J. Inform. Process. Cybernet. **EIK 28** (1992) 4, 213–223
(formerly: Elektron. Inform.verarb. Kybernet.)

Equational Dependencies

By *Joel Berman* and *W. J. Blok*

Abstract: Given a set L of connectives of propositional logic, the class of *equational dependencies* over L is defined; it subsumes various notions of dependency for relational databases considered in the literature. The natural consequence relation for equational dependencies is introduced and is shown to correspond to the relation of congruence generation on the free algebras in an equational class that depends on the set of connectives L involved. Using this fact the computational complexity of the inference problem for equational dependencies is determined for the various possible choices of the set L of connectives.

1. Introduction

Given a set L of connectives of propositional logic we introduce the class of *equational dependencies over* L. If L consists of "conjunction" ($\wedge$) only, the equational dependencies over L are just the familiar functional dependencies. Other classes of dependencies that are subsumed are the classes of *Boolean dependencies* of *Sagiv, Delobel, Parker*, and *Fagin* ([17]), the *strong, weak, and dual* dependencies of *Czédli* ([10]) and *Demetrovics* and *Gyepesi* ([11]), the *positive Boolean dependencies* of *Berman* and *Blok* ([6]), and the class of dependencies considered in *Thalheim* ([19]).

The purpose of the paper is to show that, under some weak assumptions on L, the natural consequence relation over the set of equational dependencies over L corresponds in a precise way to the relation of "congruence generation" over a free algebra in a suitable variety determined by L. This correspondence provides us with a tool to determine the computational complexity of the inference problem for equational dependencies for various choices of L.

In Section 2 we define equational dependencies over a set of connectives L, and explain what it means for a relation R to satisfy the equational dependency $p(A_1, \ldots, A_n) \sim q(A_1, \ldots, A_n)$ where $A_1, \ldots, A_n$ are the attributes of R and p and q are terms built from variables $A_1, \ldots, A_n$ and connectives from L. Section 3 contains a number of examples of equational dependencies. One example discussed is how, for an error-correcting code R, the property that R has minimum distance d can be expressed as an equational dependency that holds for R. Section 4 contains results from universal algebra concerning varieties and quasi-varieties that will be used in the later sections. The consequence relation for equational dependencies is considered in Section 5 and a characterization of consequence in terms of congruence generation on free algebras is presented. The inference problem for equational dependencies and the computational complexity of this problem is the topic of Section 6. The final Section 7 explores in more detail how the familiar notions in the theory of functional dependencies for relational databases correspond, via the theory of equational dependencies, to notions in universal algebra involving free algebras and congruence relations.

The basic notions that are needed from relational database theory and from universal algebra are defined in the paper. For background information on relational databases we refer to *Ullman* ([21]) or *Maier* ([14]), and further details on universal algebra may be found in *Burris* and *Sankappanavar* ([9]) or *McKenzie*, *McNulty* and *Taylor* ([15]). Some of the results of the paper were announced in the survey paper *Berman* ([5]).

2. Definitions and notation

A relation R over a finite set $\{A_1, \ldots, A_n\}$ of attributes is a set of functions with domain $\{A_1, \ldots, A_n\}$. Elements of R are called *tuples* and they are denoted by r, s, or t. The value of r at attribute A_i is written as $r(A_i)$. We let $\mathscr{R} = \mathscr{R}(A_1, \ldots, A_n)$ denote the class of all relations with attribute set $\{A_1, \ldots, A_n\}$. For $k \geqq 0$, $\mathscr{R}_k$ denotes the subclass of $\mathscr{R}$ of all relations that contain at most k tuples. A *valuation* is a function $v: \{A_1, \ldots, A_n\} \to \{T, F\}$. The set of all valuations on $\{A_1, \ldots, A_n\}$ is denoted $\mathrm{val}(A_1, \ldots, A_n)$. For $s, t \in R$, the valuation v_{st} is defined by $v_{st}(A_i) = T$ if $s(A_i) = t(A_i)$ and $v_{st}(A_i) = F$ if $s(A_i) \neq t(A_i)$.

A language $L = \langle f_i : i \in I \rangle$ is a family of finitary operation symbols. For a set X of variables and a language L, the set of *terms* of L over X, denoted $\mathrm{Term}_L(X)$, is the smallest set S such that $X \subseteq S$ and if $t_1, \ldots, t_{n_i} \in S$, then $f_i(t_1, \ldots, t_{n_i}) \in S$. A *connective* is any term of the language $\{\vee, \wedge, \to, \leftrightarrow, \neg, T, F\}$ over the variables $\{x_1, x_2, \ldots\}$.

An *algebra* over the language $L = \langle f_i : i \in I \rangle$ is a pair $\mathbf{B} = \langle B, \langle f_i^{\mathbf{B}} : i \in I \rangle \rangle$, where B is a nonvoid set called the *universe* of $\mathbf{B}$, and $f_i^{\mathbf{B}} : B^{n_i} \to B$ an n_i-ary operation for f_i an n_i-ary operation symbol, $i \in I$. We often write f_i for $f_i^{\mathbf{B}}$ if no confusion is likely. If $\mathbf{B}$ is an algebra over L and $p \in \mathrm{Term}_L(X)$, then p induces in a natural way a *term operation* $p^{\mathbf{B}} : B^n \to B$, also often denoted by p itself. The set of all such term operations will be denoted by $\mathrm{Term}_{\mathbf{B}}(X)$. Typically in this paper we will consider languages consisting of connectives, and algebras with universe $\{T, F\}$; the interpretation of the connectives in the algebra is of course the standard one. For example, if $c(x_1, x_2, x_3)$ is the connective $x_1 \vee (x_2 \wedge x_3)$, and $L = \langle c, T \rangle$, then in the algebra $\mathbf{B} = \langle \{T, F\}, c, T \rangle$ we have $c(F, T, F) = F$. Observe that the join operation on $\{T, F\}$ belongs to $\mathrm{Term}_{\mathbf{B}}(\{x_1, x_2\})$ since $p(x_1, x_2) = c(x_1, x_2, T) \in \mathrm{Term}_L(\{x_1, x_2\})$, and $p^{\mathbf{B}}(x_1, x_2) = x_1 \vee x_2$.

If $p(x_1, \ldots, x_n), q(x_1, \ldots, x_n) \in \mathrm{Term}_L(X)$, and if $\mathbf{B}$ is an algebra over L, then $\mathbf{B}$ is said to *satisfy the identity* $\forall \bar{x}(p \approx q)$ if p and q are equal as term operations on $\mathbf{B}$, i.e. for all $v: \{x_1, \ldots, x_n\} \to B$, $p(v(x_1), \ldots, v(x_n)) = q(v(x_1), \ldots, v(x_n))$. We write $\mathbf{B} \models \forall \bar{x}(p \approx q)$ if $\mathbf{B}$ satisfies the identity $\forall \bar{x}(p \approx q)$. For $v: X \to B$ and for $p(x_1, \ldots, x_n) \in \mathrm{Term}(X)$, $\bar{v}(p)$ denotes the element of B corresponding to $p(v(x_1), \ldots, v(x_n))$. Thus $\mathbf{B} \models \forall \bar{x}(p \approx q)$ if and only if for all $v: X \to B$, $\bar{v}(p) = \bar{v}(q)$ in $\mathbf{B}$.

Let $L = \{\vee, \wedge, \to, \leftrightarrow, \neg, T, F\}$ and $p, q \in \mathrm{Term}_L(\{A_1, \ldots, A_n\})$, where $A_1, \ldots, A_n$ are attributes. An *equational dependency* is an expression of the form $p \sim q$. If $p, q \in \mathrm{Term}_{L'}(\{A_1, \ldots, A_n\})$, where L' is a family of connectives, we also say $p \sim q$ is an *equational dependency over* L'. If $p \sim q$ is an equational dependency in the attributes $A_1, \ldots, A_n$, and $R \in \mathscr{R}(A_1, \ldots, A_n)$, we say R *satisfies* $p \sim q$, written $R \models p \sim q$, if for all s, $t \in R$, $\bar{v}_{st}(p) = \bar{v}_{st}(q)$, i.e. $p(v_{st}(A_1), \ldots, v_{st}(A_n)) = q(v_{st}(A_1), \ldots, v_{st}(A_n))$. Again p and q are evaluated in the two-element Boolean algebra $\{T, F\}$ in the standard way. If Δ is a finite set of equational dependencies in $A_1, \ldots, A_n$, say $\Delta = \{p_j \sim q_j \mid j \in J\}$, and $\mathscr{R}' \subseteq \mathscr{R}(A_1, \ldots, A_n)$, we write $\Delta \models_{\mathscr{R}'} p \sim q$ if, for all $R \in \mathscr{R}'$, $R \models p_j \sim q_j$, $j \in J$, imply $R \models p \sim q$. In the case that $\Delta = \varnothing$, we write $\models_{\mathscr{R}'} p \sim q$ and this means that $R \models p \sim q$ for all $R \in \mathscr{R}'$. For example, $\models_{\mathscr{R}} A_i \sim A_i$ for any A_i in the attribute set.

3. Examples of equational dependencies

We give some examples of equational dependencies. A relation R satisfies the equational dependency $A_1 \sim A_2$, written $R \models A_1 \sim A_2$, if for any $(s, t) \in R^2$, s and t agree in attribute A_1 precisely when s and t agree in attribute A_2. If $\wedge$ and $\vee$ are in the language, then $R \models A_1 \wedge A_2 \sim A_1 \vee A_2$ means that if $(s, t) \in R^2$ and $s(A_1) = t(A_1)$ or $s(A_2) = t(A_2)$, then $s(A_1) = t(A_1)$ and $s(A_2) = t(A_2)$. Note that $R \models A_1 \sim A_2$ if and only if $R \models A_1 \wedge A_2 \sim A_1 \vee A_2$. If T is in the language and if $p(A_1, \ldots, A_n)$ is a term, then $R \models p(A_1, \ldots, A_n) \sim T$ means that for all $s, t \in R^2$, $\bar{v}_{st}(p(A_1, \ldots, A_n)) = T$. The examples given in *Berman* and *Blok* ([6]) can be expressed in this way. That is, if $p(A_1, \ldots, A_n)$ is a term, then $R \models p$ in the sense of *Berman* and *Blok* ([6]) is the same as $R \models p \sim T$ as an equational dependency.

Recall, for X, $Y \subseteq \{A_1, \ldots, A_n\}$ and for $R \in \mathscr{R}(A_1, \ldots, A_n)$, R satisfies the functional dependency $X \to Y$ if whenever two tuples in R agree on every attribute in X, then they agree on every attribute in Y. In *Berman* and *Blok* ([6]) this is expressed as $R \models \bigwedge X \to \bigwedge Y$. In terms of equational dependencies in a language involving $\to$, $\wedge$, and T, this can be expressed $R \models (\bigwedge X \to \bigwedge Y) \sim T$. This functional dependency can also be formulated without use of the connective $\to$ since the functional dependency $X \to Y$ is equivalent to the equational dependency $\bigwedge X \sim \bigwedge X \wedge \bigwedge Y$.

In *Berman* and *Blok* ([6]) several examples of positive Boolean dependencies are given that describe combinatorial configurations. Each of these can also be expressed using equational dependencies.

As another example of equational dependencies we consider Hamming distance. A code R on an alphabet S is a set of n-tuples of elements of S, i.e. $R \subseteq S^n$. Elements $r, s \in R$ have *Hamming distance* k if r and s differ in exactly k coordinates.

Consider the condition that for a given k any two codewords in R have Hamming distance at most k. This means that for every $k + 1$ distinct coordinates $A_{i_1}, \ldots, A_{i_{k+1}}$ and every $(r, s) \in R^2$, r and s must agree on at least one of the A_{i_j}, $1 \leqq j \leqq k + 1$. Thus, $R \models A_{i_1} \vee \ldots \vee A_{i_{k+1}} \sim T$. So R has Hamming distance at most k if and only if $R \models p \sim T$ where p is the term $\bigwedge_{1 \leqq i_1 < \ldots < i_{k+1} \leqq n} A_{i_1} \vee \ldots \vee A_{i_{k+1}}$.

A code R is said to have minimum distance k if every pair of distinct codewords have Hamming distance at least k. The code R has minimum distance k if and only if for every $(s, t) \in R^2$, if s and t agree on more than $n - k$ coordinates, then s and t are identical. Let $d = n - k + 1$. Then R has minimum distance k if for every $i_1 < \ldots < i_d$, $R \models A_{i_1} \wedge \ldots \wedge A_{i_d} \sim A_1 \wedge \ldots \wedge A_n$. This can be expressed by the single equational dependency $R \models p \sim A_1 \wedge \ldots \wedge A_n$ where $p = \bigvee_{1 \leqq i_1 < \ldots < i_d \leqq n} A_{i_1} \wedge \ldots \wedge A_{i_d}$. In Section 5 we consider examples of Hamming distance in our discussion of the inference problem for equational dependencies.

4. Varieties and quasi-varieties

Our definition of an equational dependency $p \sim q$ for $p, q \in \mathrm{Term}(A_1, \ldots, A_n)$ involves the algebraic structure of the algebra $\mathbf{B}$. In this section we present definitions and results from universal algebra that are used in our investigation of equational dependencies.

Let K be a class of algebras over the language $\langle f_i : i \in I \rangle$. The class of all algebras that are homomorphic images of algebras in K is denoted by $H(K)$. The class of all algebras isomorphic to subalgebras of members of K is $S(K)$ and the class of algebras isomorphic to direct products of algebras in K is $P(K)$. By $P_u(K)$ we denote the class of all ultraproducts of algebras in K.

Let $\mathbf{B} = \langle B, \langle f_i : i \in I \rangle \rangle$ be an algebra. The *variety* generated by $\mathbf{B}$ is the class of algebras that are homomorphic images of subalgebras of products of $\mathbf{B}$, i.e., $HSP(\mathbf{B})$. We denote this class by $V(\mathbf{B})$. By a theorem of Birkhoff, V ($\mathbf{B}$) is the same as the class (called the *equational class*) of all algebras over the same language as $\mathbf{B}$ that satisfy all the identities that $\mathbf{B}$ satisfies.

Let $\mathbf{B}$ be an algebra over a language L. A *quasi-identity* in the language of $\mathbf{B}$ is either an identity $\forall \bar{x}(p \approx q)$ for $p, q \in \mathrm{Term}_L(X)$ or a statement of the form $\forall \bar{x}[(p_1 \approx q_1 \wedge \ldots \wedge p_n \approx q_n) \rightarrow p \approx q]$ with $p_1, q_1, \ldots, p_n, q_n, p, q \in \mathrm{Term}_L(X)$, and $\bar{x}$ a list of all $x_i \in X$ that appear in these terms. The quasi-variety generated by $\mathbf{B}$, denoted $Q(\mathbf{B})$, is the class of all algebras of the same similarity type as $\mathbf{B}$ that satisfy all the quasi-identities that $\mathbf{B}$ satisfies. The quasivariety $Q(\mathbf{B})$ can be described as all algebras isomorphic to subalgebras of products of ultrapowers of $\mathbf{B}$, i.e., $Q(\mathbf{B}) = ISPP_u(\mathbf{B})$. Note that always $Q(\mathbf{B}) \subseteq V(\mathbf{B})$; equality need not hold, as we will see below. We are especially interested in equational dependencies that arise from algebras $\mathbf{B}$ for which $Q(\mathbf{B}) = V(\mathbf{B})$.

A nontrivial algebra $\mathbf{A}$ is called *subdirectly irreducible* if whenever h is an isomorphism of $\mathbf{A}$ to a subalgebra of a product of algebras $\mathbf{A}_i$, then there exists an i for which π_i, the projection onto A_i, restricted to $h[A]$ is one-to-one. It is known that for any $\mathbf{A}$, $V(\mathbf{A}) = ISP(V(\mathbf{A})_{SI})$ where $V(\mathbf{A})_{SI}$ denotes all subdirectly irreducible algebras in $V(\mathbf{A})$.

Theorem 4.1. *Let* $\mathbf{A}$ *be a finite algebra. The following are equivalent.*

(i) $V(\mathbf{A}) = Q(\mathbf{A})$.
(ii) $V(\mathbf{A}) = ISP(\mathbf{A})$.
(iii) $V(\mathbf{A})_{SI} \subseteq IS(\mathbf{A})$.

Proof: $Q(\mathbf{A}) = ISPP_u(\mathbf{A})$ and since $\mathbf{A}$ is finite, $ISPP_u(\mathbf{A}) = ISP(\mathbf{A})$. Thus (i) and (ii) are equivalent. If $\mathbf{B} \in V(\mathbf{A})_{SI}$ and (ii) holds, then $\mathbf{B} \in IS(\mathbf{A})$. Finally, if (iii) holds and $B \in V(\mathbf{A})$, then $B \in SP(V(A)_{SI}) \subseteq ISPS(\mathbf{A}) = ISP(\mathbf{A})$, so (iii) $\Rightarrow$ (ii).

Even for 2-element algebras $\mathbf{B}$, there exist examples for which $V(\mathbf{B}) \neq Q(\mathbf{B})$. For example, if $\mathbf{B}_1 = \langle \{T, F\}, \neg \rangle$ with $\neg T = F$, $\neg F = T$, then $V(\mathbf{B}_1)$ contains a 3-element subdirectly irreducible algebra so by Theorem 4.1 $V(\mathbf{B}_1) \neq Q(\mathbf{B}_1)$. *Taylor* ([18]) or *Berman* ([4]) contains a discussion of the subdirectly irreducible algebras in varieties generated by 2-element algebras. If $\mathbf{B} = \langle \{T, F\}, \langle f_i : i \in I \rangle \rangle$ has the property that $f_i(T, \ldots, T) = T$ for all $i \in I$, then it is known that $\mathbf{B}$ is the only subdirectly irreducible algebra in $V(\mathbf{B})$ so $V(\mathbf{B}) = Q(\mathbf{B})$ in this case. Thus if the f_i are drawn from the set of connectives $\{\wedge, \vee, \rightarrow, \leftrightarrow\}$, then the equivalent conditions of Theorem 4.1 hold.

For an algebra $\mathbf{A}$ the set of congruence relations on $\mathbf{A}$ is denoted Con $\mathbf{A}$. The set Con $\mathbf{A}$ forms a complete lattice with respect to the partial order of inclusion. For $a, b \in A$ the least congruence relation containing (a, b) is denoted $\theta(a, b)$. If Q is a quasi-variety containing $\mathbf{A}$ and if $\theta \in$ Con $\mathbf{A}$, then θ_Q is the least congruence on $\mathbf{A}$ containing θ and for which $\mathbf{A}/\theta_Q \in Q$. Since Q is closed under S and P and Con $\mathbf{A}$ is complete, θ_Q exists in Con $\mathbf{A}$. For an algebra $\mathbf{A}$ the free algebra on n free generators in the variety generated by $\mathbf{A}$ is denoted by $\mathbf{F}_{\mathbf{A}}(n)$. Recall that the free algebra has the universal mapping property: if $x_1, \ldots, x_n$ are the free generators of $\mathbf{F}_{\mathbf{A}}(n)$ and if $\mathbf{C} \in V(\mathbf{A})$ and $c_1, \ldots, c_n \in C$, then there exists a

homomorphism h from $\mathbf{F}_{\mathbf{A}}(n)$ into $\mathbf{C}$ for which $h(x_i) = c_i$, $1 \leqq i \leqq n$. Another property of $\mathbf{F}_{\mathbf{A}}(n)$ is that if $\mathbf{A}$ is an algebra over the language L, and $p, q \in \mathrm{Term}_L(x_1, \ldots, x_n)$, $p(x_1, \ldots, x_n) = q(x_1, \ldots, x_n)$ in $\mathbf{F}_{\mathbf{A}}(n)$ if and only if $\mathbf{A} \models \forall \bar{x}(p \approx q)$. We will denote the element $p(x_1, \ldots, x_n)$ of $F_{\mathbf{A}}(n)$ often just by p.

The next theorem is an algebraic result that will be used later in the paper.

Theorem 4.2. *Let* $\mathbf{A}$ *be a finite algebra over a language* L *and let* Q *denote* $Q(\mathbf{A})$. *For arbitrary* $p, q, p_1, \ldots, p_k, q_1, \ldots, q_k \in \mathrm{Term}_L(n)$ *the following are equivalent.*

(i) $p \equiv q \left[\bigvee_{1 \leqq i \leqq k} \theta(p_i, q_i) \right]_Q$ *in* $\mathbf{F}_{\mathbf{A}}(n)$.

(ii) *For all* $\mathbf{C} \in Q$, $\mathbf{C} \models \forall \bar{x} \left[\left[\bigwedge_{1 \leqq i \leqq k} p_i \approx q_i \right] \rightarrow p \approx q \right]$.

(iii) $\mathbf{A} \models \forall \bar{x} \left[\left[\bigwedge_{1 \leqq i \leqq k} p_i \approx q_i \right] \rightarrow p \approx q \right]$.

(iv) *For all* $v : \{x_1, \ldots, x_n\} \rightarrow A$, *if* $\bar{v}(p_i) = \bar{v}(q_i)$, $1 \leqq i \leqq k$, *then* $\bar{v}(p) = \bar{v}(q)$.

Proof: The equivalence of (ii), (iii) and (iv) is easy. Assume (i) and let $c_1, \ldots, c_n \in \mathbf{C} \in Q$ with $p_i(c_1, \ldots, c_n) = q_i(c_1, \ldots, c_n)$ in $\mathbf{C}$, $1 \leqq i \leqq k$. Consider a homomorphism $h : \mathbf{F}_{\mathbf{A}}(n) \rightarrow \mathbf{C}$ with $h(x_i) = c_i$. Then $h(p_i) = h(q_i)$, $1 \leqq i \leqq k$, so $\ker(h) \geqq \left[\bigvee_{1 \leqq i \leqq k} \theta(p_i, q_i) \right]_Q$. Thus $h(p) = h(q)$ and $p(c_1, \ldots, c_n) = q(c_1, \ldots, c_n)$.

To complete the proof, assume (ii) and let γ denote the congruence $\left[\bigvee_{1 \leqq i \leqq k} \theta(p_i, q_i) \right]_Q$ of $\mathbf{F}_{\mathbf{A}}(n)$ and let $\mathbf{C}$ denote $\mathbf{F}_{\mathbf{A}}(n)/\gamma$. Then $p_i(x_1/\gamma, \ldots, x_n/\gamma) = q_i(x_1/\gamma, \ldots, x_n/\gamma)$, for $1 \leqq i \leqq k$, in $\mathbf{C}$, and hence by (ii) $p(x_1/\gamma, \ldots, x_n/\gamma) = q(x_1/\gamma, \ldots, x_n/\gamma)$. Since $p(x_1, \ldots, x_n)/\gamma = p(x_1/\gamma, \ldots, x_n/\gamma)$ and $q(x_1, \ldots, x_n)/\gamma = q(x_1/\gamma, \ldots, x_n/\gamma)$, we conclude $p \equiv q(\gamma)$.

Corollary 4.3. *Let* $\mathbf{A}$ *be a finite algebra over a language* L *such that* $V(\mathbf{A}) = Q(\mathbf{A})$ *and let* $p, q, p_1, \ldots, p_k, q_1, \ldots, q_k \in \mathrm{Term}_L(n)$. The following are equivalent.

(i) $p \equiv q \bigvee_{1 \leqq i \leqq k} \theta(p_i, q_i)$ in $\mathbf{F}_{\mathbf{A}}(n)$.

(ii) *For all* $v : \{x_1, \ldots, x_n\} \rightarrow A$, if $\bar{v}(p_i) = \bar{v}(q_i)$, $1 \leqq i \leqq k$, *then* $\bar{v}(p) = \bar{v}(q)$.

Note that in Corollary 4.3, (i) implies (ii) even without the hypothesis that $V(\mathbf{A}) = Q(\mathbf{A})$. If, however, $V(\mathbf{A}) \neq Q(\mathbf{A})$ then there are an integer n and $p, q, p_1, \ldots, p_k, q_1, \ldots, q_k \in \mathrm{Term}_L(n)$ for which (ii) holds and (i) fails. Indeed any quasi-identity that holds in $\mathbf{A}$ but fails in $V(\mathbf{A})$ will yield a counter example.

5. Equational dependencies and equational logic

We now present algebraic conditions equivalent to the equational dependency conditions of $\mathscr{R} \models p \sim q$ and of $\Delta \models_{\mathscr{R}} p \sim q$.

Let $v : \{A_1, \ldots, A_n\} \rightarrow \{T, F\}$ be an arbitrary valuation. The relation $R_v \in \mathscr{R}(A_1, \ldots, A_n)$ is defined to consist of the tuples $\langle T, \ldots, T \rangle$ and $\langle v(A_1), \ldots, v(A_n) \rangle$. So R_v has at most two tuples. We omit the easy proof of the following useful lemma.

Lemma 5.1. *Let* $\mathbf{A}$ *be an algebra with universe* $\{T, F\}$ *over a language* L *and let* $p, q \in \mathrm{Term}_L(A_1, \ldots, A_n)$. *Let* v *be a valuation* $v : \{A_1, \ldots, A_n\} \rightarrow \{T, F\}$. *Then* $R_v \models p \sim q$ *if and only if* $\bar{v}(p) = \bar{v}(q)$ and $p(T, \ldots, T) = q(T, \ldots, T)$.

Note that if $\{T\}$ is the universe of a subalgebra of $\mathbf{A}$, that is, $p(T, \ldots, T) = T$ for all $p \in \mathrm{Term}_L(X)$, then the lemma simplifies to $R_v \models p \sim q$ if and only if $\bar{v}(p) = \bar{v}(q)$.

The following generalizes Theorem 4 of *Berman* and *Blok* ([6]).

Theorem 5.2. *Let $L = \langle f_i : i \in I\rangle$ be a family of connectives, $\mathbf{B} = \langle\{T, F\}, \langle f_i : i \in I\rangle\rangle$ an algebra over L, and W the equational class generated by $\mathbf{B}$. Let $p, q \in \mathrm{Term}_L(A_1, \ldots, A_n)$ and let $\mathscr{R} = \mathscr{R}(A_1, \ldots, A_n)$. The following are equivalent.*

(i) *$\forall \bar{x}(p \approx q)$ is an identity that holds in W.*
(ii) *$\bar{v}(p) = \bar{v}(q)$ for all valuations v of $\{A_1, \ldots, A_n\}$ into B.*
(iii) *$\mathscr{R} \models p \sim q$.*
(iv) *$\mathscr{R}_2 \models p \sim q$.*

Proof: That (i) and (ii) are equivalent is a standard fact of universal algebra mentioned in Section 3. It is trivial that (iii) implies (iv). If (ii) fails by virtue of some valuation v, then (iv) fails as witnessed by the relation R_v and Lemma 5.1. If (iii) fails, then for some R and some $(s, t) \in R^2$, $\bar{v}_{st}(p) \neq \bar{v}_{st}(q)$. Hence (ii) fails.

A semantic condition for $\mathscr{R} \models p \sim q$ is given by (ii). A syntactic condition is obtained by observing that the equation $\forall \bar{x}(p \approx q)$ holds in W if and only if $\forall \bar{x}(p \approx q)$ can be derived from Γ_B by means Birkhoff's deduction rules, where Γ_B is any set of equations axiomatizing the equational theory of $\mathbf{B}$ (see *Birkhoff* ([7]) or any text on universal algebra). *Lyndon* ([13]) (see also *Berman* ([3]) shows that the equational theory of every 2-element algebra can be axiomatized by a finite number of equations. We use this syntactic condition in the next section to analyze the computational complexity of the problem $\Delta \models_{\mathscr{R}} p \sim q$.

Theorem 5.3. *Let $L = \langle f_i : i \in I\rangle$ be a language of connectives and $\mathbf{B} = \langle\{T, F\}, \langle f_i : i \in I\rangle\rangle$ the corresponding algebra over L. Assume that $f_i(T, \ldots, T) = T$ for all $i \in I$. Let $\{p_j \sim q_j : j \in J\} \cup \{p \sim q\}$ be a finite set of equational dependencies over L and let $\mathbf{F}$ denote the free algebra for $\mathbf{B}$ with free generating set $\{A_1, \ldots, A_n\}$. We write $\Delta = \{p_j \sim q_j : j \in J\}$ and $\mathscr{R} = \mathscr{R}(A_1, \ldots, A_n)$. The following are equivalent.*

(i) *$\Delta \models_{\mathscr{R}} p \sim q$.*
(ii) *$\Delta \models_{\mathscr{R}_2} p \sim q$.*
(iii) *For all $v \in \mathrm{val}(A_1, \ldots, A_n)$, if $\bar{v}(p_j) = \bar{v}(q_j)$ for all $j \in J$, then $\bar{v}(p) = \bar{v}(q)$.*
(iv) $p \equiv q \bigvee_{j \in J} \theta(p_j, q_j)$ *in* $\mathbf{F}$.

Proof: (i) implies (ii) is trivial. Conditions (iii) and (iv) are equivalent by Corollary 4.3, since $\mathbf{B}$ is the only subdirectly irreducible algebra in the variety it generates. To show that (iii) implies (i), let $R \in \mathscr{R}$ with $R \models \Delta$. Then for all $(s, t) \in R^2$, $\bar{v}_{st}(p_j) = \bar{v}_{st}(q_j)$ for all $j \in J$. So by (iii) $\bar{v}_{st}(p) = \bar{v}_{st}(q)$ and $R \models p \sim q$ as desired. To show that (ii) implies (iii), suppose that for a valuation v, $\bar{v}(p_j) = \bar{v}(q_j)$ for all $j \in J$ and $\bar{v}(p) \neq \bar{v}(q)$. Form the relation $R_v \in \mathscr{R}_2$. Lemma 5.1 gives $R_v \models \Delta$ but $R_v \not\models p \sim q$.

We illustrate this theorem with the Hamming distance example of Section 3. For $n = 4$ a relation R has Hamming distance at most 2 if

$$R \models (A_1 \vee A_2 \vee A_3) \wedge (A_1 \vee A_2 \vee A_4) \wedge (A_1 \vee A_3 \vee A_4) \wedge (A_2 \vee A_3 \vee A_4) \sim T.$$

Let A_i^* denote $A_1 \vee \ldots \vee A_{i-1} \vee A_{i+1} \vee \ldots \vee A_4$. Consider the set of equational dependencies:

$$\Delta = \{A_1 \vee A_2 \vee A_3 \vee A_4 \sim T, A_1 \vee A_2 \vee A_3 \vee A_4 \sim A_4^*, \\ A_1 \vee A_2 \vee A_3 \vee A_4 \sim A_3^*, A_3^* \wedge A_4^* \sim A_2^*, A_1^* \wedge A_2^* \sim A_4^*\}$$

The algebra $\mathbf{B}$ can be chosen to be $\langle\{T, F\}, \wedge, \vee, T\rangle$. Let $\mathbf{F}$ be the free algebra freely generated by $\{A_1, A_2, A_3, A_4\}$. Note that $\mathbf{F}$ is a free distributive lattice with a new largest element adjoined. Let γ be the join of the five principal congruences $\theta(p, q)$ on $\mathbf{F}$, where $p \sim q \in \Delta$. It is an easy computation to show $A_1^* \wedge A_2^* \wedge A_3^* \wedge A_4^* \equiv T(\gamma)$, and thus by Theorem 5.3, $\Delta \models_R A_1^* \wedge A_2^* \wedge A_3^* \wedge A_4^* \sim T$. In particular, if R is a relation such that $R \models \Delta$, then R has Hamming distance at most 2.

Similarly if $A_i^\#$ denotes $A_1 \wedge \ldots \wedge A_{i-1} \wedge A_{i+1} \wedge \ldots \wedge A_4$, then R has minimum Hamming distance at least 2 if $R \models A_1^\# \vee A_2^\# \vee A_3^\# \vee A_4^\# \sim A_1 \wedge A_2 \wedge A_3 \wedge A_4$. If

$$\Delta = \{A_4^\# \sim A_1 \wedge A_2 \wedge A_3 \wedge A_4, A_3^\# \sim A_1 \wedge A_2 \wedge A_3 \wedge A_4, \\ A_3^\# \vee A_2^\# \sim A_2^\#, A_2^\# \vee A_1^\# \sim A_4^\#\},$$

then a routine congruence computation and Theorem 5.3 yield $\Delta \models_{\mathscr{R}} A_1^\# \vee A_2^\# \vee A_3^\# \vee A_4^\# \sim A_1 \wedge A_2 \wedge A_3 \wedge A_4$. In particular, if R is a relation such that $R \models \Delta$, then R has minimum Hamming distance at least 2.

How important is the preservation condition, $f_i(T, \ldots, T) = T$ for all $i \in I$, in Theorem 5.3? This hypothesis is used in two places. The first is to guarantee that condition (iii) of Theorem 4.1 holds. The preservation condition is not crucial here, e.g. the 2-element Boolean algebra has operations that do not preserve T but still the 2-element Boolean algebra is the only subdirectly irreducible Boolean algebra. The other use of the preservation condition is in applying Lemma 5.1. The real requirement is that $p_j(T, \ldots, T) = q_j(T, \ldots, T)$ *holds for all* $j \in J$. Thus if $\mathbf{B}$ is the 2-element Boolean algebra and if, say, $\Delta = \{\neg A_1 \sim \neg A_1 \wedge A_1\}$, $p = A_1$, and $q = A_1 \vee \neg A_1$, then all four conditions of Theorem 5.3 hold but the preservation hypothesis fails. If instead $\Delta = \{\neg A_1 \sim A_1 \vee \neg A_1\}$, $p = A_1$, $q = \neg A_1$, then condition (ii) holds vacuously since for no $R \in \mathscr{R}(A_1)$ does $R \models \Delta$ hold. But (iii) fails for the valuation $v(A_1) = F$. A more detailed discussion of the problem is found in *Berman* and *Blok* ([6]).

6. The inference problem for equational dependencies

Let $L = \langle f_i : i \in I\rangle$ be a language of connectives. The *inference problem for equational dependencies over* L is to provide an algorithm that decides for an arbitrary finite set $\Delta \cup \{p \sim q\}$ of equational dependencies over L in $A_1, \ldots, A_n$ whether or not $\Delta \models_{\mathscr{R}(A_1, \ldots, A_n)} p \sim q$.

Condition (iii) in Theorem 5.3 shows that if Δ and $p \sim q$ involve n attributes, then $\Delta \models_{\mathscr{R}} p \sim q$ can be decided by testing at most 2^n valuations. Since each valuation can be tested in polynomial time the time complexity of the inference problem is in co-NP.

In Theorem 5.3, if $\Delta = \varnothing$, then the equivalence of conditions (i) and (iv) reduces the inference problem in this case to the word problem for free algebras in the variety generated by the algebra $\mathbf{B}$. Hence the inference problem for equational dependencies over L is at least as hard as the word problem for finitely generated free algebras in $V(\mathbf{B})$.

Since we are interested in algebras $\mathbf{B}$ that satisfy the hypotheses of Theorem 5.3 we restrict ourselves to algebras $\mathbf{B}$ in which $\{T\}$ is a subuniverse. Also, in order to simplify

the discussion and to reduce the number of algebras to be considered, we assume that the constant operation T is either an explicit operation of $\mathbf{B}$ or is definable, as in say $\mathbf{B} = \langle\{T, F\}, \rightarrow\rangle$, with $x \rightarrow x$ defining T.

Post ([16]) classifies all algebras having two elements with respect to the set of term operations of the algebra. The classification can be pictured by a countable lattice, partially ordered by inclusion among sets of term operations. We are concerned with a convex interval of this lattice drawn in Figure 1. The top element of this interval is the set of all operations on $\{T, F\}$ that preserve T and the bottom element of this interval is the set of operations consisting of all the projection operations and the constant operation T. A convenient set of generating operations for some of these sets of term operations is provided in Figure 1 as a labelling of the algebras. Every set $\mathrm{Term}_{\mathbf{B}}(X_1, X_2, \ldots)$ for an algebra $\mathbf{B} = \langle\{T, F\}, \langle f_i : i \in I\rangle\rangle$ in which $\{T\}$ is a subuniverse and T is a constant operation appears as a vertex in the diagram.

Theorem 6.1. *For the collection of algebras in Figure 1, the word problem for free algebras is co-NP complete for the varieties generated by an algebra above the dotted line, and is*

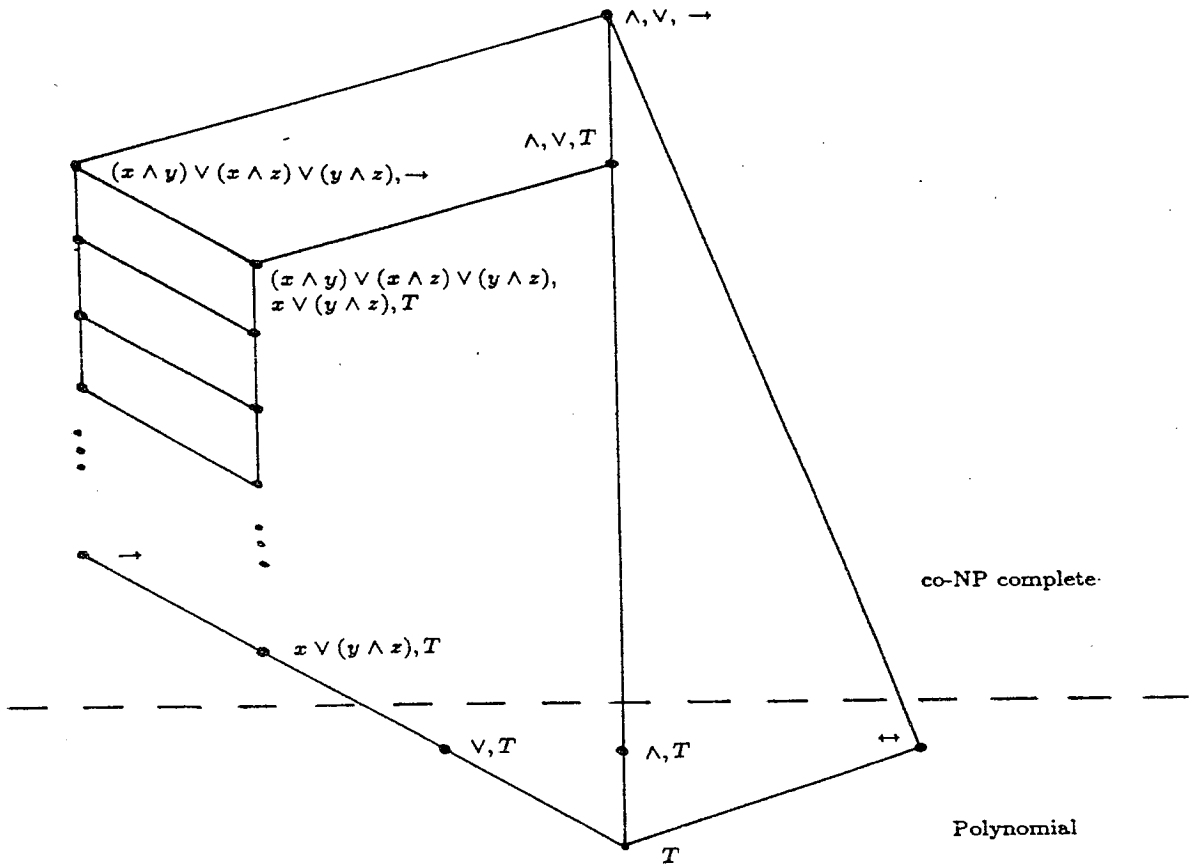

Fig. 1

polynomial for every variety generated by an algebra below the dotted line. In particular the free algebras in the varieties generated by each of $\langle\{T, F\}, T\rangle$, $\langle\{T, F\}, \leftrightarrow\rangle$, $\langle\{T, F\}, \vee, T\rangle$, *and* $\langle\{T, F\}, \wedge, T\rangle$ *have a word problem of polynomial time complexity.*

Proof: That the algebras below the dotted line have polynomial time word problems for free algebras is fairly well known. For example if $p(A_1, \ldots, A_n)$ and $q(A_1, \ldots, A_n)$ are two terms in the operation $\leftrightarrow$, then p and q are equal in the free algebra if and only if the number of occurrences of A_i in p has the same parity as the number of occurrences of A_i in q, $i = 1, 2, \ldots, n$.

For those above the line, *S. Tschantz* ([20]) and *Bloniarz*, *Hunt* and *Rosenkrantz* ([8]) have shown that the variety generated by $\langle\{T, F\}, \wedge, \vee, T\rangle$ has a co-NP complete word problem for free algebras. (One way to do this is to reduce the problem of $p \neq q$ to ONE-IN-THREE 3SAT of *Garey* and *Johnson* ([12]), p. 259.) So it suffices to consider $\mathbf{B}_1 = \langle\{T, F\}, x \wedge (y \vee z), T\rangle$. Let $\mathbf{F}_1$ be the free algebra for $V(\mathbf{B}_1)$ with free generators $A_0, A_1, \ldots, A_n$ and let $\mathbf{F}_2$ be the free distributive with free generators $A_1, A_2, \ldots, A_n$. Define a function g from F_2 to F_1 inductively by

$$g(A_i) = A_0 \vee (A_i \wedge A_i),$$
$$g(p_1 \wedge p_2) = A_0 \vee (g(p_1) \wedge g(p_2))$$
$$g(p_1 \vee p_2) = g(p_1) \vee (g(p_2) \wedge T).$$

Then it is easily verified that $g(p)$ and $A_0 \vee p$ are the same as functions from $\{T, F\}^{n+1} \to \{T, F\}$. Hence g is a 1-1 function from F_2 into F_1. If p has length k then $g(p)$ has length at most $7k$. The transformation p to $g(p)$ is Turing computable so an algorithm for the word problem for free algebras in $V(\mathbf{B}_1)$ could be used for distributive lattices. So $V(\mathbf{B}_1)$ has a co-NP complete word problem for free algebras.

Corollary 6.2. *The inference problem for equational dependencies is co-NP complete for those based on algebras above the dotted line in Figure 1, and is polynomial for those based on algebras below.*

Proof: Since the inference problem is at least as hard as the word problem for free algebras, the co-NP complete result follows from Theorem 5.3 and Theorem 6.1.

For algebras below the line, first consider $\mathbf{B} = \langle\{T, F\}, T\rangle$. A typical instance of the inference problem is, by Theorem 5.3, equivalent to deciding if $A_i \equiv A_j \bigvee_{1 \leqq k \leqq m} \theta(A_{i_k}, A_{j_k})$. This is equivalent to the question of whether (i, j) is in the transitive closure of the symmetric reflexive closure of the relation $\{(i_k, j_k) \mid 1 \leqq k \leqq m\}$. Standard transitive closure algorithms give efficient polynomial time algorithms for this.

For $\mathbf{B} = \langle\{T, F\}, \leftrightarrow\rangle$ the algebra $\mathbf{B}$ is polynomially equivalent to the two element group $\mathbf{G} = \langle\{0, 1\}, +, 0\rangle$. In this group, $p \equiv q(\theta)$ if and only if $0 \equiv p + q(\theta)$. So a typical instance of the inference problem is equivalent to deciding if $q = A_{i_1} + \ldots + A_{i_k}$ is in the normal subgroup generated by a set of terms, say, $p_1, \ldots, p_m$. This in turn is equivalent to solving the linear equation $q = c_1 p_1 + \ldots + c_m p_m$ with $c_i \in \{0, 1\}$ over the 2-element field. Standard polynomial time algorithms exist for this problem.

Finally, for the two semilattice types, we have already seen how the inference problem here is the same as the inference problem for functional dependencies. *Beeri* and *Bernstein* ([2]) contains an $O(n)$ algorithm for this problem.

7. Functional dependencies as equational dependencies

For $\mathbf{B} = \langle \{T, F\}, \wedge \rangle$ the equational dependencies using terms from $\mathbf{B}$ correspond in a very precise way to the familiar functional dependencies. For two sets $X, Y \subseteq \{A_1, \ldots, A_n\}$, $R \models \bigwedge X \sim \bigwedge Y$ means that for all $s, t \in R$, $s(A) = t(A)$ for all $A \in X$ if and only if $s(A) = t(A)$ for all $A \in Y$. In the usual notation of functional dependencies this is the same as $R \models X \to Y$ and $R \models Y \to X$. The functional dependency $X \to Y$ is equivalent to the equational dependency $\bigwedge X \sim (\bigwedge X \wedge \bigwedge Y)$. This equivalence can be made more formal: Let $S = \{A_1, \ldots, A_n\}$ be a set of attributes and $\Gamma = \{X_1 \to Y_1, \ldots, X_k \to Y_k\}$ be an arbitrary set of functional dependencies over S. Let Γ^+ denote the family of all functional dependencies that are consequences of Γ. Many algorithms have been given for deriving Γ^+ from Γ. The original rules of *Armstrong* ([1]) are that Γ^+ is the smallest class of functional dependencies containing Γ and such that the following four rules are satisfied:

(i) $X \to X \in \Gamma^+$ for all $X \subseteq \{A_1, \ldots, A_n\}$.
(ii) If $X \to Y$, $Y \to Z \in \Gamma^+$, then $X \to Z \in \Gamma^+$.
(iii) If $X \subseteq X'$, $Y' \subseteq Y$, and $X \to Y \in \Gamma^+$, then $X' \to Y' \in \Gamma^+$.
(iv) If $X \to Y$, $X' \to Y' \in \Gamma^+$, then $X \cup X' \to Y \cup Y' \in \Gamma^+$.

Let $\mathbf{F}$ denote the free algebra for the variety generated by $\langle \{T, F\}, \wedge \rangle$, with free generators $A_1, \ldots, A_n$, and let γ be the relation on F given by $(\bigwedge X, \bigwedge Y) \in \gamma$ if and only if $X \to X \cup Y$ and $Y \to X \cup Y$ are in Γ^+. The rules (i), (ii), (iii), and (iv) suffice to show that γ is a congruence relation on $\mathbf{F}$ generated by the k pairs $(\bigwedge X_i, \bigwedge X_i \wedge \bigwedge Y_i)$, $1 \leq i \leq k$. Conversely, let θ be an arbitrary congruence relation on $\mathbf{F}$. Define a binary relation $\to$ on the powerset of S by $X \to Y$ if $(\bigwedge X, \bigwedge X \wedge \bigwedge Y) \in \theta$. Using the fact that θ is a congruence relation it is easy to show that rules (i), (ii), (iii) and (iv) hold for $\to$.

Theorem 5.3 establishes a correspondence between equational dependencies over the language of $\mathbf{B}$ and congruence relations on a free algebra in $V(\mathbf{B})$. If $\mathbf{B} = \langle \{T, F\}, \wedge \rangle$, then the equational dependencies are essentially the functional dependencies. In this case, familiar concepts involving functional dependencies translate to standard lattice-theoretic notions involving congruence relations. In Table 1 we provide a lexicon for this translation. Note that the congruence relation concepts make sense for any algebra, not just the algebra $\langle \{T, F\}, \wedge \rangle$. In this table, $X_j, Y_j \subseteq \{A_1, \ldots, A_n\}$ for $1 \leq j \leq k$, $D = \{X_1 \to Y_1, \ldots, X_k \to Y_k\}$ is a set of functional dependencies, $\Delta = \{p_1 \sim q_1, \ldots, p_k \sim q_k\}$ is the corresponding set of equational dependencies over $\mathbf{B} = \langle \{T, F\}, \wedge \rangle$, and $\theta(\Delta) = \bigvee_{1 \leq j \leq k} \theta(p_j, q_j)$ in $f_{\mathbf{B}}(A_1, \ldots, A_n)$. Page numbers refer to *Maier* ([14]).

Table 1

Functional Dependency Condition	Equational Dependency Condition
D derives $X \to Y$ (p. 51)	$p \equiv q\theta(\Delta)$
D_1 and D_2 are equivalent (p. 71)	$\theta(\Delta_1) = \theta(\Delta_2)$
$D_1 \models D_2$ (p 72)	$\theta(\Delta_1) \leq \theta(\Delta_2)$
D is nonredundant (p. 72)	$\theta(\Delta)$ is a nonredundant join of the $\theta(p_j, q_j)$
D is canonical (p. 75)	$\theta(\Delta)$ is an irredundant join of maximal join irreducibles $\theta(p_j, q_j)$
D is a minimum (p. 79)	$\theta(\Delta)$ is the join $\theta(p_1, q_1) \vee \ldots \vee \theta(p_k, q_k)$, k minimal

References

[1] *Armstrong, W. W.:* Dependency Structures of data base relationships. IFIP Congress 1974, Geneva, pp. 580—583.
[2] *Berri C., P. A. Bernstein:* Computational problems related to the design of normal form relational schemas. ACM Trans. Database System 4 (1979), 30—59.
[3] *Berman, J.:* A proof of Lyndon's finite basis theorem. Discrete Math. **29** (1980), 229—233.
[4] *Berman, J.:* Algebraic properties of k-valued logics. Proc. of Tenth International Symposium on Multiple-valued Logic. IEEE 1980, pp. 195—204.
[5] *Berman, J.:* The value of free algebras, in Algebraic Logic and Universal Algebra in Computer Science. In: Lecture Notes in Comput. Sci. **425**, 1990, pp. 15—26.
[6] *Berman, J., W. J. Blok:* Positive Boolean dependencies, Inform. Process. Lett. **27** (1988), 147—150.
[7] *Birkenhoff, G.:* On the structure of abstract algebras. Proc. Cambridge Phil. Soc. **31**, 1935, pp. 433—454.
[8] *Bloniarz, P. A., H. B. Hunt III, D. J. Rosenkrantz:* Algebraic structures with hard equivalence and minimization problems. J. ACM, **31** (1984), 879—904.
[9] *Burris, S., H. P. Sankappanavar:* A Course in Universal Algebra. Graduate Texts in Mathematics, vol 78, Springer-Verlag 1981.
[10] *Czédli, G.:* On dependencies in the relational model of data. Elektron Inform.verarb. Kybernet. **17** (1981), 103—112.
[11] *Demetrovics, J., Gy. Gyepesi:* Some generalized type functional dependencies formalized as equality set on matrices. Discrete Appl. Math. **6** (1983), 35—47.
[12] *Garey M. R., D. S. Johnson:* Computers and Intractability. W. H. Freeman C. 1979.
[13] *Lyndon, R.:* Identities in two-valued calculii. Trans. Amer. Math. Soc. **71** (1951), 457—465.
[14] *Maier, D.:* The Theory of Relational Databases. Computer Science Press 1983.
[15] *McKenzie, R., G. McNulty, W. Taylor:* Algebras, Lattices, Varieties, vol. 1, Wadsworth 1987.
[16] *Post, E. L.:* The Two-valued Iterative Systems of Mathematical Logic. Annals of Mathematics Studies, No. 5, Princeton University Press 1941.
[17] *Sagiv, Y., C. Delobel, D. S. Parker, R. Fagin:* An equivalence between relational database dependencies and a fragment of propositional logic. J. ACM **28** (1981), 435—453. Corrigendum J. ACM **34** (1987), 1016—1018.
[18] *Taylor, W.:* The fine spectrum of a variety. Algebra Universalis **5** (1975), 263—303.
[19] *Thalheim, B.:* Funktionale Abhängigkeiten in relationalen Datenstrukturen. Elektron. Inform.verarb. u. Kybernet. **21** (1985), 23—33.
[20] *Tschantz, S.:* Private communication, 1984.
[21] *Ullman, J. D.:* Principles of Database Systems. 2nd ed., Computer Science Press 1983.

(Received: October 18, 1991

Authors' address:
J. Berman
W. J. Blok
University of Illinois at Chicago
Department of Mathematics, Statistics and Computer Science
Box 4348 (M/C 249)
Chicago, Illinois 60680 USA