The collapse results in database theory

Michael A. Taitslin

March 23, 2003

Chapter 1

Introduction

In the relational model of databases a database state is thought of as a finite collection of relations between elements. The 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. The domain functions and relations are infinite by their nature. For example, if the language FO of first-order logic 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. 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 *database scheme* ρ is a finite collection of relational symbols of fixed arities. A *database state* (over U) of scheme ρ (in symbols, ρ-state) 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.

A ρ-state s over an L-structure W is said to be *pseudo-finite in* W if (W, s) is a model of the first-order $L(\rho)$-theory $F(W, \rho)$ of all (W, r), where r is a finite ρ-state over W.

Let us consider ordered domains only and FO as the query language.

Roughly speaking, a query is locally generic if it is preserved by any order local monomorphism from a finite subset X of U into U for any finite state s over X.

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

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?

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 (is not relevant) for querying.

Chapter 2

Relativize Pseudo-finite Homogeneity and Isolation

Theorem 1. *Let U be a universe. Let an extended Boolean ρ-query ϕ be locally generic for finite states over U.*

Suppose, for some uncountable κ with $\kappa = \kappa^*$, there is a special model (V, J) of power κ such that J is an indiscernible sequence in V, $V \equiv U$, and the query ϕ is locally generic over pseudo-finite states over J in V. Then ϕ is equivalent over finite states over U to a restricted ρ-query.

Definition 1. *Let M be a model of T and an infinite set I be an indiscernible sequence in M. A complete L-theory T has the (M, I)-Pseudo-finite Homogeneity Property iff whenever $(N, J) \equiv (M, I)$, A and B are pseudo-finite subsets of J in the model N, C and D are finite subsets of N, and $h : (A \cup C) \to (B \cup D)$ is an elementary map in N with ω-saturated (N, A, B, h), for any $a \in N$ there is $b \in N$ such that $h \cup \{(a, b)\}$ is an elementary map in N.*

Theorem 2. *Let $M \equiv U$ and an infinite set I be an indiscernible sequence in M. Suppose the first-order theory of a universe U has the (M, I)-Pseudo-finite Homogeneity Property. Let an extended query ϕ be locally generic over finite states over U. Then ϕ is equivalent over finite states over U to a restricted query.*

Definition 2. *Let* M *be a model of* T *and an infinite set* I *be an indiscernible sequence in* (M, I). *A complete* L*-theory* T *has the second* (M, I)*-Pseudo-finite Homogeneity Property iff whenever* $(N, J) \equiv (M, I)$, A *and* B *are pseudo-finite subsets of* J *in the model* N, C *and* D *are finite subsets of* N, *and* $h : (A \cup C) \to (B \cup D)$ *is an elementary map in* (N, J) *with* ω*-saturated* (N, J, A, B, h), *for any* $a \in N$ *there is* $b \in N$ *such that* $h \cup \{(a, b)\}$ *is an elementary map in* (N, J).

Theorem 3. *Let* $M \equiv U$ *and an infinite set* I *be an indiscernible sequence in* (M, I). *Suppose the first-order theory of a universe* U *has the second* (M, I)*-Pseudo-finite Homogeneity Property. Let an extended query* ϕ *be locally generic over finite states over* U. *Then* ϕ *is equivalent over finite states over* U *to a restricted query.*

Definition 3. *Let M be a model of T and an infinite set I be an indiscernible sequence in M. A complete theory T has the (M, I)-Isolation Property iff for any special $(N, J) \equiv (M, I)$, for any pseudo-finite subset A of J in N, any finite subset C of N, and any element a of N, there is a countable $A_0 \subseteq A$ such that $\mathrm{tp}(a/(A_0 \cup C))$ isolates $\mathrm{tp}(a/(A \cup C))$ in N.*

Theorem 4. *The (M, I)-Isolation Property implies the (M, I)-Pseudo-finite Homogeneity Property.*

Definition 4. *Let M be a model of T and an infinite set I be an indiscernible sequence in* (M, I)*. A complete theory T has the second* (M, I)*-Isolation Property iff for any special* $(N, J) \equiv (M, I)$*, for any pseudo-finite subset A of J in N, any finite subset C of N, and any element a of N, there is a countable $A_0 \subseteq A$ such that* $\mathrm{tp}(a/(A_0 \cup C))$ *isolates* $\mathrm{tp}(a/(A \cup C))$ *in* (N, J).

Theorem 5. *The second* (M, I)*-Isolation Property implies the second* (M, I)*-Pseudo-finite Homogeneity Property.*

[1] John Baldwin and Michael Benedikt. Stability theory, permutations of indiscernibles, and embedded finite models. *Trans. Amer. Math. Soc.*, 352(11):4937–4969 (electronic), 2000.

Definition 5 ([1]). *Let* K *be a class of first-order* (L,P)*-formulas.* (M,I) *is called* (P,K)*-reducible iff*

for every K*-formula* $\phi(\bar{x},\bar{y})$*, there is an order formula* $\psi(\bar{w},\bar{y})$ *such that for every* $\bar{m}$ *there is a* $\bar{c}_{\bar{m}} \in I$ *such that*

$$(\forall \bar{y} \in P)(\psi(\bar{c}_{\bar{m}},\bar{y}) \equiv \phi(\bar{m},\bar{y})).$$

If K *is the class of all the* L*-formulas,* (P,K)*-reducibility is called* P*-reducibility. If* K *is the class of all the* (L,P)*-formulas,* (P,K)*-reducibility is called strong* P*-reducibility.*

Definition 6 ([1]). *A first-order* (L,P)*-formula is called* P*-restricted iff it contains no* P *or it has one of the forms* $(\forall x \in P)\Psi$ *or* $(\exists x \in P)\Psi$ *where* Ψ *is a* P*-restricted formula. A structure* (M,I) *is called* P*-restricted iff*

any (L,P)*-formula is equivalent over* (M,I) *to a* P*-restricted one.*

[2] M.A. Taitslin. A general condition for collapse results. *Annals of Pure and Applied Logic*, 113(1–3):323–330, 2001.

Lemma 6 (see [1] and [2], Theorem 2.5). *Any P-reducible and P-restricted structure is strong P-reducible.*

Theorem 7. *Let M be a model of T and an infinite set I be a dense ordered without end points indiscernible sequence in M.*

Let (M, I) be a P-reducible and P-restricted model. Then I is an indiscernible sequence in (M, I) and T has the second (M, I)-Isolation Property.

Theorem 8 (Dudakov, 2002). *Any P-reducible structure is P-restricted.*

Corollary 9. *Let M be a model of T and an infinite set I be a dense ordered without end points indiscernible sequence in M.*

Let (M, I) be a P-reducible model. Then T has the second (M, I)-Isolation Property.

Definition 7. *Let M be an L-structure. An L-formula $\phi(\bar{x}, \bar{y})$ is called an independent formula in M iff the following property holds:*

for each natural n there are $\bar{a}_1, \ldots, \bar{a}_n$ in M such that for any subset σ of $\{\bar{a}_1, \ldots, \bar{a}_n\}$ there is $\bar{b}_\sigma$ in M such that $\{\bar{a}_i \mid \phi(\bar{a}_i, \bar{b}_\sigma)\}$ is σ.

M lucks the independence property iff there is no independent formula in M.

Theorem 10 ([1], Theorem 5.2). *Let M be a model of T and an infinite set I with order type a complete dense linear order be an indiscernible sequence in M.*

If M has no independent formula then (M, I) is P-reducible.

Corollary 11 ([1]). *If U lacks the independence property then for any extended query ϕ which is locally generic for finite database states over U, there is a restricted query ψ which is equivalent to ϕ over finite database states over U.*

Chapter 3

An expansion of Presburger's arithmetics by a unary function such that the first-order theory of the expansion is decidable, the expansion has an independent formula and (M, I)-Isolation Property does not hold

[3] A. Blumensath and E. Graedel. Automatic structures. In *Proc. 15th IEEE Symp. on Logic in Computer Science*, pages 51–62, 2000.

[4] M. Benedikt, L. Libkin, T. Schwentick, and L. Segoufin. A model-theoretic approach to regular string relations. In *Proc. 16th IEEE Symp. on Logic in Computer Science*, pages 431–442. IEEE Computer Society Press, 2001.

It was proved in [3] that the first-order theory of $< \mathbb{N}, <, +, |_p >$ where $p > 1$ is a natural number and

$$x|_p\, y \leftrightarrow (\exists u)(\exists k)(x = p^u \& y = kx)$$

is decidable. It is trivial and was observed in [4] that $< \mathbb{N}, <, +, |_p >$ has an independent formula. I prove that in many cases, (M, I)-Isolation Property does not hold for the first-order theory of $< \mathbb{N}, <, +, |_p >$.

Let $f_p(x)=y$ iff $(y|_p\ x\&\neg(py|_p\ x))$. Then

$$(y|_p\ x)\leftrightarrow(f_p(x)\geqslant y\&f_p(y)=y).$$

So the first-order theories of $<\mathbb{N},<,+,|_p>$ and of $<\mathbb{N},<,+,f_p>$ are mutually interpreted.

For simplicity, suppose $p=2$.

First, we point out that $<\mathbb{N},<,+,|_2>$ and $<\mathbb{N},<,\in>$ are mutually interpreted where $x\in y$ iff

$$(\exists z)(\exists u)(y=z+x+u\&z<x\&2x|_2u).$$

Second, we observe that $x<y$ we may use only for x and y such that $x\in x$ and $y\in y$. Indeed, $u<v$ iff

$$(\exists x)(x\in v\&x\notin u\&(\forall y)(x<y\rightarrow(y\in u\leftrightarrow y\in v))).$$

Thus, we will investigate the structure $< \mathbb{N}, <, \in >$ where $<$ is defined on $\{x \mid x \in x\}$ only.

Theorem 12. *Let* $M \equiv < \mathbb{N}, <, \in >$*. Suppose* I *is a subset of*

$$\{x \in M \mid x \in x\}$$

such that I *is an indiscernible sequence in* M *and* $<$ *is a dense order without end points over* I*. Then*

$$\mathrm{Th}(< \mathbb{N}, <, \in >)$$

has no (M, I)*-Isolation Property.*

Chapter 4

Collapse results for expansions of Presburger's arithmetics

The full text of the section was published in [7].

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

[5] A. L. Semenov. Logical theories of one-place functions on the natural number series. *Izv. Akad. Nauk SSSR Ser. Mat.*, 47(3):623–658, 1983. Russian.

[6] S.M. Dudakov. Collapse results for expansions of Presburger's arithmetics by one concordant with addition unary function. Russian, 2001.

[7] M.A. Taitslin. Translation results in database theory. In *Complex systems: data processing, simulation, and optimization*, pages 5–23. Tver State University, Tver, 2002. Russian.

The key conditions for decidability of the first-order theories of these domains was pointed out by A.L.Semenov (see [5]).

It was proved by S.M.Dudakov in [6] that in the Semenov's case, locally generic extended queries express as well as pure order ones. So the collapse result is true for any f which is interpretable in a Semenov's domain.

Let $A(x) = \sum_{i=1}^{k} n_i f(x + m_i)$. The Semenov'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 either $|A(x)| < \delta$ for all natural x, or $|\ A(x + \delta)|\ > f(x)$ 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

A unary function f which the Semenov's conditions hold for is called a *concordant with addition monotone function*. In [5] it was proved that for any structure

$$(\mathbb{N},\ <\ 0,\ 1,\ +,\ \{[\ /n] \mid n = 1, 2, \dots\},\ f,\ f^{-1}) \tag{4.1}$$

of signature

$$\{\ <,\ 0,\ 1,\ +,\ \{[\ /n] \mid n = 1, 2, \dots\},\ f^{(1)}, f^{-1}\} \tag{4.2}$$

where f is a concordant with addition monotone function, $f(0) = 0$, $[x/n]$ is such a natural number y, that $ny \leqslant x$ and $n(y+1) > x$, and $f^{-1}(n) = m$ iff $f(m) \leqslant n < f(m{+}1)$, any first-order formula is equivalent to a quantifier-free formula.

Let

$$f(x) = 2^{([\log_2 x]^2)} \tag{4.3}$$

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.

The same quantifier elimination holds for

$$f(x) = 2^{P([\log_2 x])} \tag{4.4}$$

where $P(x) = d_n x^n + d_{n-1} x^{n-1} + \cdots + d_1 x + d_0$, $d_0, d_1, \ldots, d_n$ are integer numbers, $n > 1$, and $d_n > 0$.

The same quantifier elimination also holds for

$$f(x) = \ell^{g([\log_\ell x])} \tag{4.5}$$

where ℓ is a natural number, $\ell > 1$, g is a unary function from natural numbers into natural numbers, g is periodical by module m for any positive natural m, and $g(n+1) - g(n) > n$.

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

For function f defined by (4.3), we will consider the structure

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

of signature (4.2).

Let $[u]$ denote $f^{-1}(f(u))$. If $2^i \leqslant u < 2^{i+1}$ then $[u] = 2^i$. It is easy to see that $[u] \leqslant u < 2[u]$.

Lemma 13. *Any first-order formula is equivalent to a quantifier-free formula in* $\mathbf{S}$.

Notation $t(x)$ means that term t does not contain any different from x variable. Then $t(a)$ denote the value of $t(x)$ when the value of x is a. $t(\bar{x})$ and $t(\bar{a})$ have the same sense. Similar, notation $\Phi(\bar{x}, \bar{y})$ denote that formula Φ does not contain free variables which are not in variable sequences $\bar{x}$ and $\bar{y}$.

Let I be an indiscernible sequence in M without end points. The sequence is called divided (in M) if for any terms $t_1(x)$ and $t_2(x)$, either there is a natural number n such that $t_2(a) < n$ for any a from I, or for any $a < b$ from I, $t_1(a) < t_2(b)$. The definition was proposed by S.M.Dudakov (see [6]). The structure (M, I) is called divided if I is divided in M.

Theorem 14 (Dudakov,[6]). *For any unary f, there are M and I such that M is elementary equivalent to*

$$(\mathbb{N},\ <\ 0,\ 1,\ +,\ \{[\ /n] \mid n = 1, 2, \dots\},\ f) \tag{4.6}$$

of signature

$$\{\ <,\ 0,\ 1,\ +,\ \{[\ /n] \mid n = 1, 2, \dots\},\ f^{(1)}\} \tag{4.7}$$

and I is divided in M.

Below we suppose that M is elementary equivalent to (4.6) where f is an arbitrary unary function.

Let (M, I) be divided. Let $\bar{x}$ be a sequence of elements of M. Let Y_0 be empty and X_0 be the set of all the elements of $\bar{x}$. Let Y_{i+1} be the set of all the elements of $b \in I$ such that there exist term $t(\bar{x})$, unbounded terms $t_1(y)$ and $t_2(y)$, and $\bar{a} \in (X_0 \cup Y_i)$ such that

$$t_1(b) < t(\bar{a}) < t_2(b). \tag{4.8}$$

For t and $\bar{a}$, there is no $b \in I$ or there is the only $b \in I$ such that (4.8) holds for some unbounded t_1 and t_2. So Y_i is countable for any i. Let Y be the union of all Y_i. We call Y the linked part of I for $\bar{x}$ and denote Y by $[I, \bar{x}]$.

Let us consider any divided (M, I). Fix two variable sequences $\bar{x}$ and $\bar{y}$. A quantifier-free formula is called divided iff it is a Boolean of atomic divided formulas. An atomic formula is called divided iff it has the form

$$t_1(\bar{x}) + t_2(\bar{y}) < t_3(\bar{x}) + t_4(\bar{y}).$$

The theory of (M, I) is divided iff for any first-order formula $\Phi(\bar{x}, \bar{y})$, there exist a quantifier-free divided formula $\Psi(\bar{x}, \bar{y})$ such that for any $\bar{b}$, $\bar{a} \in (I \setminus [I, \bar{b}])$, and $\bar{c} \in ([I, \bar{b}] \cup \bar{b})$, $\Phi(\bar{a}, \bar{c})$ is equivalent to $\Psi(\bar{a}, \bar{c})$. A function f is divided iff there exists a divided (M, I) such that the theory of (M, I) is divided and M is equivalent to (4.1). The next theorem is a trivial corollary of [2], Theorems 3.1 and 3.2.

Theorem 15. *For any structure* (4.6) *where* f *is a divided function, any locally generic extended query is equivalent to a restricted one.*

Theorem 16 (Dudakov, [6]). *The theory of divided* (M, I)*, where* M *is equivalent to* (4.6) *and* f *is a concordant with addition monotone function, is divided.*

Now I propose a sufficient condition for f to be divided. The condition is a generalization of the Dudakov's proof of the last theorem. The function (4.3) supports the condition. So the function (4.3) is divided, and the collapse theorem holds for the expansion of the Presburger's arithmetics by the function. The same is true for all the functions (4.4).

Let M be a structure (4.1) of signature (4.2). A unbounded f is called weakly monotone in M iff $f(0) = 0$, for any natural m there is a natural number k such that $f(x) > mx$ for $x > k$, and for any $x < y$ from M, $f(x) \leqslant f(y)$. For a weakly monotone f, there is a strong monotone sequence of natural numbers $s_f^{(0)}, s_f^{(1)}, \ldots, s_f^{(n)}, \ldots$ such that

- $s_f^{(0)} = 0$ and $f(s_f^{(i)}) < f(s_f^{(i+1)})$ for any natural i
- for any natural i and $s_f^{(i)} \leqslant x < s_f^{(i+1)}$, $f(x) = f(s_f^{(i)})$

For a weakly monotone f, let

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

We will consider the structures (4.1) where f is a weakly monotone function.

A weakly monotone function f is called strong divided iff it has the following properties:

- any sm-formula with m quantifiers is equivalent to a disjunction of sm-formulas with less than m quantifiers and sm-formulas

$$(\exists[u_1])\dots(\exists[u_m])\Theta(\overline{[u]},\overline{x}),$$

where $\Theta(\overline{[u]},\overline{x})$ is a conjunction of a formula

$$\rho(x_1,\dots,x_k,[u_1],\dots,[u_m]),$$

which says that $x_1,\dots,x_k,[u_1],\dots,[u_m]$ have given remainders for given positive natural numbers, and sm-formulas of the forms

$$t_2(\overline{x})+m_2f([u_1])<t_1(\overline{x})+m_1f(u_1), \tag{4.9}$$

$$n_2t(\overline{x})<mf([u_1])<n_1t(\overline{x})\&\Psi(\overline{[u]},\overline{x}), \tag{4.10}$$

$$t_2(\overline{x})+m_2[u_1]<t_1(\overline{x})+m_1[u_1], \tag{4.11}$$

and

$$n_2t(\overline{x})<m[u_1]<n_1t(\overline{x})\&\Psi(\overline{[u]},\overline{x}), \tag{4.12}$$

where m, m_1, m_2, n_1, and n_2 are natural numbers, and $\Psi(\overline{[u]},\overline{x})$ is a quantifier-free sm-formula

- for any natural numbers m, and $n<m$, there is a natural number $g(n,m)$ such that, for any natural numbers a and i, if $s_f^{(i)}=f^{-1}(na)$, then

$$s_f^{(i+g(n,m))}>f^{-1}(ma)$$

- there is a natural number j such that either forms (4.11) and (4.12) never appear or $s_f^{(i+1)}\geqslant 2s_f^{(i)}$ for any natural $i\geqslant j$

- there is a natural number k such that for any natural number i,

$$f^k(s_f^{(i)})\geqslant s_f^{(i+1)}$$

Any concordant with addition monotone function is strong divided. Functions (4.3), (4.5) and (4.4) are strong divided.

Theorem 17. *Any strong divided function f is divided.*

Bibliography

[1] John Baldwin and Michael Benedikt. Stability theory, permutations of indiscernibles, and embedded finite models. *Trans. Amer. Math. Soc.*, 352(11):4937–4969 (electronic), 2000.

[2] M.A. Taitslin. A general condition for collapse results. *Annals of Pure and Applied Logic*, 113(1–3):323–330, 2001.

[3] A. Blumensath and E. Graedel. Automatic structures. In *Proc. 15th IEEE Symp. on Logic in Computer Science*, pages 51–62, 2000.

[4] M. Benedikt, L. Libkin, T. Schwentick, and L. Segoufin. A model-theoretic approach to regular string relations. In *Proc. 16th IEEE Symp. on Logic in Computer Science*, pages 431–442. IEEE Computer Society Press, 2001.

[5] A. L. Semenov. Logical theories of one-place functions on the natural number series. *Izv. Akad. Nauk SSSR Ser. Mat.*, 47(3):623–658, 1983. Russian.

[6] S.M. Dudakov. Collapse results for expansions of Presburger's arithmetics by one concordant with addition unary function. Russian, 2001.

[7] M.A. Taitslin. Translation results in database theory. In *Complex systems: data processing, simulation, and optimization*, pages 5–23. Tver State University, Tver, 2002. Russian.