COLLAPSE RESULTS IN DATABASE THEORY

MICHAEL A. TAITSLIN

In the relational model of databases a database state is thought of as a finite collection of relations between elements.

Database relations (tables) are always going to be finite. However, it is often convenient to assume that there is an infinite *domain*—for example, the integer or rational numbers or the strings—such that the data elements are chosen from this domain.

Sometimes, the domain is called the universe.

Functions and relations defined over the entire domain, like $<$ and $+$, may also be used in querying, for example, if the language of first-order logic FO is used as the query language, its formulas may use the database relations as well as the domain relations, while variables range over the entire domain.

These functions/relations are infinite by their nature.

Some special problems arise if we use such an approach. In the talk I discuss some of the problems.

In 1990 I visited Yuri Gurevich in Ann Arbor.
Every day we spent time to walk and to discuss problems.
I asked Yuri the following questions:

Let us consider an ordered universe. Is it possible to improve the expressive power of the query language if its formulas use the order relation as well as the database relations?
Does the answer depend on the universe?

Of course, the queries must be about the database, not about the order. Such a query is called *generic* or invariant. It means that the query is preserved by any permutation of the domain.

Next day, Yuri answered the questions. The answer for the first one is positive, and the answer for the second one is negative.

If we only know that the domain is ordering and know nothing as to the order, we can formulate a query that we cannot formulate without the order.

Never was the Yuri's result published by Yuri. But it is well known.

Now let us consider the ordered domain only.

Let me permit to present more formal definitions.

More formally, a structure of a relational signature L is a non-empty set with a mapping that assigns to every relational symbol in L a relation of the same arity over the set. Let U be an infinite structure over the signature L. This structure is called the *universe.* In this talk, we always consider ordered universes. This means that L includes a binary relational symbol $<$ whose interpretation in U satisfies the axioms of linear order. Let us denote $L_0 = \{<\}$.

A *database scheme* SC is a finite collection of relational symbols of fixed arities. A *database state* (over U) is an assignment to these relational symbols of concrete relations of corresponding arities over U. These relations are called *database relations.* A database state is called a *finite database state* if all the relations are finite.

We fix a database scheme SC and denote $L_0^+ = L_0 \bigcup SC$, and $L^+ = L \bigcup SC$.

A *database query* can formally be defined as a mapping that takes in a database state (of a fixed database scheme), and produces a new relation, of a fixed arity, over U. Thus, every query has an arity. Specifically, queries of arity 0 are called *Boolean queries.* A Boolean query defines a mapping from database states to $\{0, 1\}$, or, in other words, a set of database states of a given database scheme.

Queries can be formulated using *query languages,* the simplest being the language of first-order logic FO. Formulas (queries) of this language use $=$, as well as the relational symbols of the signature and of the database scheme. Thus, a database state essentially defines a structure of a larger signature with U as the domain; then a formula with n free variables defines an n-ary relation over U; sentences define Boolean queries. We say that two L^+-formulas with n free variables are equivalent over finite states over U if they define the same n-ary query.

We consider two languages for querying. Queries of the first one are FO formulas of the signature L_0^+—we call them *restricted.* Queries of the second language are FO formulas of the signature L^+—we call them *extended.*

Speaking informally, for querying, the restricted query language uses stored information only but the extended one also uses a general knowledge, for example, the knowledge of the addition of the natural numbers in the case when the universe is $(\mathbb{N}, <, +)$.

A k-ary query Q is said to be *locally generic over finite states* if $\bar{a} \in Q(s)$ iff $f(\bar{a}) \in Q(f(s))$, for any partial $<$-isomorphism $f : X \to U$ with $X \subseteq U$, for any finite state s over X, and for any k-tuple $\bar{a}$ in X.

Roughly speaking, a query is locally generic if it is preserved by any order local monomorphism any finite subset of U in U.

The problem we are interested in the paper is:
does the general knowledge improve the expressive power of the locally generic queries over finite states?

In contrast to the Gurevich's result, the answer is negative in many cases.

Usually the theorem that the answer is negative for a universe is called the collapse result or the collapse theorem for the universe.

The collapse result holds for

$$(\mathbb{N}, <, +).$$

The universe is called the Presburger's arithmetics.

Then, how much higher than $+$ in $(\mathbb{N}, <)$ can we go?

Obviously, $+, \times$ make locally generic extended queries impossible to express as pure order ones because, for example, the query "the cardinality of P is even" is locally generic and expressible in the FO extended language but not in the restricted language.

But the first-order theory of the arithmetics is undecidable, and the language is bad for querying.

We **conjecture** that it is the general case.

We **conjecture** that

if, in an expansion of $(\mathbb{N}, <, +)$, locally generic extended queries express more than restricted ones, the first order theory of the expanded domain is undecidable.

Moreover, we **conjecture** that

if, in an expansion of

$$(\mathbb{N}, <, +),$$

locally generic extended queries express more than restricted ones, then there is a first-order formula $\Phi(x, y)$ in the language of the expansion such that for any finite subset A of natural numbers there is such a natural number b that

$$A = \{a \in \mathbb{N} \mid \Phi(a, b)\}.$$

In particular, the Random Graph is interpretable in such an expansion.

The real big step in the direction to prove the conjecture is the excellent result of Baldwin and Benedikt.

They proved the collapse result for locally generic queries over ordered domains without any independence formula.

An independence formula

$$\phi(\bar{x}, \bar{y})$$

has the following property:

for each natural n there are

$$\bar{a}_1, \ldots, \bar{a}_n$$

such that for any subset I of

$$\{\bar{a}_1, \ldots, \bar{a}_n\}$$

there is $\bar{b}_I$ such that

$$\{\bar{a}_i \mid \phi(\bar{a}_i, \bar{b}_I)\}$$

is I.

To support the conjecture, we are considering the natural number as the domain, the addition and a unary function as the domain functions, and the usual ordering as the domain relation. So we are considering the universes

$$(\mathbb{N}, <, +, f^{(1)}).$$

The key conditions for decidability of the first-order theories of these domains was pointed out by A.L.Semenov.

Let

$$A(x) = \sum_{i=1}^{k} n_i f(x + m_i).$$

The Semionov's conditions are:

- f is strongly monotone
- for any integer numbers

  $$n_1, \ldots, n_k, m_1, \ldots, m_k$$

  there is natural δ such that

  $$|\ A(x+\delta)|\ > f(x)$$

  for all natural x
- for any integer numbers

  $$n_1, \ldots, n_k, m_1, \ldots, m_k,$$

  either $A(x) > 0$ for almost all natural x,
  or $A(x) < 0$ for almost all natural x,
  or $A(x) = 0$ for all x.
- f module m is periodical for all positive m

It was proved by S.M.Dudakov that in the Semenov's case locally generic extended queries express as well as pure order ones.

The proof is interesting because Oleg Belegradek, Ya'acov Peterzil, and Frank Wagner proved that

$$(\mathbb{Z}, <, +)$$

has no quasi-o-minimal expansions.

If I understand their proof correctly, the same holds for

$$(\mathbb{N}, <, +).$$

So the collapse result is true for f which is interpretable in a Semionov's domain.

Let us try to refute the conjecture.

It is not so easy to present an example of f such that the first-order theory of

$$(\mathbb{N}, <, +, f)$$

is decidable and f is not interpretable in a Semionov's domain.

For example, if $f(x)$ is a polynomial of degree more than 1 from x and $[\log_2 x]$, the theory is undecidable because the multiplication is definable in the theory.

Let

$$f(x) = 2^{([\log_2 x]^2)}$$

for $x > 0$, and $f(0) = 0$. Here

$$[\log_2 x] = i \Leftrightarrow 2^i \leqslant x < 2^{i+1}.$$

If

$$2^i \leqslant x < 2^{i+1}$$

then

$$f(x) = 2^{(i^2)}.$$

This f is not interpretable in

$$(\mathbb{N}, <, +, 2^x)$$

because the first-order theory of

$$(\mathbb{N}, <, +, 2^x, f)$$

is undecidable.

Theorem. *For this f, the theory is decidable. But the collapse result also holds.*

Proof. Let $f^{-1}(0) = 0$ and for $x > 0$,

$$f^{-1}(x) = i \Leftrightarrow f(i-1) < f(i) \leqslant x < f(2i).$$

If

$$2^{(i^2)} \leqslant x < 2^{((i+1)^2)}$$

then $f^{-1}(x) = 2^i$.

We will consider the structure

$$\mathbf{S} = (\mathbb{N}, 0, 1, <, +, [x/n], f, f^{-1} \mid n = 2, 3, \dots)$$

where

$$[x/n] = i \Leftrightarrow ni \leqslant x < n(i+1).$$

Lemma 1. *If any existential formula is equivalent to a quantifier-free formula in* $\mathbf{S}$ *then any first-order formula is equivalent to a quantifier-free formula in* $\mathbf{S}$*.*

Proof. Induction on the number of quantifiers in a prenex form. If the first quantifier is $(\forall u)$ then the formula is equivalent to $\neg(\exists u)\Psi$ where Ψ is quantifier-free. ∎

An existential formula is called sm-formula iff it has the form:

$$(\exists u_1)\dots(\exists u_m)\Phi$$

where Φ is a conjunction of formulas of the forms $t_1 < t_2$ and $t_1 = t_2$ where t_1 and t_2 have the form

$$(\sum_{i=1}^{k}(m_i f(x_i) + n_i x_i)) + m$$

where m_i, n_i, and m are natural numbers, and x_i are variables.

Lemma 2 (**Semionov**). *Any existential formula is equivalent to a disjunction of sm-formulas in* **S**.

Lemma 3. *Any sm-formula is equivalent to a quantifier-free formula in* **S**.

Proof. Induction on the number m of quantifiers.

Let $u_1 > \cdots > u_m$.

Let $[u] = f^{-1}(f(u))$.

If $[u_1] = [u_2]$, the proof is trivial.

Suppose $[u_1] > [u_2]$.

Case 1. Φ is equivalent to

$$t + (\sum_{i=2}^{m}(m_i f(u_i) + n_i u_i)) = m_1 f(u_1) + n_1 u_1 + k[u_1] \& \Psi$$

where t does not contain $u_1, \ldots, u_m$.

Either

$$u_1 \in \{0, 1, \ldots, m_1 + n_1 + k + \sum_{i=2}^{m}(m_i + n_i)\},$$

or

$$u_1 > m_1 + n_1 + k + \sum_{i=2}^{m}(m_i + n_i).$$

In the first case, the proof is trivial.

In the second one, we have:

$$[u_1] = 2^v;\ f(u_1) = 2^{(v^2)};$$

$$m_1 f(u_1) + n_1 u_1 + k[u_1] - \sum_{i=2}^{m} (m_i f(u_i) + n_i u_i) <$$

$$2^v 2^{v^2} + 2^{2v} < 2^{(v+1)^2};$$

$$f^{-1}(t) = 2^v = [u_1];\ f(f^{-1}(t)) = f(u_1);$$

$$m_1 f(f^{-1}(t)) = m_1 f(u_1);\ k f^{-1}(t) = k[u_1].$$

Case 2. Φ is equivalent to

$$t_1 + (\sum_{i=2}^{m}(m_i f(u_i) + n_i u_i)) < m_1 f(u_1) + n_1 u_1 + k[u_1] <$$

$$t_2 + (\sum_{i=2}^{m}(m'_i f(u_i) + n'_i u_i))\&\Psi$$

where t_1 and t_2 do not contain $u_1, \ldots, u_m$, and Ψ does not contain $f(u_1)$.

If $[u_1] \neq f^{-1}(t_1)$ and $[u_1] \neq f^{-1}(t_2)$, the proof is trivial. In another case, we replace $[u_1]$ with $f^{-1}(t_1)$ or $f^{-1}(t_2)$ and replace $f(u_1)$ with $f(f^{-1}(t_1))$ or $f(f^{-1}(t_2))$. ∎

The proof of the collapse theorem is similar to the Dudakov's one. ∎

M.A.Taitslin, Department of Computer Science, Tver State University, Tver, Russia 170000
E-mail address: Michael.Taitslin@tversu.ru