do not have access to prior generations of our DBMSes (which is why the previous discussion was in the form of a Gedanken experiments), we can approximate this with an experiment using data already collected on the current version available for each DBMS, to *simulate* the prior generations having a fewer number of operators, and thus a smaller set of realizable query plans. (The effect will certainly be smaller in our simulation, as we are nonetheless using the most recent *query optimizer*n in all of the simulated generations.)

Our data, drawn from the 8,840 query instances and their 112,118 Q@Cs in both the exploratory and confirmatory experiments of the previous study, consists of pairs of adjacent plans (let's refer to them as the *lower* plan and the *upper* plan) for the same query at the *lower* and *upper* cardinalities) that are consecutive (that is, separated by the minimum cardinality, either 10,000 or 300 rows), each with an actual execution time. All of Q@Cs having cardinalities between the upper plan of one pair and the lower plan of the next pair are associated with that same plan, guaranteed by the process in which we chose those Q@Cs for actually timing.

For this generational experiment, we associate with each Q@C a generation (a positive integer) that is the earliest generation containing all the operators in the plan. We also only consider Q@C pairs (with adjacent Q@Cs) where the lower plan has a generation distinct from the upper plan. Say the upper generation is earlier than that of the lower plan, as illustrated in Figure 8. In this figure, as we scan from left to right in increasing cardinality, we first encounter Plan A from generation 2 at the first cardinality of 10K rows, then Plan B sometime later, then even later, a pair of Q@Cs with Plan B at 900K rows and Plan A at the adjacent 910K rows, then much later Plan A again at 1320K rows (this is the measured Q@C that has Plan A that is closest to the measurement at 910K rows).

In this case, the thinking goes that, had we been in the DBMS generation 2, the optimizer would have selected Plan A throughout, because Plan B simply wasn't possible (as it involves an operator not present in generation 2). Then when generation 3 was created by adding an operator, the optimized chose Plan B for 900K. The reason Plan B was chosen is that it is faster than Plan A at that cardinality, shown in the figure by extrapolating the time down from 910K down to 900K (we will revisit this extrapolation shortly).

In this particular case, Plan B is more appropriate (faster) than Plan A, but that is not the only possibility. Sometimes the later generation with more operators available chooses the *wrong* plan, one that is slower than the plan chosen by the earlier generation, due to the suboptimality we've observed. (We'll examine an example shortly.)

The question then becomes, does a plan change by a subsequent generation (enabled by one or more additional operators) represent a win (runs faster) or a loss (runs slower, because of suboptimality)? More broadly, do the Q@Cs in the aggregate enabled

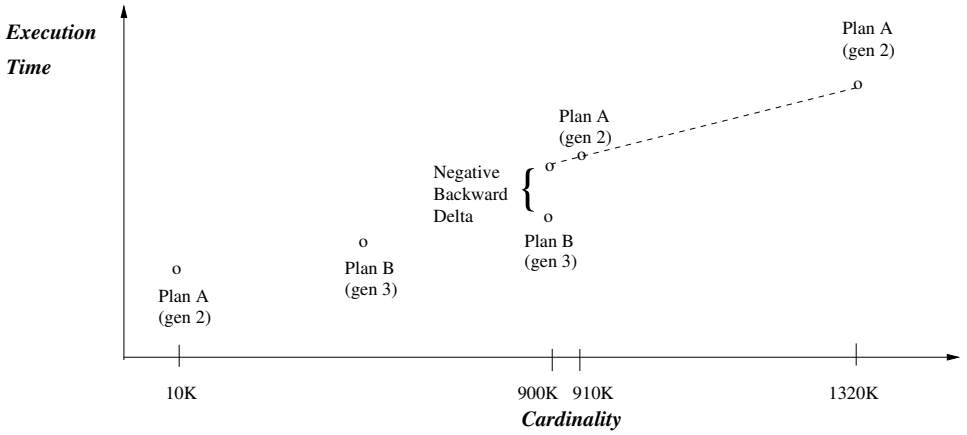

Fig. 8. An Example of a Negative Backward Delta

by each succeeding generation continue to overcome the increasing burden of suboptimality?

7.3. Characterizing DBMS Generations

To which generation do we assign each DBMS operator?

Note that we don't actually know the specific order in which the operators were added to each DBMS. But even if we did, that order was somewhat arbitrary, with a host of considerations going into those decisions over the years. Given that we are using the same optimizer for each such defined generation, we'll adopt a more systematic ordering of the operators. We start by gathering all *single-operator* plans, all two-operator plans, and so forth, and order the generation by the prevalence of their appearance in these query plans.

Specifically, for each DBMS we designate the first generation to contain the single operator that maximizes the number of plans at change points, that is, maximizing the number of Q@Cs, containing just that operator. So for example, all the plans generated by MySQL that have exactly one operator involve just the Full Table Scan operator. So no choice was needed: we designate the first generation as just having that one operator. Any plan with just that operator can be constructed by MySQL generation 1, as well as by any subsequent generation (as each generation includes that initial operator).

We then examine the plans containing exactly two operators (in the case of MySQL, one of which is full table scan, as a two-operator plan not having that operator would necessarily be constructed by generation 3 or higher). Using MySQL again, there are two sets of such plans, those with the full table scan and full table scan with join operators (888 Q@Cs) and those with the full table scan and ref operators (112 Q@Cs). We pick full table scan with join as the operator added by Generation 2, given its prevalence of Q@Cs (at total of 1705 in the full set). Each subsequent generation adds that operator that maximizes the number of plans that operator will eventual enable. So Generation 3 adds the eq_ref operator, as that operator enables 633 plans eventually. Generation 4 adds the ref operator and Generation 5 adds the index operator. There are six plans that involve all five operators.

The generations of MySQL thus can be characterized from the Q@Cs we encountered: five distinct combinations of two operators, covering 1086 Q@Cs, six combinations of three operators (1126 Q@Cs), two combinations of four operators (104 Q@Cs),

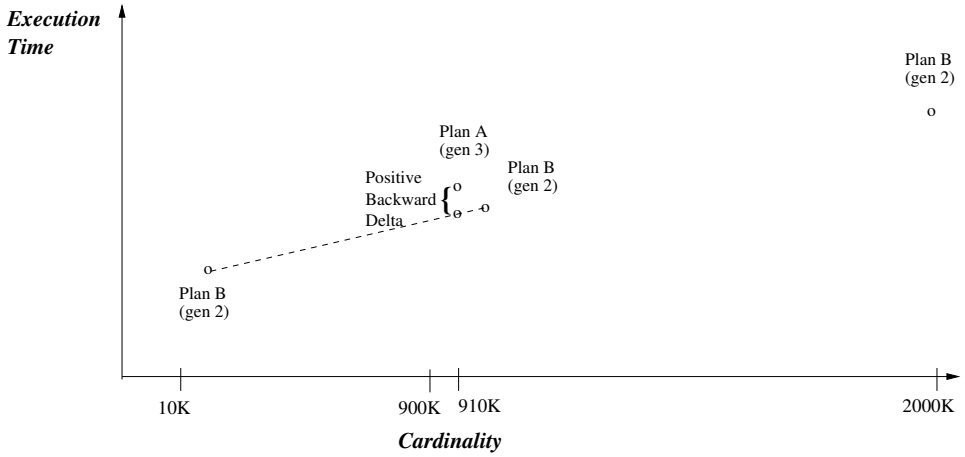

Fig. 9. An Example of a Positive Backward Delta

and (naturally) exactly one combination of all five operators (6 Q@Cs). Thus for MySQL our data implies a simulation of five distinct generations, each adding an operator.

$$
\begin{aligned}
g_1 &= \{\textit{full table scan}\} \\
g_2 &= \{\textit{full table scan}, \textit{scan with join}\} \\
g_3 &= \{\textit{full table scan}, \textit{scan with join}, \textit{eq\_ref}\} \\
g_4 &= \{\textit{full table scan}, \textit{scan with join}, \textit{eq\_ref}, \textit{ref}\} \\
g_5 &= \{\textit{full table scan}, \textit{scan with join}, \textit{eq\_ref}, \textit{ref}, \textit{index}\}
\end{aligned}
$$

For the four DBMSes in our study, the number of generations ranged from five to thirteen.

7.4. Using Change Points

We now consider how to use data already collected, that is the *change points* discussed in Section 6.3: consectutive cardinalities (10K apart; 300 for one DBMS) with different plans. Specifically, how can change points be used to evaluate the effectiveness of different generations of a DBMS?

Define $\textit{ops}(p)$ for a given plan p to be the set of operators present in that plan, with some operators perhaps repeated in that plan. For example, plan p_1 may be composed of a *select* of an *agg* of a *select* over a relation. Hence, $\textit{ops}(p_1) = \{\textit{select}, \textit{agg}\}$. A generation g is *applicable* to a plan p if $\textit{ops}(p) \subseteq g$, and so it is the case that g_4 is applicable to p_1. By definition, if a generation is applicable to a given plan, it is applicable to all subsequent generations. The smallest such generation is termed the *minimally applicable generation*, or *mingen*, so $\textit{mingen}(p_1) = g_3$.

First consider pairs for which $\textit{mingen}(\textit{lower}) > \textit{mingen}(\textit{upper})$. An example is shown in Figure 9. In this case, the set of operators in Plan A for the Q@C at cardinality 900K requires at least generation 3, whereas the set of operators for Plan B for the Q@C at cardinality 910K requires but generation 2. (Note that the generations here are consecutive, that that is not required. Sometimes the generations of pairs of Q@Cs are wildly different.)

To examine the wisdom of picking Plan A for 900K (whose time was measured as the *higher* point shown (the one above the dashed line), we find the closest Q@C also

having Plan A. There are two illustrated here, at the minimum and maximum cardinalities, of 10K and 2000K, respectively, with the closer one at 10K. (If there is no such change point, we can't do the extrapolation).

We use the slope of the dashed line between the measured query time of Plan B at cardinalities 10K and 910K to get an estimated query time at 900k. This realizes an estimate of how fast the query using Plan B would have run on the database having the variable table with a cardinality of 900K. In this particular case, Plan B looks like it would have run *faster* (in fact, was was faster even at 910K in this particular example).

As illustrated in Figure 1, plans can be discontinuous, because their asymptotic complexity is $O(n \lg n)$ (though this often turns into a sequence of linear relationships, as illustrated by this figure). This behavior will, for plan B present at the two granularities of 10K and 910K, slightly overestimate the runtime of that plan at 900K, and thus *underestimate* the penalty of going with Plan A rather than Plan B. That said, given that we are extrapolating from the actual run time at a very close cardinality, our extrapolated estimate should be close.

We term this a *backward extrapolation*, because we are extrapolating backward from a cardinality of 910K to 900K.

With this estimate, we can compute the *relative delta*, defined as the measured time of Plan A (the chosen one) at the lower cardinality minus the extrapolated time of Plan B at that same cardinality, divided by the measured time of Plan B. The relative delta is thus scaled by original measured time at that granularity, and thus has a value between -1 and 1.

In this case, the relative delta is positive (the measured time is greater than the extrapolated time, which indicates that the optimizer chose the *wrong plan*: here, Plan B should have been chosen. A negative relative delta implies that the additional operator(s) available to the minimally applicable generation of the upper plan were indeed beneficial, that the measured time was lower. (In such cases, we again divide by the larger value.) A positive relative delta implies the presence of suboptimality: the wrong plan was chosen by the DBMS generation, perhaps due to the greater number of operators available. (Note that this is a slightly more expansive definition of suboptimality than that used earlier and illustrated in Figure 1.)

There are thus seven orthogonal possibilities for each change point (that is, adjacent pairs of Q@Cs, a total of 66,769): (i) the pair share the same generation, that is, $\mathit{mingen}(\mathit{lower}) \neq \mathit{mingen}(\mathit{upper})$ (62,903 Q@C pairs, 94.2%), which we don't consider further, (ii) there is no other cardinality having the same plan as that with the later generation, in which case we also ignore this Q@C (768 Q@C pairs, 1.2%), (iii) the extrapolation yielded a computed query time that was negative, which we also drop (10 pairs, 0.01%); (iv) the extrapolation was from above and indicates no suboptimality, termed a *negative backward relative delta*, as exemplified in Figure 8, examined earlier (492 pairs, 0.7%); (v) the extrapolation indicates no suboptimality, termed a *positive backward relative delta*, exemplified in Figure 9 (583 pairs, 0.9%); (vi) the extrapolation was from below (consider Figure 8 but with Plan A from generation 5; we would then need to extrapolate from the closest Plan B, which is *at a smaller cardinality*, in a *forward direction*, to compute a *negative forward relative delta* (1218 pairs, 1.8%); and (vii) a *positive forward relative delta* (795 pairs, 1.2%). This analysis is for a single pair of adjacent Q@Cs for a single query running on a specific DBMS, providing a relative delta for the minimally applicable generation. This is a percentage difference, and so is not affected by the absolute magnitude of the run time nor of the cardinality in question. Indeed, because it is a percentage difference, the relative delta is not affected by the query nor even which DBMS is involved. We associate this

relative delta with the later generation, for it is that generation which had the choice between the two plans for that query.

7.5. Trends Across DBMS Generations

Consider what might the *average relative delta* behave across the generations, that is, how the average of the relative deltas associated with each generation will change with successive generations. The average relative delta is a characterization of the aggregate impact of that generation, providing a quantitative estimate of the benefit of adding that operator. (We use average so that each point is not impacted by the number of pairs over which that point is computed.)

What does our causal model in Figure 5 say about this? That causal model asserts that as the number of operators increases, the latent measures in plan space complexity increase, which therefore impacts suboptimality, also in a positive direction.

Our Gedanken experiment in Section 7.1 takes this predicted behavior and predicts taht as the generations of the DBMS contain successively greater numbers of operators, suboptimality will also increase.

Drilling down into this interaction, the next question is: By how much is the increase in efficiency enabled by a new operator (for those Q@Cs utilizing a plan containing that operator) greater than the decrease in efficiency resulting from suboptimality (for a subset of those Q@Cs) resulting from adding that operator? Given that each new operator will improve a shrinking subset of Q@Cs, while the suboptimality occurs in a portion over an expanding subset of Q@Cs, it seems that there must be a point where the decrease due to suboptimality obviates the increase enabled by the new operator. How close are modern DBMSes to that limit?

This analysis implies that each additional operator will result in a net *decrease* in average relative delta, with the possible result that the advantage of that additional operator exceeded the net impact of suboptimality. We thus predict that the first derivative of a relationship between average relative delta over increasing generation will be negative, a natural result of the efforts of DBMS developers to increase performance over successive generations of their DBMS.

That said, each new operator is applicable to a successively smaller portion of the queries, and perhaps over a successively smaller portion of the cardinality space. (We note in passing that the space of queries over which our particular experiment ranged is a rather small, though oft-used, portion of the exceeding complex SQL definition: over 3000 pages [20].) As already noted, our causal model predicts that the prevalence of suboptimality would increase as the number of operators available increased. We thus predict that the *second derivative* of the average relative delta for increasing generation will be *positive*, that is, that the curve will start to level off. If that indeed occurs, it raises the possibility of the curve either asymptotically approaching a horizontal line (the first derivative approaching 0), or worse: the first derivative changing to positive, with the average time over the queries we are studying actually *increasing* with DBMS generation.

We partition the relevant four sets of change points into two groups: (i) those with a *positive* relative delta (either forward or backward), denoting a *suboptimal decision* by the query optimizer at the later generation (correcspeonding to a higher generation number) and (ii) those with a *negative* relative delta (either forward or backward) relative delta, indicating a *non-suboptimal decision* by the query optimizer at the later generation. (We can't state unequivicably that the plan is optimal, because there may be yet another plan involving the operators within that generation that is even faster.)

Let's first examine those change points for which the query optimizer made a good decision, the non-suboptimal change points; see Table IV.

Table IV. Non-Suboptimal Change Points

Generation	*Number of Change Points*	*Average Relative Delta*	*Cumulative Relative Delta*
1	—	—	0.00
2	—	—	0.00
3	—	—	0.00
4	24	0.28	0.28
5	282	0.12	0.13
6	309	0.27	0.20
7	336	0.32	0.24
8	43	0.20	0.24
9	4	0.12	0.24
10	17	0.29	0.23
11	357	0.18	0.23
12	6	0.25	0.23
13	0	—	0.23
Cumulative	1378	—	0.23

The first thing to note is that this data aggregates the results over the four DBMSes, which had generations ranging from five to 13. Thus this is the first study we are aware of that compares the generations trends of DBMSes over quite disparate code bases, thus getting at fundamental trends.

Overall, only a small number of change points, 1378, or about 2% of the total, satisfied the requirements listed above, including having a different generation number for the lower and upper plans. In fact, there were no such change points for the first three generations nor the last generation.

While the numbers jump around a good bit, the thing to focus on is the *trend* as the generations evolve. The basic trend is a slightly decreasing relative delta, with the cumulative starting off at a high of -0.28 (recall that high negative is good) and slowly decreasing to -0.23, resulting from larger negative numbers trending to lower negative numbers.

This matches our prediction arising from the structural causal model that it gets harder to squeeze out performance gains as operators are added to the DBMS over successive generations.

We now examine those change points for which the query optimizer made a poor decision, in that we can surmize that there was a better plan (the one right next to it in the adjacent Q@C pair). Table V provides the same information across the DBMS generations for the suboptimal change points: those for which the relative deltas are positive, indicating a poor decision. While there are still only a small number of change points, there are a greater number of *suboptimal* change points, with more relatively showing up at more recent generations. But more strikingly, the cumulative relative delta has the opposite behavior to the non-suboptimal change points: it *increases* over the generation, starting at 0.08 and ending at 0.25, a value *higher* than that of the non-suboptimal change points.

This is also as predicted: the optimizer is struggling with both more options (plans over the available operators) to select from and a diminished opportunity to make a significant improvement.

Looking at this struggle from a different perspective, consider the *maxium number of change points per query*, or *CPQ*, from the perspective of DBMS generations. For each query, count the number of Q@Cs at each generation (so for instance query Q1 could have 5 Q@Cs at generation 1, 7 Q@Cs at generation 2, 17 at generation 5, etc.). We then then take the maximum CPQ over the queries (so perhaps 17 is the maximum for generation 5 over all queries). Our hypothesis is that as the generation increases,

Table V. Suboptimal Change Points

Generation	*Number of Change Points*	*Average Relative Delta*	*Cumulative Relative Delta*
1	—	—	0.00
2	—	—	0.00
3	—	—	0.00
4	2	-0.19	-0.19
5	211	-0.08	-0.08
6	86	-0.16	-0.10
7	263	-0.30	-0.20
8	273	-0.30	-0.23
9	3	-0.04	-0.23
10	168	-0.32	-0.24
11	639	-0.24	-0.24
12	53	-0.35	-0.25
13	12	-0.40	-0.25
Cumulative	1710	—	-0.25

Table VI. Max Change Points Per Query, Per Generation

Generation	*Maximum CPQ*
1	1
2	66
3	115
4	136
5	132
6	141
7	42
8	172
9	2
10	12
11	55
12	29
13	1
Cumulative	172

the maximum query flutter will increase, which then influences the maximum CPQ. The results are shown in Table VI.

What we notices is that the maximum CPQ *does* increase with generation, up through generation 8, after which it falls off dramatically. In retrospect, this fall-off makes sense. There can be only 200 Q@Cs for a query, because we look for query plan changes only every 10,000 tuples up to a maximum of 2M tuples. As the generation increases, there is less "room" for more plans (as some fraction of the previously generated plans will still be quite good or even optimal), and so less opportunity for flutter which will show up as a large maximum CPQ. And indeed, at generation 8, there is a query that has an astonishingly high number of change points, 172, all of that generation.

There's another limitation we're hitting, one that is much more fundamental.

The center graph shows the "net benefit," that is, the cumulative average relative delta of both the suboptimal decision and non-suboptimal decision change points. There is a dramatic rise at generation 5, with the rest of the generations presenting a relatively smooth growth, moving from negative relative delta (not suboptimal) to positive relative delta (suboptimal), crossing around generation 10. This highly aggregated result, extracting the 3000-odd change-point pairs from almost 100-thousand Q@Cs having the specified properties of interest, implies that these four particular

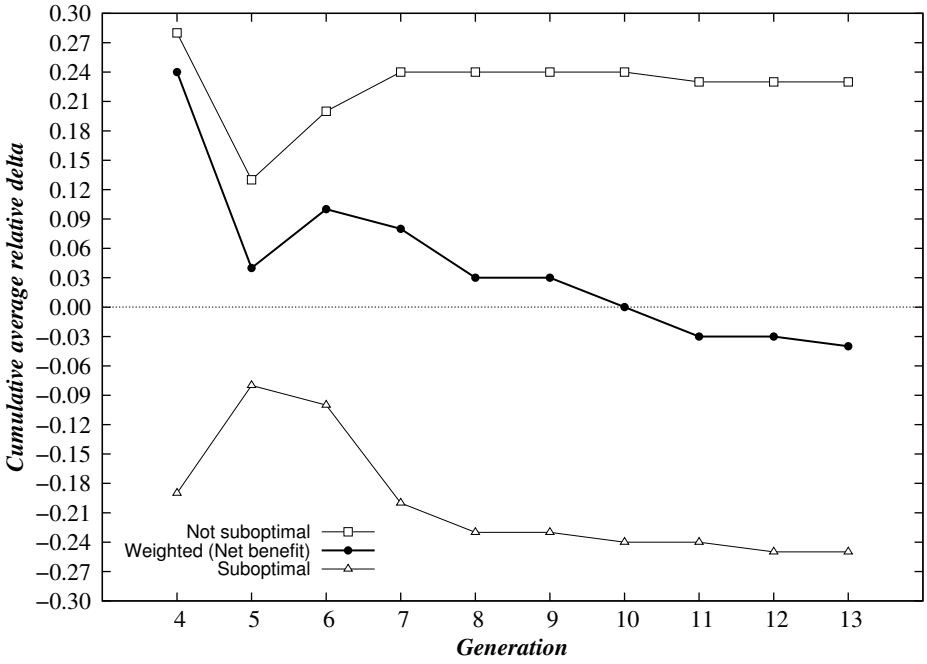

Fig. 10. Cumulative Average Relative Delta over Generations

DBMSes, taken together, might be close to hitting the wall, where adding an operator actually slows down the average query.

More specifically, we can't say whether any of these DBMSes have actually transitioned to where, for this class of queries, the errors of the suboptimal change points overwhelm the benefits of the non-suboptimal change points. Presumably, the DBMS vendors have done extensive tests to ensure that the operator added to each generation did in fact effect a speedup on the representative workloads that they use in evaluating their optimizer enhancements. However, we *can* say that (a) the trends observed strongly point to a decreasing benefit and an increasing cost. as predicted by the simple arguments made above, and (b) if current DBMSes haven't yet reached the point of diminishing returns, that possibility exists.

We should emphasize all the provisos mentioned earlier. We are looking at quite simple queries, over a quite limited range of data, with only a small percentage (3%) of change points identified in our experimental protocol.

8. ENGINEERING IMPLICATIONS

We studied a particular phenomenon, that of the optimizer selecting a wrong plan, indicated by the existence of a query plan that performs more efficiently than the DBMS's chosen plan, for the same query. From the engineering perspective, it is of critical importance to understand the prevalence of suboptimality and its causal factors. The genesis of our model was a sense that suboptimality is caused in part by the inherent complexity of these system and the concomitant unanticipated interactions between various rules in the optimizer.

Through a series of experiments managed by DBLAB, we uncovered several surprising results that provide systematic clues as to where current optimizers come up short and how they can be further improved.

— For many queries, a majority of the ones we considered, the optimizer picked the wrong plan for at least one cardinality, even when the cardinality estimates were completely accurate and even for our rather simple queries.
— A quarter of the queries exhibited significant suboptimality ($>= 20\%$ of the runtime) at some cardinality.

These two results indicate that there is still research needed on this topic. Fortunately, the causal model helps point out specifically where that research should be focused.

— Many queries exhibited *query fluttering*, in which the query optimizer returned to a previous plan at a higher cardinality.
— Some queries exhibited significant *query thrashing*, with a plan change at almost every cardinality. While this phenomenon was first visualized by Haritsa et al. [8; 9] on some complex queries, we have shown that it is present even in a surprising percentage of simple queries.
— Furthermore, some queries exhibited many changes *to a suboptimal plan* as the cardinality was varied.

These particular queries, as well as those of the right-hand side of Figure 4 exhibiting a large degree of suboptimality, can be a starting point for identifying the root cause(s) of query thrashing. The phenomenon can be investigated initially on a per-DBMS basis. Our methodology could then be used to test proposed causal mechanisms of query thrashing across DBMSes, to ascertain the generality of any proposed solutions.

— The causal model and our experimental results suggest that more research is needed to improve the cost model of *discontinuous operators*.
— We also show that it may well be useful to explicitly take *cardinality estimate uncertainty* into account.
— This research indicates that aggregates are *not* a problem, so that aspect of query optimization is in good shape.

TODO: *Rick: evaluate following shorten:*

We see that costing of discontinuous operators is a root cause of query suboptimality, and is thus particularly challenging to a query optimizer. If the cost model is even a little bit off, the optimizer might be on the wrong side of the "jump," thereby selecting the wrong plan. That the presence of discontinuous operators has such a high regression coefficient provides a quite specific guideline: more research is needed to improve the accuracy of the cost model for such operators, such as careful calibration that tunes the cost model with more accurate resource knowledge, including the global memory capacity available, as well as to improve the algorithms that allocate those buffer pages to specific operators.

Concerning the plan search process, another identified root cause of suboptimality, in cases where the DBMS is not as sure about the cardinalities of the underlying relations or the speed of the disk (e.g., if such relations migrated frequently [23]), perhaps the optimizer should explicitly take uncertainty into account. Indeed, others have started to argue that uncertainties in the query planning process should be acknowledged and exploited [2].

We mentioned dynamic query optimization in Section 3 earlier. Dynamic query-reoptimization normally requires a significant amount of information to be recorded during query execution, which can incur non-negligible overhead on the overall query performance [1; 16]. We envision that by utilizing the proposed predictive model for suboptimality, it may be possible to enhance reoptimization techniques such that given a particular query, a particular data distribution, and a specific plan operator, just the

important statistics that affect the operator's performance can be identified and should be recorded, thereby reducing the overhead of bookkeeping irrelevant information.

Hence, the methodology introduced in this paper suggests fairly specifically where additional engineering is needed (the cost model of discontinuous operators and accommodating cardinality estimate uncertainty) and is not needed (costing of aggregation).

TODO: *Rick:discuss implications of generational study*

We should emphasize that this section of this paper, considering engineering implications of the underlying causal model, contrasts with the rest of the paper, whose focus is on the science and on understanding these complex systems at a fundamental level. Good engineering should be built on solid scientific results.

9. SUMMARY

This paper studies an important component of a DBMS, the query optimizer. This component is an amazingly sophisticated piece of code, but is still not well understood after decades of research and development.

This paper makes the following contributions in an attempt to gain new understanding of this component.

— Shows that even for simple queries, the prevalence of *query suboptimality*, *query flutter*, and *query thrashing*, three problems that have not been systematically investigated across DBMSes, is high, and thus there is still research needed on this mature topic.
— Introduces a new *methodological perspective* that treats DBMSes as experimental subjects within *empirical generalization*.
— Proposes *operationalizations* of several relevant measures that apply even to proprietary DBMSes, as well as an overarching *predictive model* that attempts a causal explanation of suboptimality, encoding some of what is known about query optimization.
— Utilizes a novel *research infrastructure*, DBLAB, to express, run, and analyze experiments.
— Tests *eight hypotheses* deductively derived from the causal model. A correlational analysis and a regression analysis provided *strong support* for our model, across DBMSes, thus qualitatively confirming what was informally known.
— Provides a *space of hypotheses* that in concert through further investigation can refine and support (or not) this model or suggest a better model.
— Uncovers compelling *evidence* (a) that suboptimality correlates with two operationalizations of query complexity, (b) that suboptimality correlates with two operationalizations of plan space complexity, (c) that query complexity is a contributor to plan space complexity, and (d) that schema complexity, as operationalized by the presence of primary key attributes, moderates these three interactions.

TODO: *Keep the following shorten?*

— For the kinds of queries we looked at, the factors that we identified in our model: optimizer complexity, query complexity, and plan space complexity, in concert predicted a significant portion of the variance of suboptimality. It is doubtful that any other factor, as yet unknown, will itself predict as much variance as the factors we studied in this paper. That said, it is certain that there remain several unknown causal factors; identifying those factors will undoubtedly also have important engineering implications.
— Articulates for the first time a possible *upper bound* on the number of operators a DBMS may be able to support, given that the empirical evidence suggests that additional operators speed up a smaller and smaller portion of the query/cardinality space

while incurring an increasing chance of suboptimality over the remaining space, which is growing.
— Applies a novel experiment over the pairs of adjacent Q@Cs to show **TODO:** *Rick:what?*
— Provides a *path toward scientific progress* in the understanding of a key enabling technology. It is important to emphasize that our model doesn't apply to just one implementation of the algorithm or to one DBMS. Rather, it is quite broad, applying to any DBMS with a cost-based optimizer.
— Thereby identifies *specific directions for engineering interventions*.

This paper thus suggests a framework of casual model elaboration and directed engineering efforts that could reduce the prevalence of query suboptimality to an acceptably low level.

10. FUTURE WORK

There are several directions we wish to take this work. With the model refinements we propose here, additional directed pointers to engineering efforts should emerge.

TODO: *Rick: keep following shorten?*

We want to look into query flutter and thrashing in greater detail, as those phenomena provide concrete indicators of problematic optimizer behavior. One possible methodological approach is to utilize SQL optimizer hints, such as "`+ SEMIJOIN`" in MySQL and "`enable_hashjoin(false)`" in PostgreSQL, to encourage the optimizer to produce more plans at a given cardinality, that can then be timed to make more explict the entire plan space (recall that CEPS, the cardinality of the effective plan space, is no greater and probably much smaller than the cardinality of the plan space).

The investigations in this paper were based on very simple SPJ (select-project-join) queries. We wish to also consider *schema complexity*: foreign keys and indexes and *query complexity*: complex predicates, subqueries, and user-defined data types, methods, and operators. We also want to manipulate DBMSes internally (at least for those that are open-source), turning on and off the rules and observing suboptimality. Our causal model is extensible, in that we can add other factors, as long as their proper operationalization can be established, and additional causal links. So for sub-queries, we can add nesting level, type of sub-query (scalar, correlated, etc.), and number of "effective joins" (as some queries can be rewritten into joins with the outer level). It would also be interesting to study how a suboptimal subquery can affect the suboptimality of the containing query. For instance, is it true that if many subqueries are themselves suboptimal, does that causally impact whether the overall query is suboptimal? Finally, we may be able to determine whether certain query rewrite techniques are employed within each DBMS, introducing another "DBMS complexity" factor into our model.

Kabra and DeWitt have identified another source of complexity: inaccurate statistics on the underlying tables and insufficient information about the runtime system: "amount of available resources (especially memory), the load on the system, and the values of host language variables." [16, p. 106]. Might there be other unanticipated interactions, that are unknown simply because they haven't been looked for?

By employing an extensible causal model, many complex factors can be studied via a systematic, statistically sound, scientific manner to better understand the causal factors and their interactions.

In addition to the refinements of the causal model just discussed, our research has provided specific directions for implementation interventions: refining the cost model

of discontinuous operators, improving buffer allocation algorithms, and accommodating cardinality estimate uncertainty.

TODO: *Rick: revise: talk about generational study, how we're hitting a wall, and how it might be necessary*

The methodology introduced in this paper and the causal model that results has suggested both the scope of the problem of query suboptimality and a number of specific engineering efforts that can now be carried out. Our ultimate goal is a refined causal model that can fully explain how query suboptimality arises in cost-based optimizers, thereby enabling engineering solutions that reduce and eventually effectively eliminate query suboptimality.

Appendix A provides further details on the experiments.

REFERENCES

R. Avnur and J. M. Hellerstein, "Eddies: Continuously Adaptive Query Processing", in *Proceedings of the ACM SIGMOD Conference*, pp. 261–272, 2000.

B. Babcock and S. Chaudhuri, "Towards a Robust Query Optimizer: A Principled and Practical Approach," in *Proceedings of the ACM SIGMOD Conference*, pp. 119–130, Baltimore, Maryland, 2005.

D. J. Campbell, "Task Complexity: A Review and Analysis," *Academy of Management*, 13(1), pp. 40–52, 1988.

S. Chaudhuri, "An Overview of Query Optimization in Relational Systems," in *Proceedings of the ACM PODS Conference*, pp. 34–43, Seattle, WA, 1998.

S. Currim, R. T. Snodgrass, Y. -K. Suh, and R. Zhang, "DBMS Metrology: Measuring Query Time," in *Proceedings of the ACM SIGMOD Conference*, pp. 421–432, June 2013.

S. Currim, R. T. Snodgrass, Y. -K. Suh, and R. Zhang, "DBMS Metrology: Measuring Query Time," ACM *Transactions on Database Systems*, 42+8 pages, October, 2016.

G. Graefe, "Query Evaluation Techniques for Large Databases," *ACM Computing Surveys* 25(2), pp. 73–170, June 1993.

D. Harish, P. Darera, and J. R. Haritsa, "On the Production of Anorexic Plan Diagrams," in *Proceedings of the VLDB Conference*, pp. 1081–1092, 2007.

J. R. Haritsa, "The Picasso Database Query Optimizer Visualizer," *PVLDB* 3(2):1517–1520, 2010.

A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research," *MIS Quarterly* 28(1):75–105, 2004.

A. Hulgen and S. Sudarshan, "AniPQO: almost non-intrusive query optimization for nonlinear cost functions," in *Proceedings of the VLDB Conference*, pp. 766-777, 2003,

Y. Ioannidis, "Query Optimization," *ACM Computing Surveys* 23(1):121–123, June 1996.

Y. Ioannidis, "The History of Histograms (abridged)", in *Proceedings of the VLDB Conference*, pp. 19–30, September 2003.

ISO, "ISO SQL:2008 International Standard," 2008.

M. Jarke and J. Koch, "Query Optimization in Database Systems," *ACM Computing Surveys* 16(2):111–152, June 1984.

N. Kabra and D. J. DeWitt, "Efficient mid-query re-optimization of sub-optimal query execution plans," in *Proceedings of the ACM SIGMOD Conference*, pp. 106–117, 1998.

M. V. Mannino, P. Chu, and T. Sagar, "Statistical Profile Estimation in Database Systems," *ACM Computing Surveys*, 20(3), pp. 192–221, 1988.

J. Melton, **Advanced SQL:1999**, Morgan Kaufmann, 2003.

J. Melton and A. R. Simon, **Understanding the New SQL: A Complete Guide**, Morgan Kaufmann, 1993.

J. Melton (editor), ISO/IEC 9075, Database Language SQL:2011 Part 2: SQL/Foundation, December, 2011.

D. L. Moody, "Metrics for Evaluating the Quality of Entity Relationship Models," *Proceedings of International Conference on Conceptual Modeling*, pp. 211–225, Springer, Singapore, 1998.

R. Ramakrishnan and J. Gehrke, **Database Management Systems**, Third Edition, 2003.

F. R. Reiss and T. Kanungo, "A Characterization of the Sensitivity of Query Optimization to Storage Access Cost Parameters," in *Proceedings of the ACM SIGMOD Conference*, pp. 385–396, 2003.

P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lori, and T. G. Price, "Access Path Selection in a Relational Database System," in *Proceedings of the ACM SIGMOD Conference*, pp. 23–34, 1979.

M. Winslett, "David DeWitt Speaks Out," *ACM SIGMOD Record* 31(2), pp. 50–62, June 2002.

Table VII. Experiments 1–7: Detailed Run Statistics **TODO:** *Rick: Please properly double-blind protocol names.*

	Experiment	*Data Sets used*	*Lab Shelves*	*What was Examined?*	*Was query Timed?*	*Number*	*Number of Retained QEs*
1	Monotonicity	A	6.0	all cardinalities	yes	12,000	12,000
2	Initial Exhaustive	A	5.19,6.0	all cardinalities	yes	320,000	244,787
3	Exhaustive	A	7.1	all cardinalities	no	320,000	319,980
4	Exhaustive with Keys	B	6.0	change points	no	—	—
5	Initial Exploratory	A + B	5.19 + 5.2 + 6.0	change points	yes	88,420	68,891
6	Exploratory	A + B	7.1	change points	yes	125,600	114,377
7	Confirmatory	A + B + C +D	7.1	change points	yes	995,580	890,631

A. DETAILS ON THE EXPERIMENTS

Table I lists the run statistics of the seven experiments used in this paper. In this appendix we provide more detailed information on the experiments.

Table VII gives some of those details. We'll walk through the columns in succession.

The third column states the data set used in each experiment, that is, the specific tables being queried. There are four data sets, named A, B, C, and D.

We first discuss the features shared between the four data sets. As introduced in Section 2, the queries referenced tables `ft_HT1`, `ft_HT2`, `ft_HT3`, and `ft_HT4`. All four tables contain four columns, each of type integer. The specific values of the rows for all four columns depend on the percentage of skewed data. Section 5.4 provides the algorithm for generated the values for different values of skew; this algorithm is used in each column, so all four columns in any row will have identical values. **TODO:** *Young: is this true?*

There was one version of the last three tables, for use with MySQL, with cardinality 60K, and one version for the rest of the DBMSes, with cardinality 2M.

We generate 200 versions of `ft_HT1`, termed the *variable table*. For MySQL, these version contain 300, 600, 900, 1200, ..., 59,700, and 60,000 rows; for the rest of the DBMSes, these versions contain 10,000, 20,000, 30,000, ..., 1,970,000, and 2M rows, as introduced in Section 2.

We now differentiate the data sets, elaboring on the discussion in Section 6.2–6.3. Data Set C is perhaps the simplest to describe: it specifies no primary key, has no duplicate rows, and has no skew (of course, for any of the four tables). Data Set A differs from Data Set C only in that there is skew. As summarized at the end of Section 5.4, we use five skewness values: 0 (approximately), 0.001, 0.1, 0.5, and 1.0.

Data Set B is similar to Data Set A, adding the specification of the first column as the primary key. And Data Set D is similar to Data Set B, adding the specification that the other three columns should each be associated with a secondary index, only only that one column. We see the confirmatory examined a much larger variation of data sets than the exploratory studies.

The next column of Table VII concerns the *Lab Shelf*. DBLAB utilizes the metaphor of a bookshelf of lab notebooks. Here, each shelf is associated with a version of DBLAB itself. For the experiments in this paper, we used at vearious times over the last three years labh shelves (that is, program versions) 5.19, 5.2, 6.0, and 7.1. Versions 5.19 and 5.2 were very similar; both implemented TTPv1. Version 6.0 also implemented TTPv1, but collected more query measures that were not relevant for this paper. Version 7.1 implemented TTPv2.

The DBLAB system also includes support for *experiment scenarios*, which are Java code that actually performs the experiment, such as varying the cardinality and running different queries on the data. The only difference in the scenario code across these experiments was in accommodating the details of the data set (that is, creating secondary indexes and data skew) and adding robustness features, such as using exponential backoff when the network connection to the DBMS was temporarily lost.

The bottom line is that while the lab shelf and experiment scenario varied somewhat, the only important aspect was the *Protocol* column of Table I.

We now turn to the fifth column, "what was examined?" Here there are just two possibilities, all 200 cardinalities or just the cardinalities at which the query plan changed. The sixth column, "wheat was timed?", indicates that experiments three and four, Exhaustive and Exhaustive with Keys, described in Sections 5.5 and 6.3, just collected query plans, and thus took much less time to run.

The following sets of queries were used in the seven experiments.

TODO: *Young: how does the following look? Any changes needed?*

QSa. 100 queries over the four tables, generated as described in Section 5.3
QSb. 100 queries, generated the same way
QSc. 100 queries
QSd. 100 queries
QSe. 100 queries
QSf. 100 queries
QSg. 100 queries
QSh. 100 queries
QSi. 100 queries
QSj. 100 queries
QSl. A primary key query set consisting of 230 queries from *QSa–QSf*
QSm. A primary key query set consisting of 160 queries: 40 queries drawn each) from *QSg–QSj*
QSn. A primary key query set 4 with 110 new queries
QSo. A no-skew query set consisting of 100 queries without aggregates for no skew
QSp. A subquery query set consisting of 100 queries, each with a subquery
QSq. A subquery query set consisting of 100 queries, each with a subquery
QSr. A primary key query set consisting of 100 queries drawn from *QSl* and *QSm*
QSab'. 100 queries drawn from *QSa* and *QSb* for which the where selection involved a primary key column

TODO: *Young what does this mean: "This query set was revised to QSI and QSm."? Als, it doesn't seem like*

Experiment 1 (Monotonicity) used the first 50 queries from QSa for one DBMS plus the first six queries from QSa and two queries each from *QSd* and *QSe* for MySQL, for a total of 60 query instances.

Experiment 2 (Initial Exhaustive) used the first 50 queries from *QSa* for the three other DBMSes plus the ten queries for MySQL from Experiment 1, for a total of 160 query instances. Experiment 3 (Exhaustive) used the same 160 queries.

Experiment 4 (Exhaustive with Keys) used the first 50 queries from *QSa* for all four DBMSes, for a total of 200 query instances.

Experiment 5 (Initial Exploratory) used the (100) queries from *QSa–QSf* (for one DBMS), the first 20 queries from *QSa–QSf* (for another DBMS), the first 10 (primay key) queries from *QSa* (for all four DBMSes), and the first 10 (non-primary key) queries from *QSa* (for the other two DBMSes), for a total of 780 query instances.

Experiment 6 (Exploratory) used *QSa* and *QSb*, plus the first 100 (primary key) queries from *QSl* and *QSm*, for all four DBMSes, for a total of 1200 query instances.

Experiment 7 (Confirmatory) used *QSc–QSr*, along with *QSp* for primary key and subquery, *QSp* for subquery, *QSp* and *QSq* for primary key and secondary index and subquery), all across all four DBMSes, for a total of 7,640 query instances.