

MIT/LCS/TM-193

ALGEBRAIC DEPENDENCIES

Mihalis Yannakakis  
Christos H. Papadimitriou

February 1981

# ALGEBRAIC DEPENDENCIES

*Mihalis Yannakakis*

Bell Laboratories  
Murray Hill, New Jersey 07974

*Christos H. Papadimitriou\**

Laboratory for Computer Science  
Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139

## ABSTRACT

We propose a new kind of data dependencies called algebraic dependencies, which generalize all previously known kinds. We give a complete axiomatization of algebraic dependencies in terms of simple algebraic rewriting rules. In the process we characterize exactly the expressive power of tableaux, thus solving an open problem of Aho, Sagiv and Ullman; we show that it is NP-complete to tell whether a tableau is realizable by an expression; and we give an interesting dual interpretation of the chase procedure. We also show that algebraic dependencies over a language augmented to contain union and set difference can express arbitrary domain-independent predicates of finite index over finite relations. The class of embedded implicational dependencies recently — and independently — introduced by Fagin is shown to coincide with our algebraic dependencies. Based on this, we give a simple proof of Fagin's Armstrong relation theorem.

KEYWORDS: Relational database model, data dependencies, functional, multivalued, transitive, join, template, and algebraic dependencies, embedded implicational dependencies, tableaux, extended relations, complete axiomatization, project-join expressions, chase.

# ALGEBRAIC DEPENDENCIES

*Mihalis Yannakakis*

Bell Laboratories  
Murray Hill, New Jersey 07974

*Christos H. Papadimitriou\**

Laboratory for Computer Science  
Massachusetts Institute of Technology  
Cambridge, Massachusetts 02139

## 1. INTRODUCTION

The relational model for databases [Codd 1970, Ullman 1979] has gained recognition as a valuable formal framework for understanding the semantics, design, and even implementation, of databases. At the heart of the research on relational databases lies the notion of *data dependency*. Data dependencies are domain-independent (i.e., invariant under consistent renamings of domain elements) predicates on databases. Starting with functional [Armstrong 1974] and multivalued [Fagin 1977] dependencies, a dozen of different kinds of data dependencies have been proposed in the literature [Nicolas 1978, Paradaens 1979, Sagiv and Walecka 1979, and others]. New, more and more general, kinds of data dependencies have been put forward in a rather arbitrary and heuristic fashion. This reflected two major frustrations of the research in this area: First, no natural, stable closure of this process was in sight. Secondly, the elegant complete axiomatizations of functional [Armstrong 1974] and multivalued dependencies [Beeri et al. 1977] did not appear to carry over to the more general kinds; thus the further generalizations were futile attempts at "enriching the language" enough so as to obtain a complete axiomatization.

Two important ideas that appeared to point towards a unified theory are the *tableaux* of [Aho et al. 1979], and the related concept of the *chase* [Maier et al. 1979] as a proof system for data dependencies. The tableaux, however, were introduced as models of queries. They were known to be strictly more powerful than the algebraic system that motivated them, and their exact power remained a mystery. Also, the chase was applied in a rather narrow way to functional and join dependencies, as a strictly combinatorial process. No connections to the underlying algebraic system were revealed.

More recently, [Sadri and Ullman 1980] proposed a new kind of data dependencies, the *template dependencies*. Template dependencies generalized most known data dependencies. They are defined in terms of tableaux, and as a consequence the rules of the chase provide an adequate axiomatization for them. However, template dependencies failed to model the functional dependencies, in some sense the most natural and fundamental kind. This inadequacy dramatized the fact that equality had been missing from most attempts at generalizing the notion of data dependencies. It was this absence of equality that caused an annoying dichotomy between the treatment of functional dependencies on the one hand, and that of multivalued dependencies and their relatives on the other.

In this paper we outline some new ideas and results that appear to comprise definitive positive answers to the main quests and open problems of the theory of data dependencies, as exposed above. We introduce a new kind of data dependency, the *algebraic dependency*. This dependency is a natural generalization of all data dependencies existing in the literature (*including* the functional



dependencies) and is stated as an algebraic equation with operations projection and join. We achieve this unified treatment of functional dependencies with other data dependencies by considering *extended* relations, i.e., relations with arbitrarily many copies of each column. Because of its generality and simplicity, the algebraic dependency is a stable, natural concept. We present several pieces of evidence to this effect. We show that algebraic dependencies are equivalent in expressive power to tableaux — thus solving the open problem in [Aho et al. 1979] — and to algebraic equations with *equijoins* — an operator long forgotten since [Codd 1972]. More importantly, we show that deductions of algebraic dependencies are axiomatized by an extremely simple and natural set of algebraic axioms. All past proven (or conjectured) axiomatizations of data dependencies are derived as tedious special cases from ours. To further reinforce the belief that algebraic identities are a natural way of stating data dependencies, we show that *all domain-independent predicates* of finite index over data-bases can be expressed as algebraic identities, with union and difference allowed in addition to projection and join.

Our proof of the completeness of our axiomatic system is quite involved, and proceeds in several stages. It entails understanding the expressive power of tableaux, algebraic tautologies, and also an algebraic interpretation of the chase. It has some interesting side-products. For example, we exhibit two algebraic expressions which, although very different in structure, have the same tableau. We also show that the embedded join dependencies (EJD) are *deductively complete*, in the sense that any algorithm for testing whether a set of EJD's implies another EJD can be modified to work for general algebraic dependencies — thus theoretically justifying the apparent difficulty in obtaining such an algorithm.

It is well-known (e.g., [Nicolas, 1978]) that data dependencies can be expressed in a fragment of first-order logic. This fragment has equality, one relation symbol  $-R-$  of arity  $|a(R)|$ , and typed variables. Independently of the authors, Fagin [Fagin 1980] studied a further fragment of first-order logic, which consists roughly of Horn clauses quantified in the  $\forall\exists$  fashion. Fagin called this fragment of first-order logic *embedded implicational dependencies*, and showed that it generalizes all previously proposed kinds of data dependencies. Fagin showed that sets of embedded implicational dependencies are invariant under a version of the Cartesian product. Based on this, he went on to prove that any set of embedded implicational dependencies possesses an *Armstrong relation*; that is, a universal counterexample to any non-valid implication. Fagin's proof of this result is quite complex, and invokes certain results from logic. Fagin did not provide a complete axiomatization of his class.

Surprisingly, we show that the algebraic dependencies defined in this paper coincide with the embedded implicational dependencies of Fagin. This testifies to the naturalness of our class. Furthermore, the main result of [Fagin 1980] — the existence of an Armstrong relation — follows very easily using our algebraic approach (see Section 6).

The remaining of this paper is organized as follows: In Section 2 we introduce an axiomatic system for expression identities, which is complete for simple expressions. In Section 3 we introduce extended relations, and prove the equivalence between project-join expressions over extended relations, project-equijoin expressions, and tableaux. In Section 4 we introduce algebraic dependencies and an axiom capturing the semantics of extended relations. We show that the axiomatic system of Section 2 together with this axiom comprise a complete axiomatization of algebraic dependencies. This relies heavily on the results of Sections 2 and 3. In Section 5 we show constructively that algebraic dependencies with project, join, union and set difference can express arbitrary domain-independent predicates with finite index. Finally, in Section 6 we study the relation of algebraic to embedded implicational dependencies.



## 2. EXPRESSIONS OVER PROJECTION AND JOIN

A relation  $R$  is a table. Its columns correspond to *attributes*; the set of attributes of  $R$ ,  $a(R)$ , is a subset of a finite set (called the *universe*),  $U = \{A, B, C, \dots\}$ . The rows of  $R$  are called *tuples*. The attributes  $A, B, \dots$  have disjoint *domains*  $D(A), D(B), \dots$ . Thus  $R \subseteq \prod_{A \in a(R)} D(A)$ . If  $X \subseteq a(R)$ , and  $t \in R$ ,  $t_X$  is  $t$  restricted to columns of  $X$ . The *projection*  $\pi_X(R) = \{t_X : t \in R\}$ . The (*natural*) *join* is  $R_1 \bowtie R_2 = \{t \in \prod_{A \in a(R_1) \cup a(R_2)} D(A) : t_{a(R_1)} \in R_1, \text{ and } t_{a(R_2)} \in R_2\}$ .

We shall deal with *expressions* over projection and join involving the variable  $R$  ranging over relations on  $U$ . If  $\phi_1$  and  $\phi_2$  are expressions, then  $\phi_1(R) \subseteq \phi_2(R)$  denotes the identity inclusion, implicitly quantified over all  $R$ . This has meaning only if  $a(\phi_1(R)) = a(\phi_2(R))$ .  $\phi_1 = \phi_2$  means  $\phi_1 \subseteq \phi_2$  and  $\phi_1 \supseteq \phi_2$ . We are interested in devising a complete axiomatization of the deductive theory of sentences of the form  $\phi_1 \subseteq \phi_2$ , where  $\phi_1$  and  $\phi_2$  are project-join expressions over a single relational variable  $R$ . We shall be interested in axioms that are *algebraic* in nature, that is, they are rules that either modify expressions syntactically (e.g., commutativity, associativity, etc.) or state that a sentence implies a syntactic variant (e.g., monotonicity). In addition to the ordinary *modus ponens*

$$\frac{A \Rightarrow B}{A}$$

we also employ the transitivity of set inclusion as a deductive tool.

One important desirable feature of the axioms considered is that their applicability can be decided in polynomial time by tree isomorphism techniques. This should be a feature of any "reasonable" axiomatic system. A second positive property sought is that the axioms be reasonably "syntactic" and "local", in the sense that they should be stated in terms of local pattern matching on the expression tree, and not reflect global or semantic considerations. The axioms A1 through A7 that we are proposing below satisfy these criteria. Furthermore, they can be easily rendered to the format  $\sigma_1 \wedge \sigma_2 \wedge \dots \wedge \sigma_k \Rightarrow \sigma_{k+1}$  (where  $\sigma_1, \dots, \sigma_{k+1}$  are sentences and  $\sigma_{k+1}$ 's syntax depends in a straightforward way on that of  $\sigma_1, \dots, \sigma_k$ ) familiar from previous work on dependency theory [Armstrong 1974, Beeri et al. 1977, Sagiv and Walecka 1979].

It is not hard to see that projection and join satisfy the following identities for all  $R_1, R_2$  and  $R_3$  (recall that by writing  $\pi_X(R_1)$  we are implicitly requiring that  $X \subseteq a(R_1)$ ). Similarly,  $R_1 \subseteq R_2$  assumes that  $a(R_1) = a(R_2)$ .

- A1. (Idempotency of Projection)
  - (a)  $\pi_X(\pi_X(R_1)) = \pi_X(R_1)$ .
  - (b)  $\pi_{a(R_1)}(R_1) = R_1$ .
- A2. (Idempotency of Join)
  - (a)  $R_1 \bowtie \pi_X(R_1) = R_1$ .
  - (b)  $\pi_{a(R_1)}(R_1 \bowtie R_2) \subseteq R_1$ .
- A3. (Monotonicity of Projection)
  - $R_1 \subseteq R_2 \Rightarrow \pi_X(R_1) \subseteq \pi_X(R_2)$ .
- A4. (Monotonicity of Join)
  - $R_1 \subseteq R_2 \Rightarrow R_1 \bowtie R_3 \subseteq R_2 \bowtie R_3$ .
- A5. (Commutativity of Join)
  - $R_1 \bowtie R_2 = R_2 \bowtie R_1$ .
- A6. (Associativity of Join)
  - $(R_1 \bowtie R_2) \bowtie R_3 = R_1 \bowtie (R_2 \bowtie R_3)$ .
- A7. (Distributivity of Projection over Join)
  - Let  $X \subseteq a(R_1)$ .

- (a)  $\pi_{X \cup Y}(R_1 \bowtie R_2) \subseteq \pi_{X \cup Y}(R_1 \bowtie \pi_Y(R_2))$ .
- (b) If  $a(R_1) \cap a(R_2) \subseteq Y$  then equality holds in (a).

Axioms A1-A6 hardly need any discussion, since they follow directly from the definitions of the two operations. Axiom A7, the only one that is not totally trivial, simply states that projecting one operand of a join may restrict the common attributes of the two operands, and therefore enrich the result of the join. A7(b) says that if, nevertheless, the common attributes remain the same despite the projection, then the result of the join remains unaffected. We have

*Proposition 2.1.* Axioms A1-A7 are sound.  $\square$

To illustrate the application of the axioms, we will give two examples. In the first one we derive a basic property of project-join expressions which we will use later on. The second example shows how the pseudotransitivity rule for multivalued dependencies can be derived from the axioms.

*Example 2.1* For all expressions  $\phi$ ,  $\pi_{a(\phi)}(R) \subseteq \phi(R)$  (1).

We prove this property by induction on the structure of  $\phi$ . For the basis,  $\phi = R$ , and (1) follows from axiom A1b. For the inductive step, assume that (1) can be derived from the axioms for all expressions  $\sigma$  with fewer operations than  $\phi$ .

*Case 1*  $\phi = \pi_X \sigma$ , for some set of attributes  $X (= a(\phi))$  and some expression  $\sigma$ .

From the inductive hypothesis,  $\pi_{a(\sigma)}(R) \subseteq \sigma(R)$  is derived from the axioms. From A3 we have  $\pi_X(\pi_{a(\sigma)}(R)) \subseteq \pi_X \sigma(R) = \phi(R)$ , and from A1 we get  $\pi_{a(\phi)}(R) \subseteq \phi(R)$ .

*Case 2*  $\phi = \sigma_1 \bowtie \sigma_2$ , for some expressions  $\sigma_1, \sigma_2$ , with  $a(\phi) = a(\sigma_1) \cup a(\sigma_2)$ .

From the inductive hypothesis,  $\pi_{a(\sigma_1)}(R) \subseteq \sigma_1(R)$  and  $\pi_{a(\sigma_2)}(R) \subseteq \sigma_2(R)$ . We have now:

$$\begin{aligned} \pi_{a(\phi)}(R) &= && \text{, by A2, A3} \\ \pi_{a(\phi)}(R \bowtie R) &\subseteq && \text{, by A7a} \\ \pi_{a(\phi)}(\pi_{a(\sigma_1)}(R) \bowtie \pi_{a(\sigma_2)}(R)) &\subseteq && \text{, by A3, A4 and i.h.} \\ \pi_{a(\phi)}(\sigma_1(R) \bowtie \sigma_2(R)) &= && \text{, by A1b} \\ \sigma_1(R) \bowtie \sigma_2(R) &= \phi(R) . && \square \end{aligned}$$

*Example 2.2* Let us show how we can derive the pseudotransitivity property of multivalued dependencies [Beeri et al. 1977]. This property states that, if  $X, Y, Z \subseteq U$  then

$$X \twoheadrightarrow Y, Y \twoheadrightarrow Z \text{ imply } X \twoheadrightarrow Z - Y ,$$

or, in algebraic terms,

$$\begin{aligned} XY \bowtie X(U - XY) &\subseteq R \text{ and} \\ YZ \bowtie Y(U - YZ) &\subseteq R \text{ imply} \\ X(Z - Y) \bowtie XYW &\subseteq R \text{ where } W = U - XYZ. \end{aligned}$$

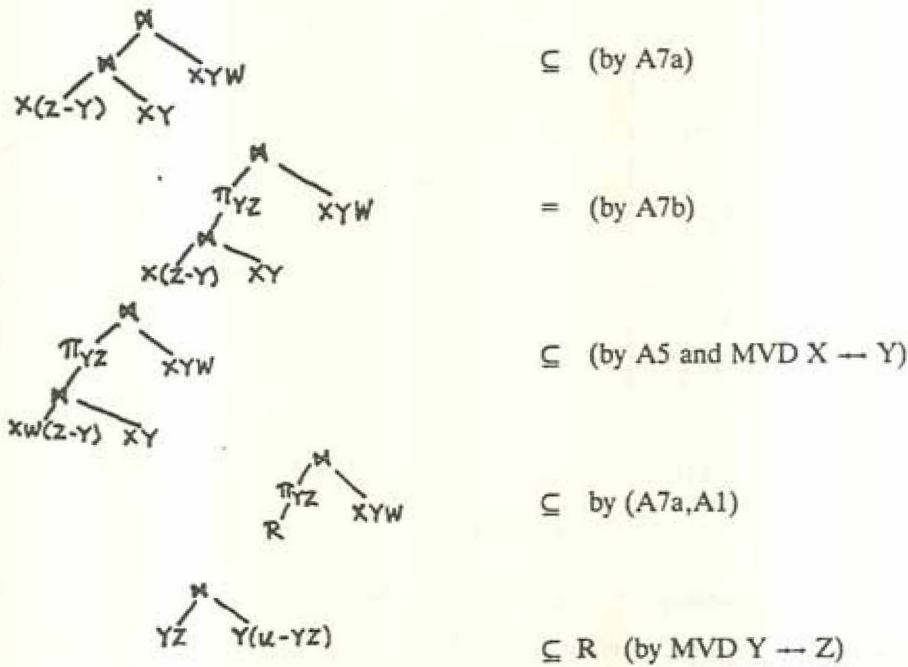
(As is customary, union of sets of attributes is represented by concatenation, and we denote  $\pi_S(R)$  by  $S$ ).

The sequence of applications of the axioms establishing the implication is shown below. The expressions are shown as trees.

$$\begin{array}{c} \begin{array}{c} \bowtie \\ / \quad \backslash \\ X(Z-Y) \quad XYW \end{array} \end{array} = \text{ (by A2a)}$$

$$\begin{array}{c} \begin{array}{c} \bowtie \\ / \quad \backslash \\ X(Z-Y) \quad \begin{array}{c} \bowtie \\ / \quad \backslash \\ XY \quad XYW \end{array} \end{array} \end{array} = \text{ (by A6)}$$





Monotonicity (axioms A3 and A4) are implicitly used almost at every step. Although the proof of this simple fact looks *ad hoc*, and quite formidable, we shall describe in Section 4 a systematic procedure for producing such derivations.  $\square$

Do these properties completely axiomatize project-join identities? The answer is "no", but for very subtle reasons. To understand why, we will have to introduce *tableaux*. A tableau  $T$  is a mapping from relations to relations — a fragment of first-order logic, see [Aho et al. 1979]. For each  $A \in U$  we define its *standard domain*  $\bar{D}(A) = \{A, a_1, a_2, \dots\}$ .  $A$  is called the *distinguished symbol* of  $\bar{D}(A)$ ;  $a_1, a_2, \dots$  are called *nondistinguished*. A tableau  $T$  over  $U$  is a finite relation  $T \subseteq \bar{D}(A) \times \bar{D}(B) \times \dots \times \bar{D}(Z)$ . For example,  $T = \{(a_1, B, c_1), (A, b_1, c_1), (a_2, b_2, c_2), (a_2, B, c_2), (A, B, c_2)\}$  is a tableau over  $\{A, B, C\}$ . We represent a tableau as shown in Figure 1a. The top row, called the *summary*  $s(T)$  of  $T$ , contains all distinguished symbols appearing in  $T$ , each in the corresponding column. The set  $a(T)$  of attributes of  $T$  is the set of attributes in which  $T$  has a distinguished symbol. Tableaux represent mappings from relations to relations. Let  $R \subseteq \bar{D}(A) \times \dots \times \bar{D}(Z)$  be a relation with  $a(R) = U$ . A *valuation*  $\rho$  is a mapping from  $\bar{D}(A)$  to  $D(A)$  for each  $A \in U$ . Then the mapping  $f_T$  corresponding to the tableau  $T$  is defined by

$$f_T(R) = \{\rho(s(T)) : \rho(w) \in R \text{ for all } w \in T\},$$

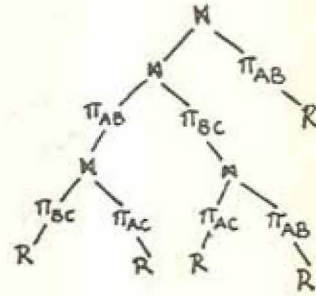
where  $\rho$  is extended to act on vectors in a componentwise manner. It turns out that every expression  $\phi(R)$  over projection and join, when considered as a mapping from relations to relations, has a corresponding tableau  $T_\phi$  such that, for all  $R$ ,  $f_{T_\phi}(R) = \phi(R)$ .  $T_\phi$  is constructed as follows:

1.  $T_R = \{(A, B, \dots, Z)\}$
2.  $T_{\pi_X(\phi)} = T_\phi$ , with all occurrences of each distinguished symbol in  $U-X$  replaced by a new nondistinguished symbol
3.  $T_{\phi_1 \bowtie \phi_2} = T_{\phi_1} \cup T_{\phi_2}$ .

For example, the tableau of Figure 1(a) can be readily seen to be  $T_\phi$ , for the expression  $\phi$  of Figure 1(b). Unfortunately, it is shown in [Aho et al. 1979] that not all tableaux are  $T_\phi$  for an appropriate  $\phi$ . In fact, we shall soon show that it is NP-complete to recognize those that do. Naturally, if  $T_{\phi_1} = T_{\phi_2}$  then  $\phi_1(R) = \phi_2(R)$  is tautologically true. Under what circumstances is  $f_{T_1}(R) \subseteq f_{T_2}(R)$  tautologically true? Let  $h$  be a set of mappings from  $\bar{D}(A)$  to  $\bar{D}(A)$  for each  $A \in U$ , such that  $h(A) = A$  and  $h(T_2) \subseteq T_1$ . Then  $h$  is called a *homomorphism* from  $T_2$  to  $T_1$ . [Aho et al. 1979] show the following Lemma:

A	B	-
$a_1$	B	$c_1$
A	$b_1$	$c_1$
$a_2$	$b_2$	$c_3$
$a_2$	B	$c_2$
A	B	$c_3$

(a)



(b)

Figure 1

Lemma 2.1  $f_{T_1}(R) \subseteq f_{T_2}(R)$  is tautologically true iff there is a homomorphism from  $T_2$  to  $T_1$ .  $\square$

We now return to the question, whether axioms A1 to A7 are sufficient for proving expression identities of the form  $\phi_1(R) \subseteq \phi_2(R)$ . It follows from the above discussion that, besides the Axioms A1-A7 the following is undoubtedly true

$$T: \text{If } T_{\phi_1} = T_{\phi_2}, \text{ then } \phi_1(R) = \phi_2(R).$$

It turns out that, surprisingly,  $T$  is independent of A1-A7. To see this, consider the two expressions shown in Figure 2.

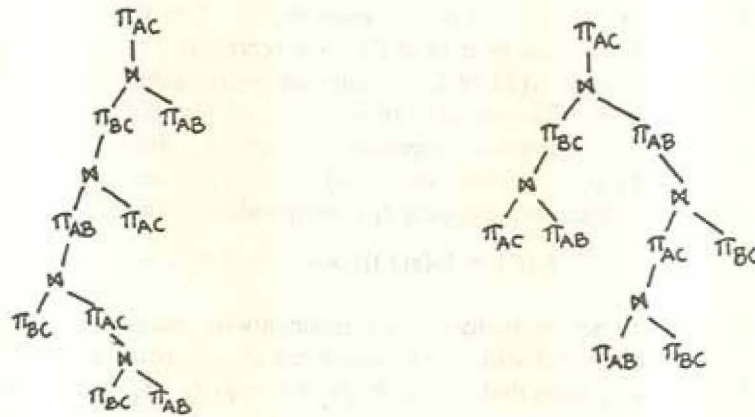


Figure 2

They both have the same tableau, namely the one shown below.



A		C
$a_1$	$b_1$	$c_1$
$a_2$	$b_2$	$c_1$
$a_3$	$b_2$	$c_2$
$a_3$	$b_3$	$C$
$A$	$b_1$	$c_3$

However, we show below (Corollary 2.1) that they cannot be shown equivalent by A1-A7 alone. Let  $\phi$  be an expression. Let  $Cl(\phi)$  be the equivalence class of expressions that can be shown equivalent to  $\phi$  via the axioms A1, A5, A6 and A7(b) alone. In other words,  $Cl(\phi)$  contains all "simple syntactic variants" of  $\phi$ . All expressions in  $Cl(\phi)$  have the same tableau,  $T_\phi$ . We construct a digraph  $D_\phi = (N, E_\phi)$  with node set the set  $N$  of nondistinguished symbols in  $T_\phi$ , and with  $(a_i, a_j) \in E_\phi$  iff the projection that created  $a_i$  is an ancestor of that which created  $a_j$  in all expressions in  $Cl(\phi)$ . For the two expressions shown in Figure 2, the corresponding digraphs are as shown in Figure 3.

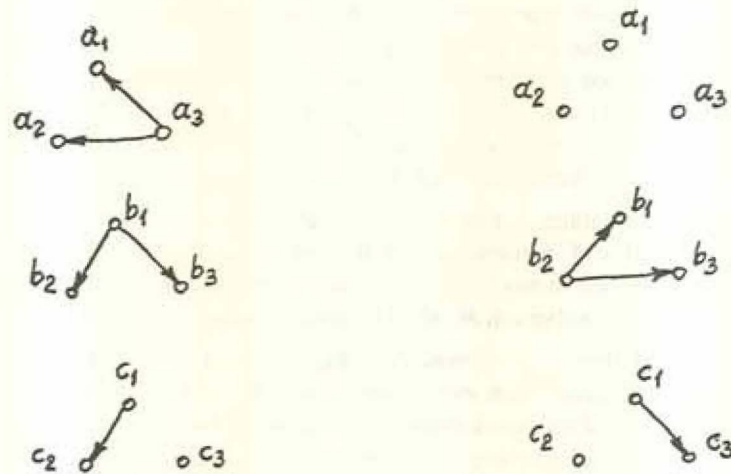


Figure 3

We saw in Lemma 2.1 that  $\phi_1(R) \subseteq \phi_2(R)$  iff there is a homomorphism  $h$  from  $T_{\phi_2}$  to  $T_{\phi_1}$ . We define now a restricted version of tautological inclusion. We write  $\phi_1(R) \stackrel{\subseteq}{=} \phi_2(R)$  iff there is a homomorphism  $h$  from  $T_{\phi_2}$  to  $T_{\phi_1}$  such that for all nondistinguished symbols  $a_i, a_j$  of  $T_{\phi_2}$ , we have  $(a_i, a_j) \in E_{\phi_2}$  iff  $a_i = a_j$  or  $(h(a_i), h(a_j)) \in E_{\phi_1}$  or  $h(a_i)$  is distinguished.  $\phi_1(R) \stackrel{=}{=} \phi_2(R)$  stands for  $\phi_1 \stackrel{\subseteq}{=} \phi_2$  and  $\phi_2 \stackrel{\subseteq}{=} \phi_1$ .

**Lemma 2.2** Axioms A1-A7 are sound even if  $\subseteq$  is replaced by  $\stackrel{\subseteq}{=}$  and  $=$  by  $\stackrel{=}{=}$ .

*Proof* By inspection. For example to show that  $\pi_{XY}(R_1 \bowtie R_2) \stackrel{\subseteq}{=} \pi_{XY}(R_1 \bowtie \pi_Y(R_2))$ , map the part

of the right-hand tableau that corresponds to  $R_1$  to itself via the identity homomorphism, and likewise for  $\pi_Y(R_2)$ ; only map the nondistinguished symbols introduced by  $\pi_Y$  to those introduced by  $\pi_{XY}$  in the left-hand side. This homomorphism obviously preserves edges of  $D_\phi$ .  $\square$

*Theorem 2.1* Suppose that  $\phi_1(R) \subseteq \phi_2(R)$ , but  $\phi_1(R) \not\subseteq \phi_2(R)$ . Then  $\phi_1(R) \subseteq \phi_2(R)$  cannot be proved from A1-A7.

*Proof* Since all axioms hold for  $\overset{\sim}{\subseteq}$  as well, and since  $\overset{\sim}{\subseteq}$  is transitive just as  $\subseteq$ , no proof can distinguish between  $\subseteq$  and  $\overset{\sim}{\subseteq}$ .  $\square$

*Corollary 2.1* The two expressions  $\phi_1$  and  $\phi_2$  shown in Figure 2 cannot be shown equivalent by A1-A7.

*Proof* Obviously  $\phi_1(R) = \phi_2(R)$ ; however,  $\phi_1(R) \not\subseteq \phi_2(R)$ , since there are only identity homomorphisms from  $T_{\phi_1}$  to  $T_{\phi_2}$  and back, and still  $D_{\phi_1} \neq D_{\phi_2}$ .  $\square$

This inability of the axioms to capture all aspects of expression equivalence has its roots at the inability of project-join expressions to represent arbitrary tableaux. The intricate combinatorics of this problem are dramatized by the following result.

*Theorem 2.2* Given a tableau, it is NP-complete to decide whether it corresponds to a project-join expression.

To prove the theorem we shall make use of *simple* tableaux. A *repeated symbol* of a tableau  $T$  is a symbol that appears in more than one rows. A tableau is called *simple* if it has at most one repeated nondistinguished symbol per column.\* For example, the tableau of Fig. 1a is not simple because it contains two repeated symbols in the first and third columns ( $A, a_2, c_1, c_3$ ). An expression is called simple if its tableau is simple. [ASSU] gives an algorithm that determines whether a simple tableau comes from an expression and constructs such an expression if it does. The basic ideas behind their algorithm are summarized in the following lemma.

*Lemma 2.3* Let  $T$  be a simple tableau. Let  $G(T)$  be a labeled graph with one node for every row of  $T$  and an edge  $(u, v)$  labeled  $w$  if in some column the rows  $u$  and  $v$  have the same nondistinguished symbol and  $w$  has a distinguished symbol. The tableau  $T$  corresponds to an expression if and only if there is no connected nontrivial subgraph  $H$  of  $G(T)$  with all edges of  $H$  labeled with nodes in  $H$ .

*Proof (only if)* Suppose that there is a connected subgraph  $H$  of  $G(T)$  with all edges labeled with nodes in  $H$ , and suppose that there is an expression  $\phi$  with  $T_\phi = T$ . Each row of  $T$  corresponds to a leaf of  $\phi$ . Let  $x$  be the lowest common ancestor of all nodes in  $H$ . Let  $y_1, y_2, \dots$  be the sons of  $x$ , and  $F_1, F_2, \dots$  the subtrees of  $\phi$  rooted at them. From our choice of  $x$ , at least two of the  $F_i$ 's contain nodes from  $H$ . Since  $H$  is connected, there is an edge  $(u_1, u_2)$  of  $H$  with  $u_1$  and  $u_2$  belonging to different subtrees, say  $u_1 \in F_1, u_2 \in F_2$ . Since  $u_1$  and  $u_2$  have the same nondistinguished symbol in some column  $A$ , the projection which created this symbol must take place above  $x$ ; i.e., some projection in the path from  $x$  to the root does not contain  $A$ . Let  $w$  be the label of  $(u_1, u_2)$ . Since  $w \in H$ ,  $w$  is a descendant of  $x$  and therefore will not have in  $T_\phi$  a distinguished symbol in column  $A$ .

*(if)* If the condition of the lemma is satisfied, then the algorithm of [ASSU] succeeds in finding an expression  $\phi$  with  $T_\phi = T$ . We will describe here briefly, for later use, how such an expression is constructed. Let  $G_1, G_2, \dots, G_k$  be the connected components of  $G(T)$ ;  $k \geq 2$ . Let  $T_1, T_2, \dots, T_k$  be the subtableaux of  $T$  corresponding to the sets of nodes of the various components. From  $T_i$  we construct  $T_i'$  by changing a nondistinguished symbol into a distinguished if it appears also in some other  $T_j$ . Since  $T$  is simple this can happen with at most one symbol for each column and  $T$  cannot have a distinguished symbol in such a column (from the construction of  $G(T)$ ). Now,  $T_1', \dots, T_k'$  are simple tableaux, and  $G(T_i')$  is a subgraph of  $G(T)$  for each  $i$ . Therefore, the  $T_i'$ 's

\* This definition is slightly more general than the one in [Aho et al. 1979].



satisfy the condition of the lemma and we can find expressions  $\phi_1, \dots, \phi_k$  such that  $T_{\phi_i} = T_i'$ . The expression  $\phi$  for  $T$  is  $\pi_X(\bigvee_i \phi_i)$ , where  $X$  is the set of columns in which  $T$  has a distinguished symbol. We call the expression that is constructed in this way, the *canonical expression* for  $T$ .  $\square$

*Lemma 2.4.* Let  $T$  be a tableau and suppose that  $T$  has at most one repeated nondistinguished symbol in each column of  $a(T)$  and at most two repeated nondistinguished symbols in each column of  $U - a(T)$ . Then if  $T$  corresponds to an expression, there is an expression  $\phi = \pi_{a(T)}\sigma$  where  $T_\phi = T$  and  $\sigma$  is a simple expression with  $a(\sigma) = U$ .

*Proof.* Let  $\psi$  be an expression with  $T_\psi = T$ . We can assume without loss of generality that all projections that create distinct (i.e. non repeated) nondistinguished symbols take place at the leaves. Let  $A$  be a column in which  $T$  has two repeated nondistinguished symbols  $a_1, a_2$  and let  $v_1, v_2$  be the two nodes in which the projections that create these symbols take place. At least one of the two nodes, say  $v_1$ , is not a descendant of the other. If we postpone the projection that creates  $a_1$  until the root, then  $T$  does not change. Doing the same with all columns of  $U - a(T)$  we get an expression  $\phi = \pi_{a(T)}\sigma$ , where  $\sigma$  is simple with  $a(\sigma) = U$  and  $T_\phi = T_\psi = T$ .  $\square$

Let  $T$  be a tableau as in Lemma 2.4. We construct a graph  $G(T)$  as follows. The nodes of  $G(T)$  are the rows of  $T$ .  $G(T)$  has an edge  $(u, v)$  labeled  $w$  if in some column  $u$  and  $v$  have the same nondistinguished symbol and  $w$  has a distinguished symbol. In addition,  $G(T)$  has two sets of edges  $S_1(A), S_2(A)$  for each column  $A$  in which  $T$  has two repeated nondistinguished symbols  $a_1, a_2$  ( $A \in U - a(T)$ ).  $S_1(A)$  contains an edge  $(u, v)$  labeled  $w$  for each pair of rows  $u, v$  that have symbol  $a_1$  in column  $A$  and each row  $w$  that has  $a_2$  in column  $A$ , and similarly with  $S_2(A)$ . Lemma 2.4 then implies that  $T$  comes from an expression if and only if we can delete either  $S_1(A)$  or  $S_2(A)$  for each column  $A$  in which  $T$  has two repeated nondistinguished symbols so that the remaining graph satisfies the condition of Lemma 2.3. The proof of Theorem 2.2 is based on this combinatorial property.

*Proof of Theorem 2.2.*

It is obvious that the problem is in NP: Guess an expression  $\phi$ , construct  $T_\phi$ , and verify that  $T_\phi = T$ . For the NP-hardness part we shall describe a reduction from the 3-SAT problem (satisfiability of a Boolean formula in conjunctive normal form with 3 literals per clause). Let  $C_1, C_2, \dots, C_p$  be the clauses of a Boolean formula  $F$  over the variables  $x_1, x_2, \dots, x_n$ . The universe  $U$  has  $12p + n$  attributes; the first  $n$ ,  $X_1, X_2, \dots, X_n$  correspond to the  $n$  variables, and the rest are divided into  $p$  groups of 12 attributes each - one group for each clause. We will construct a tableau  $T$  over  $U$  such that  $T$  corresponds to an expression iff  $F$  is satisfiable. The attributes  $a(T)$  of  $T$  are  $U - \{X_1, \dots, X_n\}$ . The tableau  $T$  has the form of Lemma 2.4 with two repeated nondistinguished symbols  $x_i$  and  $\bar{x}_i$  in each column  $X_i$ ,  $i = 1, \dots, n$ . For each clause,  $T$  has 16 rows. In Figure 4 we show the symbols of these rows for a clause  $C = y_1 \vee y_2 \vee y_3$  (where  $y_i = x_i$  or  $\bar{x}_i$ ) in the columns that correspond to this clause and  $X_1, X_2, X_3$ ; the entries in the rest of the columns are distinct nondistinguished symbols.

The portion of the graph  $G(T)$  corresponding to the rows for the clause  $C$  is shown in Figure 5.

In the figure we have labeled edges due to columns  $X_i$ , by the nondistinguished symbols rather than the rows. From our previous discussion,  $T$  corresponds to an expression iff deletion of all edges due to  $x_i$  or  $\bar{x}_i$  for each  $i = 1, \dots, n$ , results in a graph satisfying the condition of Lemma 2.3. Let us associate the deletion of all edges due to  $y_i$  ( $y_i = x_i$  or  $\bar{x}_i$ ) with the assignment of truth value 1 to the literal  $y_i$ . We claim that a truth assignment  $\tau$  satisfies  $F$  if and only if deletion of the set  $S(\tau)$  of corresponding edges results in a graph satisfying Lemma 2.3.

(only if). Let  $\tau$  be a satisfying truth assignment for  $F$  and suppose that  $G' = G(T) - S(\tau)$  contains a nontrivial connected subgraph  $H$  all of whose edges are labeled with nodes in  $H$ .  $G'$  contains a clique for each false literal and all these cliques are node-disjoint and disconnected from each other. Therefore, in order that  $H$  satisfies the previous condition, it must contain at least one edge from a clause construction that is not labeled by a literal. Let  $C = y_1 \vee y_2 \vee y_3$  be such a clause.

	$X_1$	$X_2$	$X_3$											
$C$				□	□	□	□	□	□	●		●		●
$A_1$													●	●
$A_2$										●				●
$A_3$											●	●		
$B_1$	$y_1$												●	
$B_2$		$y_2$												●
$B_3$			$y_3$								●			
$D_1$	$y_1$			●										
$D_2$		$y_2$			●									
$D_3$			$y_3$			●								
$E_1$				●	●					□	□			
$E_2$						●	●					□	□	
$E_3$								●	●					□
$F_1$		$\bar{y}_2$			●									
$F_2$			$\bar{y}_3$				●							
$F_3$	$\bar{y}_1$							●						

Figure 4

A □ denotes a distinguished symbol; a ● denotes a repeated nondistinguished symbol; blank denotes a distinct nondistinguished symbol.

*Case 1*  $y_1 = y_2 = y_3 = 1$  in  $\tau$ .

Then,  $C$  and the  $A$  and  $B$  nodes are isolated from the rest of  $G'$ . Since the edges that connect them are labeled with  $E$ -nodes they cannot be in  $H$ . But then none of the other edges (all of them labeled  $C$ ) of the construction for  $C$  can be either in  $H$ .

*Case 2*  $C$  has a false literal.

Since  $\tau$  is a satisfying truth assignment  $C$  has also a true literal. Then for some  $i = 1, 2, 3$  we have  $y_i = 1, y_{i+1} = 0$  (addition is mod 3). From the symmetry of the construction we can assume without loss of generality that  $y_1 = 1, y_2 = 0$ . Then the nodes  $D_1, E_1, F_1$  are isolated from the rest of  $G'$ . Therefore, the edges  $(C, A_2), (A_3, B_3)$  are not in  $H$ . Deletion of these edges isolates  $C, A_1, B_1, A_3$  from the rest of the graph. Therefore, no edge labeled  $C$  is in  $H$ , and  $H$  cannot contain again any edge from the construction for  $C$  that is not labeled by a literal.

(if). Let  $\tau$  be a truth assignment and suppose that deletion of  $S(\tau)$  leaves a graph satisfying the



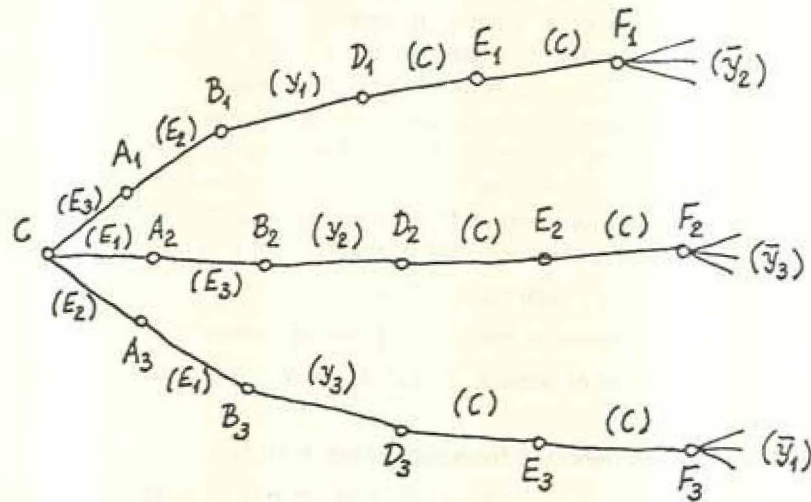


Figure 5

condition of Lemma 2.3. Let  $C$  be a clause of  $F$ . If  $\tau$  does not satisfy any literal of  $C$ , then the construction for  $C$  is a connected graph  $H$  all of whose edges are labeled with nodes in  $H$ . (Note that an edge  $(B_i, D_i)$  has label  $F_{i-1}$ ).  $\square$

It turns out that Theorem 2.2, besides characterizing the complexity of compiling expressions from their tableau also reveals that most probably there can be no usable axiomatization of expression equivalence.

*Corollary 2.2* Given an expression  $\phi$  it is NP-complete to test whether there exists a  $\phi' \in Cl(\phi)$  such that  $T_\phi = T_{\phi'}$ .

*Proof* In the construction in the proof of Theorem 2.2, two expressions  $\phi$  and  $\phi'$ , both having the same tableau  $T$ , satisfy  $\phi' \in Cl(\phi)$  iff  $\phi$  and  $\phi'$  come from different truth assignments for  $F$ . The Corollary now follows from the observation that it is NP-complete to decide, given a Boolean formula  $F$  and a truth assignment  $\tau$  satisfying  $F$ , whether there is *another* truth assignment that satisfies  $F$ .  $\square$

Therefore the apparent difficulty in axiomatizing expression equivalence can be viewed as a consequence of the difficulty involved in transforming an accepting non-deterministic computation to another by formal manipulations. In the next two Sections we show how to overcome this difficulty by replacing  $T$  by another, purely algebraic, axiom. This difficulty does not arise in the case of *simple* expressions. This is reflected in the following result:

*Theorem 2.3* Any identity of the form  $\phi_1(R) \subseteq \phi_2(R)$  for simple expressions  $\phi_1, \phi_2$  can be proved by A1-A7.

The proof of the theorem proceeds in two steps. At first we show that a simple expression  $\phi$  can be shown, using the axioms, to be equivalent to the canonical expression for its tableau  $T_\phi$ , and then we prove the theorem for  $\phi_1, \phi_2$  canonical simple expressions.

*Lemma 2.5* Let  $\phi$  be a simple expression and  $\phi^*$  the canonical expression for  $T_\phi$ . Then  $\phi = \phi^*$  can be shown using A1-A7.



*Proof* We prove the lemma by induction on the depth of  $\phi$ . The basis is trivial. For the inductive step, let  $\phi = \pi_a(\phi_1 \bowtie \phi_2 \bowtie \dots \bowtie \phi_t)$  where  $a = a(\phi) = a(T_\phi)$ . Let  $F_1, F_2, \dots, F_t$  be the trees for  $\phi_1, \dots, \phi_t$ . If the last projection of  $\phi$  creates in some column  $A$  a nondistinguished symbol that appears in leaves of only one of the  $F_i$ 's then we move this projection below the join and incorporate it into the corresponding  $\phi_i$  using A7b. Thus, let us assume that each nondistinguished symbol created in the last projection appears in leaves of at least two of the  $F_i$ 's. Let  $M_i$  be the rows of  $T_\phi$  that correspond to the leaves of  $F_i$ , for  $i = 1, \dots, t$ . Let  $T_i$  be the subtableau of  $T_\phi$  defined by the rows of  $M_i$ , and let  $T_i'$  be obtained from  $T_i$  by changing a nondistinguished symbol into a distinguished if it appears in a row of another  $M_j$ . From our assumption above,  $T_i'$  is the tableau of  $\phi_i$ . Let  $N_1, \dots, N_k$  be the nodes in the connected components of  $G(T_\phi)$ .

*Claim 1* For all  $i$ , there is a  $j$  such that  $N_i \subseteq M_j$ .

*Proof of Claim 1* Similar to the proof of the (only if) part of Lemma 2.3.  $\square$

Thus, each  $M_j$  is the union of some  $N_i$ 's. Let  $M_1 = N_1 \cup \dots \cup N_t$ . (Similar arguments hold for the rest of the  $M_j$ 's.)

*Claim 2* The  $N_i$ 's are disconnected from each other in  $G(T_1')$ .

*Proof of Claim 2* Let  $v_1 \in N_1, v_2 \in N_2$  and suppose that there is an edge  $(v_1, v_2)$  labeled  $u$  in  $G(T_1')$ . Then  $v_1$  and  $v_2$  have the same nondistinguished symbol  $b$  in a column  $B$  in which  $u$  has a distinguished symbol. Since  $(v_1, v_2) \notin G(T_\phi)$ , the row  $u$  has in  $T_\phi$  a nondistinguished symbol  $b'$  which appears also in a row of another  $M_j$ . Thus, there are at least two repeated nondistinguished symbols in column  $B$  contradicting the simplicity of  $T_\phi$ .  $\square$

Thus, in  $G(T_1')$  each  $N_i$  is a union of connected components. Let  $\phi_1^*$  be the canonical expression for  $T_1'$ . By inductive hypothesis,  $\phi_1 = \phi_1^*$  can be derived from the axioms. From the construction in the proof of Lemma 2.3,  $\phi_1^* = \pi_{a(T_1')}(\bowtie \psi_j)$  where each  $\psi_j$  corresponds to a component of  $G(T_1')$ . Using associativity of join and Claim 2,  $\phi_1^*$  can be transformed to  $\sigma_1 = \pi_{a(T_1')}(\psi_1' \bowtie \psi_2' \bowtie \dots \bowtie \psi_i')$  where  $\psi_i'$  is the join of the  $\psi_j$ 's that correspond to the components whose union is  $N_i$ . Let  $X_i$  be the set of attributes in which some row in  $N_i$  has a distinguished symbol or a common nondistinguished symbol with another  $N_j$ .  $a(\psi_i') - X_i$  is the set of attributes in which two rows in different components of  $G(T_1')$  that are contained in  $N_i$  have a common nondistinguished symbol that does not appear in any other  $N_j$ . Thus,  $[a(\psi_i') - X_i] \cap a(\psi_j') = \emptyset$  for all  $j \neq i$ , and we can replace  $\psi_i'$  in  $\sigma_1$  by  $\pi_{X_i} \psi_i'$  using A7b. The tableau of  $\pi_{X_i} \psi_i'$  is obtained from the rows of  $N_i$  by changing nondistinguished symbols that appear in other  $N_j$ 's into distinguished. Let  $\tau_i$  be the canonical form for this expression. The canonical expression for  $T_\phi$  is  $\phi^* = \pi_a(\tau_1 \bowtie \tau_2 \bowtie \dots \bowtie \tau_k)$ . By induction hypothesis, we can derive  $\tau_i = \pi_{X_i} \psi_i'$  and consequently,  $\sigma_1 = \pi_{a(T_1')}(\tau_1 \bowtie \dots \bowtie \tau_i)$ . Proceeding similarly with the rest of the  $M_j$ 's we can transform  $\phi$  into  $\pi_a[\pi_{a(T_1')}(\tau_1 \bowtie \dots \bowtie \tau_i) \bowtie \dots \bowtie \pi_{a(T_m')}(\tau_m \bowtie \dots \bowtie \tau_k)]$ .

Let  $Y_1 = \bigcup_{i=1}^t a(\tau_i) - a(T_1')$ ;  $Y_1$  is the set of attributes in which two  $N_i$ 's in  $M_1$  have a common nondistinguished symbol that does not appear in another  $M_j$ . Using A7b we can move this projection to the root. Doing the same with the rest of the  $M_j$ 's transforms  $\phi$  into  $\pi_a[(\tau_1 \bowtie \dots \bowtie \tau_i) \bowtie \dots \bowtie (\tau_m \bowtie \dots \bowtie \tau_k)] = \phi^*$  (by A6).  $\square$

### *Proof of Theorem 2.3*

Let  $\phi_1, \phi_2$  be two simple expressions with  $\phi_1 \subseteq \phi_2$ . Let  $\phi_1^*, \phi_2^*$  be the canonical expressions for the tableaux  $T_{\phi_1} = T_1, T_{\phi_2} = T_2$ . From Lemma 2.5 we can prove  $\phi_1 = \phi_1^*$  and  $\phi_2 = \phi_2^*$  using A1-A7. Thus, it suffices to prove  $\phi_1^* \subseteq \phi_2^*$ . We will use induction on the structure of  $\phi_1^*$ .

From Lemma 2.1 there is a homomorphism  $h$  from  $T_2$  to  $T_1$ . Suppose  $h(b) = B$  for a repeated nondistinguished symbol  $b$  of  $T_2$  in column  $B$ . Let  $T_2'$  be obtained from  $T_2$  by changing  $b$  into  $B$ .  $T_2'$  is simple and comes also from an expression since  $G(T_2')$  is a subgraph of  $G(T_2)$ . An expression  $\psi_2'$  for  $T_2'$  can be obtained from  $\phi_2^*$  as follows. Let  $v$  be the node of  $\phi_2$  in which the projection that creates  $b$  takes place. From the construction of a canonical expression, all



projections that create a nonrepeated nondistinguished symbol take place at the leaves of  $\phi_2^*$ . Therefore, no projection above  $v$  creates a nondistinguished symbol in column  $B$ . The expression  $\psi_2'$  is obtained from  $\phi_2^*$  by including  $B$  in all projections at node  $v$  and its ancestors. Since  $h(b) = B, B \in a(\phi_1) = a(\phi_2)$  and  $a(\psi_2') = a(\phi_2^*)$ .

We can show  $\psi_2' \subseteq \phi_2^*$  using A7 (and A3, A4): Let  $u$  be the lowest ancestor of  $v$  in  $\phi_2^*$  such that the subexpression corresponding to the tree rooted at  $u$  has  $B$  in its attributes. The expressions  $\psi_2'$  and  $\phi_2^*$  differ at the projections along the path from  $u$  to  $v$ . Let  $u_1$  be the son of  $u$  in this path (see Figure 6), and  $X_1$  the set of attributes of the subexpression of  $\phi_2^*$  rooted at  $u_1$ . From the choice of  $u$ , no subtree joined at a node in this path has  $B$  in its attributes. Therefore, using A7b, we can postpone in  $\phi_2^*$  the projection that creates  $b$  until  $u_1$  while preserving equality.

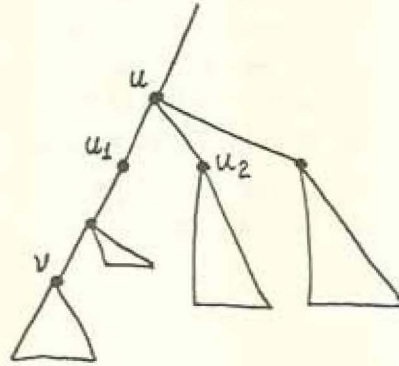


Figure 6.

Changing then the projection of  $u_1$  from  $\pi_{X_1}$  to  $\pi_{X,B}$  (to get  $\psi_2'$ ) will shrink the expression by A7a; thus,  $\psi_2' \subseteq \phi_2^*$ .

Let now  $\bar{T}_2$  be obtained from  $T_2$  by changing all repeated nondistinguished symbols  $b$  of  $T_2$  such that  $h(b) = B$  into the corresponding distinguished symbols. As above, we can find an expression  $\bar{\psi}_2$  with  $T_{\bar{\psi}_2} = \bar{T}_2$  and prove that  $\bar{\psi}_2 \subseteq \phi_2^*$ . Let  $\bar{\phi}_2^*$  be the canonical expression for  $\bar{T}_2$ . From Lemma 2.5 we can show  $\bar{\psi}_2 = \bar{\phi}_2^*$ , and therefore also  $\bar{\phi}_2^* \subseteq \phi_2^*$ , using the axioms.

The restriction of  $h$  to the symbols of  $\bar{T}_2$  is a homomorphism from  $\bar{T}_2$  to  $T_1$  that maps all repeated nondistinguished symbols into nondistinguished symbols. Let  $G_1 = G(T_1), G_2 = G(\bar{T}_2)$ . Let  $(u,v)$  be an edge in  $G_2$ ; then  $u$  and  $v$  have a common nondistinguished symbol  $b$  in some column  $B$  in  $a(\bar{T}_2)$ . Since  $h(b)$  is a nondistinguished symbol, we have that either  $h(u) = h(v)$  or  $(h(u), h(v))$  is an edge of  $G_1$ . Therefore, the image (under  $h$ ) of a connected subgraph of  $G_2$  is connected.

Let  $N_1, \dots, N_k$  be the nodes of the connected components of  $G_2$ , and  $M_1, \dots, M_l$  those of  $G_1$ . We have,  $h(N_i) \subseteq M_j$ , for each  $i$ , some  $j$ . The canonical expressions for  $\bar{T}_2, T_1$  are  $\bar{\phi}_2^* = \pi_a(\sigma_1 \bowtie \dots \bowtie \sigma_k)$ , and  $\phi_1^* = \pi_a(\tau_1 \bowtie \dots \bowtie \tau_l)$  where  $a = a(\bar{\phi}_2^*) = a(\phi_1^*)$ . Using associativity and commutativity of join we can write  $\bar{\phi}_2^*$  as  $\pi_a[(\sigma_1 \bowtie \dots \sigma_i) \bowtie \dots (\dots \bowtie \sigma_k)]$ , where the  $\sigma_i$ 's of those  $N_i$ 's that are mapped into the same  $M_j$  are grouped together. Let  $W_j$  be the set of attributes in which two  $N_j$ 's mapped into  $M_i$  have a common repeated nondistinguished symbol that does not appear in a row mapped to a different  $M_j$ . Using A7b we can move the projection that creates these nondistinguished symbols below the first join. That is,  $\bar{\phi}_2^*$  is transformed into  $\pi_a(\bar{\tau}_1 \bowtie \dots \bar{\tau}_l)$  ( $l \leq l$ ) where  $a(\bar{\tau}_i)$  is the set of columns in which some  $N_j$  mapped into  $M_i$  has (1) a distinguished symbol or (2) a common repeated nondistinguished symbol with a row mapped in another  $M_j$ .

Let  $\bar{X}_i$  be the first set of columns and  $\bar{Y}_i$  the second set of columns. Note that from the construction of  $G(\bar{T}_2)$ ,  $\bar{Y}_j \cap a(\bar{T}_2) = \emptyset$  and therefore  $\bar{X}_i \cap \bar{Y}_i = \emptyset$ . The tableau of  $\bar{\tau}_i$  is formed by taking the rows of  $\bar{T}_2$  in the  $N_j$ 's that are mapped to  $M_i$  and changing repeated nondistinguished symbols in  $\bar{Y}_i$  into distinguished. The attributes of  $\tau_i$  ( $1 \leq i \leq t$ ) are (1) those columns in which a row of  $M_i$  has a distinguished symbol, and (2) the columns in which a row of  $M_i$  has a common nondistinguished symbol with a row of another  $M_j$ . Let  $X_i$  be the first set of columns and  $Y_i$  the second set. Clearly, we have  $\bar{X}_i \subseteq X_i$  and  $\bar{Y}_i \subseteq Y_i$ . Let  $Z_i = X_i \cup \bar{Y}_i$ . Let  $S(r)$  be the set of columns in which a row  $r$  of  $T_1$  has distinguished symbols, and  $F_i$  the family of such sets  $S(r)$  for all rows  $r$  in  $M_i$ . From A2 and Example 2.1 we have  $\bar{\phi}_2^* = \pi_a(\boxtimes_i \bar{\tau}_i) = \pi_a\left(\left[\boxtimes_i \bar{\tau}_i\right] \boxtimes \left[\boxtimes_{S \in F_i} \pi_S(R)\right]\right) =$  (by A5, A6)  $\pi_a(\boxtimes_i \tau_i')$ , where  $\tau_i' = \bar{\tau}_i \boxtimes \left(\boxtimes_{S \in F_i} \pi_S(R)\right)$ . Now,  $a(\tau_i') = Z_i$ ; the tableau of  $\tau_i'$  is that of  $\bar{\tau}_i$  with some additional rows, each of which has only distinguished and nonrepeated nondistinguished symbols, and therefore is simple. The tableau of  $\pi_Z \tau_i$  is obtained from the rows of  $M_i$  by changing repeated nondistinguished symbols in  $\bar{Y}_i$  into distinguished and therefore is also simple. Now,  $h$  gives a homomorphism from the tableau of  $\tau_i'$  to that of  $\pi_Z \tau_i$  (with the new rows in  $T_{\tau_i'}$  mapped to the corresponding rows in the tableau of  $\pi_Z \tau_i$ ). Thus,  $\pi_Z \tau_i \subseteq \tau_i'$  can be proved from the axioms.

We have  $\phi_1^* = \pi_a(\tau_1 \boxtimes \dots \boxtimes \tau_t) \subseteq$  (by A2)  $\pi_a(\tau_1 \boxtimes \dots \boxtimes \tau_t) \subseteq$  (by A7a)  $\pi_a(\pi_Z \tau_1 \boxtimes \dots \boxtimes \pi_Z \tau_t) \subseteq$  (by A3, A4)  $\pi_a(\tau_1' \boxtimes \dots \boxtimes \tau_t') = \bar{\phi}_2^* \subseteq \phi_2^*$ .  $\square$



### 3. EXTENDED RELATIONS

Let  $R$  be a relation with attributes  $a(R) = U = \{A, B, \dots, Z\}$ . The *extension*  $\bar{R}$  of  $R$  is a relation with  $a(\bar{R}) = \bar{U} = \{A_1, B_1, \dots, Z_1, A_2, B_2, \dots, Z_2, A_3, \dots\}$ , and such that  $\bar{R} = \{(t; t; \dots) : t \in R\}$ .  $\bar{R}$  is therefore an *infinite collection of copies* of  $R$ . The attributes  $A_1, A_2, \dots$  of  $\bar{R}$  are said to be *associated with* (or *copies of*) the attribute  $A$ . We can have projection and join applied to extended relations. We adopt the convention that projection to a finite subset of  $\bar{U}$  is applied first to  $\bar{R}$ . If  $\phi_1$  and  $\phi_2$  are expressions over  $\bar{U}$ , we write  $\phi_1(\bar{R}) \subseteq \phi_2(\bar{R})$  to denote the identify inclusion *under the assumption that  $\bar{R}$  is an extended relation*, that is, the elements of each tuple of  $\bar{R}$  corresponding to  $A_i, A_j$  are restricted to be the same.

If it appears that by the above definitions we are introducing infinitary operations in our algebraic language, we really are not. We could achieve the same effect by considering expressions over project, join and a new operation called say, *duplicate*, where  $duplicate(R) = \{(t; t) : t \in R\}$ . A formalism similar to ours, only with a limited number of copies (namely, 2) of each attribute allowed, was used in [Sciore 1979].

Extended relations can be used to express dependencies that were previously thought of as non-algebraic. For example, the functional dependency  $A \rightarrow B$  can be written as

$$\pi_{AB_1}(\bar{R}) \bowtie \pi_{AB_2}(\bar{R}) \subseteq \pi_{AB_1, B_2}(\bar{R}).$$

Expressions on extended relations play a very important role in our development. To show their inherent stability as a concept, we prove that they are equivalent in expressive power to two important existing systems: project-equi-join expressions, and tableaux.

So far, a relation for us was a set of sets of mappings, one for each attribute of  $U$ . In the customary mathematical sense, however, a relation is a subset of a Cartesian product; that is, the columns are ordered. We shall use the term *ordered relation* for these. Our relations will be sometimes called *attributed* for distinction.

The equi-join operator was defined by Codd on ordered relations. To compare the power of equi-join to expressions over extended relations we will associate each column of an ordered relation with an attribute in  $U$ . If  $R_1$  is an ordered relation with each column associated with an attribute, and  $R_2$  an attributed relation with the number of columns of  $R_1$  equal to  $|a(R_2)|$ , we say that  $R_1 = R_2$  if we can order the attributes of  $R_2$  so that the resulting ordered relation is equal to  $R_1$ . (Note that this implies that corresponding columns in  $R_1$  and  $R_2$  are associated with the same attribute, since the domains are disjoint.) If  $R_1, R_2$  are relations, and  $I_1 = \{i_1, \dots, i_n\}, I_2 = \{j_1, \dots, j_n\}$  are sets of columns of  $R_1, R_2$  respectively with columns  $i_k, j_k$  associated with the same attribute for  $k = 1, \dots, n$ , the *equi-join of  $R_1, R_2$  on  $I_1, I_2$*  is the relation  $R_1 \bowtie_{I_1, I_2} R_2 = \{(t_1, t_2) : t_1 \in R_1, t_2 \in R_2,$

and  $t_{U_1} = t_{U_2}\}$ . The columns of  $R_1 \bowtie_{I_1, I_2} R_2$  are associated with the same attributes as in  $R_1, R_2$ .

Thus, the definition of equi-join includes the notion of repetition of columns. A project-equi-join expression is an expression built using projection and equi-join over the single variable symbol  $R$  ranging over all relations with  $|U|$  columns, each of which is associated with an attribute of  $U$ .

**Theorem 3.1** (1) If  $\phi$  is a project-equi-join expression, then there is a project-join expression  $\phi'$  such that for all relations  $R, \phi'(R) = \phi(R)$ . (2) Conversely, for every project-join expression  $\phi'$  there is a project-equi-join expression  $\phi$  such that for all relations  $R, \phi'(R) = \phi(R)$ .

*Proof* (1) The proof is by induction on the structure of  $\phi$ . The basis ( $\phi(R) = R$ ) is trivial. For the induction step, suppose first that  $\phi = \pi_I \sigma$ , for some expression  $\sigma$  and set of columns  $I$ . By the induction hypothesis  $\sigma(R) = \sigma'(R)$  for some  $\sigma'$ . Let  $X$  be the set of attributes of  $a(\sigma')$  that correspond to the columns in  $I$ . We take  $\phi' = \pi_X \sigma'$ . If  $\phi(R) = \sigma_1(R) \bowtie_{I_1, I_2} \sigma_2(R)$ , let  $\sigma_1', \sigma_2'$  be such that  $\sigma_1(R) = \sigma_1'(\bar{R}), \sigma_2(R) = \sigma_2'(\bar{R})$ . By changing the names of some attributes we can choose  $\sigma_1', \sigma_2'$  so that  $a(\sigma_1') \cap a(\sigma_2') = X$  where each attribute in  $X$  corresponds to a column in  $I_1$  of  $\sigma_1(R)$  and the corresponding column in  $I_2$  of  $\sigma_2(R)$ . For each attribute  $A_i$  in  $X$  we introduce



an attribute  $A_j$  that does not appear in  $\sigma_1'$  or  $\sigma_2'$  and change every projection in  $\sigma_1'$  that includes  $A_i$  to include also  $A_j$ . Let  $\tau_1$  be the resulting expression. For every  $R$ ,  $\tau_1(\bar{R})$  is  $\sigma_1'(\bar{R})$  with some new columns which are copies of the columns in  $X$ . From the definition of the equijoin then,  $\phi(\bar{R}) = \tau_1(\bar{R}) \bowtie \sigma_2'(\bar{R})$ .

(2) Let  $\phi'$  be a project-join expression. Using equijoin of  $R$  with itself a sufficient number of times we can create a relation which contains one column for every attribute  $A_i$  that appears in  $\phi'$ .  $\square$

It will be useful for our further discussion to introduce tableaux also for project-join expressions over extended relations; we call them *extended tableaux*. An extended tableau  $T$  is a (usual) tableau over a finite subset  $X$  of  $\bar{U}$ , and defines a mapping  $\bar{f}_T$  from relations  $R$  over  $U$  to relations over the set  $a(T) \subseteq \bar{U}$  of attributes in which  $T$  has a distinguished symbol: this mapping is obtained by taking the projection of  $\bar{R}$  onto  $X$  and applying to it the mapping  $f_T$  defined by  $T$  in Section 2 as a usual tableau. From an expression  $\phi(\bar{R})$  we can construct an extended tableau  $T_\phi$  over the set  $X$  of attributes which appear in  $\phi$  — note that since projection of  $\bar{R}$  to a finite set of attributes is applied first, the set  $X$  is finite. The tableau  $T_\phi$  is constructed from  $\phi$  as in Section 2 by treating  $\phi$  as a usual expression on a relation symbol over  $X$ .

To understand how the semantics of the extension of a relation enter into the determination of  $\bar{f}_T$  we must introduce an operation on tableaux called *chase* [Maier et al. 1979]. If  $\Sigma$  is a set of functional dependencies and  $T$  a tableau, the chase of  $T$  under  $\Sigma$ ,  $chase_\Sigma(T)$  is the tableau obtained from  $T$  applying the following transformation repeatedly whenever possible: If  $X - Y$  is a functional dependency in  $\Sigma$  and two tuples  $u, v$  of  $T$  satisfy  $u_X = v_X$  then for every attribute  $A$  in  $Y$  we make  $u_A$  and  $v_A$  identical; if one of them is distinguished, so is the resulting symbol. The final tableau  $chase_\Sigma(T)$  is unique up to renaming of nondistinguished symbols. Now let  $F$  be the set of functional dependencies  $A_i - A_j$  for every two distinct copies  $A_i, A_j$  of the same attribute  $A$  of  $U$ . We will show that the semantics of extended relations are captured essentially by these functional dependencies. If  $T$  is an extended tableau, the chase of  $T$  under  $F$  can be constructed in a very simple way as follows. For every attribute  $A \in U$  form a graph  $G_A(T)$  with the tuples of  $T$  as nodes and an edge between two tuples that have the same symbol in some copy of  $A$ . For each connected component  $K$  of  $G_A(T)$  and each column  $A_i$  of  $T$  make the entries of the tuples in  $K$  identical; the common symbol is distinguished if some tuple of  $K$  has a distinguished symbol in column  $A_i$ . In the resulting tableau  $chase_F(T)$ , columns corresponding to copies of the same attribute are identical up to renaming of symbols.

*Lemma 3.1*  $\bar{f}_T(R) = \bar{f}_{chase_F(T)}(R)$ .

*Proof* [Aho et al. 1979] show that if a relation  $I$  satisfies a set  $\Sigma$  of functional dependencies then  $f_T(I) = f_{chase_\Sigma(T)}(I)$ . The lemma then follows by noting that  $\pi_X(\bar{R})$ , where  $X$  are the columns of  $T$ , satisfies the set  $F$  of functional dependencies.  $\square$

If  $T$  is an extended relation with set of columns  $X$ , and  $X'$  is a superset of  $X$ , we can form an (extended) tableau  $T'$  with set of columns  $X'$  and  $a(T) = a(T')$  by adding to  $T'$  new columns with distinct nondistinguished variables. Obviously,  $\bar{f}_T(R) = \bar{f}_{T'}(R)$  for every  $R$ . Thus, if  $T_1$  and  $T_2$  are two extended tableaux with sets of columns  $X_1, X_2$  respectively, we can consider them as tableaux over the same set of columns  $X_1 \cup X_2$  by adding new columns. The following lemma says essentially that the set  $F$  of functional dependencies captures the semantics of extended relations, at least as far as comparison of tableaux (and therefore also expressions) is concerned.

*Lemma 3.2* Let  $T_1, T_2$  be two tableaux with the same set  $X$  of columns and  $a(T_1) = a(T_2)$ .  $\bar{f}_{T_1}(R) \subseteq \bar{f}_{T_2}(R)$  for every relation  $R$  over  $U$  if and only if  $f_{chase_F(T_1)}(I) \subseteq f_{chase_F(T_2)}(I)$  for every relation  $I$  over  $X$ .

*Proof*

(if)  $\bar{f}_{T_i}(R) = \bar{f}_{chase_F(T_i)}(R) = f_{chase_F(T_i)}(\pi_X \bar{R})$ , for  $i = 1, 2$ . Let  $I = \pi_X \bar{R}$ ; then  $\bar{f}_{T_1}(R) = f_{chase_F(T_1)}(I) \subseteq f_{chase_F(T_2)}(I) = \bar{f}_{T_2}(R)$ .

(only if) [Aho et al. 1979] show that if  $\Sigma$  is a set of functional dependencies then  $f_{chase_\Sigma(T_1)}(I) \subseteq f_{chase_\Sigma(T_2)}(I)$  for every relation  $I$  iff  $f_{T_1}(I) \subseteq f_{T_2}(I)$  for every relation  $I$  satisfying  $\Sigma$ .



Suppose now that  $f_{chase_F(T_1)}(I) \not\subseteq f_{chase_F(T_2)}(I)$  for some relation  $I$  over  $X$ . Then there is a relation  $I$  over  $X$  satisfying  $F$  such that  $f_{T_1}(I) \not\subseteq f_{T_2}(I)$ . In  $I$ , columns corresponding to renamings of the same attribute of  $U$  are copies of each other. Let  $R$  be a relation over  $U$  obtained from  $I$  by keeping one column for each attribute of  $U$  and adding columns with distinct new symbols for attributes of  $U$  that don't have copies in  $X$ . It is easy to see then that  $\bar{f}_{T_1}(R) = f_{T_1}(\pi_X \bar{R}) \not\subseteq f_{T_2}(\pi_X \bar{R}) = \bar{f}_{T_2}(R)$ .  $\square$

Our proof of the equivalence between tableaux and extended expressions is based on a useful lemma. Call an expression *shallow* if it has the form  $\pi_X(\bowtie_i \pi_{X_i} \bar{R})$  where the  $X_i$ 's are finite subsets of  $\bar{U}$ . That is, a shallow expression is one whose tree has (at most) one node with outdegree greater than one. A tableau that corresponds to a shallow expression is called also *shallow*. Each column of a shallow tableau has either (i) a distinguished symbol and no repeated nondistinguished symbol, or (ii) one repeated nondistinguished symbol and no distinguished symbol. And conversely, a tableau  $T$  that satisfies these conditions is shallow: Let  $X_i$  be the set of attributes in which the  $i$ -th row has either a distinguished or a repeated nondistinguished symbol. Then  $T = T_\phi$ , where  $\phi = \pi_{a(T)}(\bowtie_i \pi_{X_i})$ . Thus, shallow tableaux are a very special case of simple tableaux.

**Lemma 3.3** Let  $T$  be an extended tableau. Then there exists a shallow extended tableau  $T'$  such that  $\bar{f}_T(R) = \bar{f}_{T'}(R)$  for all  $R$ .

*Proof* Let  $T_1$  be the chase of  $T$  under  $F$ . In  $T_1$  columns corresponding to copies of the same attribute of  $U$  are renamings of each other. Therefore, if  $A_i, A_j$  are two copies of the same attribute  $A$ , then the corresponding columns in  $T_1$  have the same number of repeated symbols and in exactly the same sets of rows. Let  $a_1, a_2, \dots, a_n$  be the repeated symbols in a column that corresponds to a copy of attribute  $A$ , and  $S_1, S_2, \dots, S_n$  the sets of rows in which they appear. (One of the  $a_i$ 's might be a distinguished symbol.) Suppose that the rows of  $S_1, \dots, S_k$  have only nondistinguished symbols in the columns that correspond to copies of  $A$  and  $S_{k+1}, \dots, S_n$  have a distinguished symbol in at least one such column. We introduce  $k$  new attributes of  $U$  that are copies of  $A$ . A row in  $S_i (1 \leq i \leq k)$  has a repeated nondistinguished symbol in the  $i$ -th new copy of  $A$  and distinct nondistinguished symbols in the other copies. For each attribute in  $a(T_1) (= a(T))$  we change all repeated nondistinguished symbols into new distinct ones. Finally, we delete all old attributes that are not in  $a(T_1)$ . Let  $T'$  be the constructed extended tableau. In Figure 7 we show an example of this transformation.

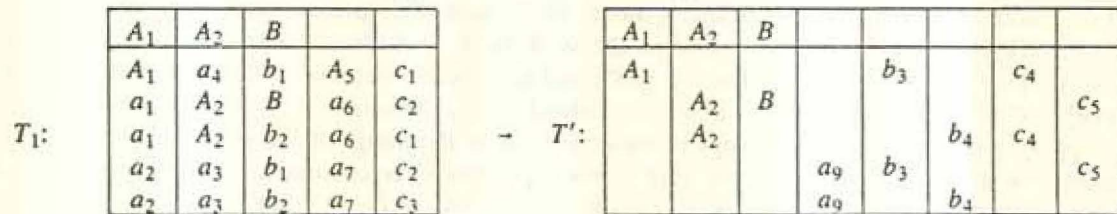


Figure 7

Blanks indicate distinct nondistinguished symbols.

Clearly,  $T'$  is a shallow extended tableau. Let  $X, X'$  be the sets of columns of  $T$  (or  $T_1$ ) and  $T'$  respectively. Let  $T_2$  be obtained from  $T_1$  by adding for each column  $A_i$  in  $X' - X$  a renamed copy of an attribute  $A_j \in X$  with all symbols nondistinguished. Clearly,  $chase_F(T_2) = T_2$ . Let  $T''$  be obtained from  $T'$  by restoring the columns of  $X - X'$  that we deleted. Now  $T_2$  and  $T''$  have the



same set of columns. Let  $T_3 = \text{chase}_F(T'')$ . From the construction of the chase that we described before Lemma 3.1,  $T_3$  is identical to  $T_2$  up to renaming of symbols: just note that for every  $A \in U$ ,  $G_A(T_2)$  and  $G_A(T'')$  are both unions of the same disjoint cliques  $S_1, \dots, S_n$ . Therefore,  $T_2(I) = T_3(I)$  for every relation  $I$  over  $X \cup X'$ , and from Lemmas 3.1 and 3.2 we have  $\bar{f}_T(R) = \bar{f}_{T_1}(R) = \bar{f}_{T_2}(R) = \bar{f}_{T_3}(R) = \bar{f}_{T'}(R) = \bar{f}_T(R)$  for every relation  $R$  over  $U$ .  $\square$

The tableau  $T'$  constructed in the proof of Lemma 3.3 is called the *canonical shallow tableau* for  $T$  and the corresponding shallow expression  $\phi'$  is the *canonical shallow expression* for  $T$ . It is unique up to renaming some of the attributes that are not in  $a = a(\phi') = a(T)$ :  $\phi' = \pi_a(\bowtie_i \pi_{X_i}(\bar{R}))$  is characterized by the fact that no  $X_i$  contains two or more copies of an attribute  $A \in U$  unless they are all in  $a = a(\phi')$ ; furthermore, if two copies  $A_i, A_j \in a$  of  $A$  belong to some  $X_i$ , then they are in exactly the same  $X_j$ 's.

Lemma 3.3 says that the difficulties that arose in the case of usual project-join expressions, due to the fact that not all tableaux come from expressions, cease to exist when we go to extended relations since every extended tableau corresponds to an expression (and one of a very simple form actually) over extended relations.

We will now show a converse to Lemma 3.3 which characterizes the power of (usual) tableaux in algebraic terms.

**Theorem 3.2** Let  $T$  be a (usual) tableau. Then there exists a (shallow) project-join expression  $\phi$  with  $a(\phi) = a(T) \subseteq U$  such that for all relations  $R$ ,  $f_T(R) = \phi(\bar{R})$ . Conversely if  $\phi$  is an expression over extended relations such that  $a(\phi) \subseteq U$ , then there is a (usual) tableau  $T$  such that  $f_T(R) = \phi(\bar{R})$  for all relations  $R$ .

*Proof* The first part follows immediately from Lemma 3.3 since every (usual) tableau  $T$  is also an extended tableau with  $\bar{f}_T = f_T$ .

For the second part, let  $\phi$  be an expression over extended relations with  $a(\phi) \subseteq U$ . From  $\phi$  we construct an extended tableau  $T_\phi$ . Let  $T_1$  be the chase of  $T_\phi$  under  $F$ . We have  $\phi(\bar{R}) = \bar{f}_{T_1}(R)$  for every  $R$ . Let  $T$  be the (usual) tableau obtained from  $T_1$  by keeping only one copy of each attribute in  $U$  (the one with the distinguished symbol if there is one). We claim that  $f_T(R) = \bar{f}_{T_1}(R)$  for every  $R$ . In proof, let  $R$  be a relation over  $U$ . (1) Let  $t \in \bar{f}_{T_1}(R)$ . Let  $X$  be the set of columns of  $T_1$ . There is a valuation  $\rho$  of the symbols of  $T_1$  such that  $\rho(w) \in \pi_X(\bar{R})$  for all  $w \in T_1$ , and  $t = \rho(s(T_1))$ . The restriction  $\rho'$  of  $\rho$  to the symbols of  $T$  satisfies  $\rho'(u) \in R$  for each  $u \in T$  and  $t = \rho(s(T_1)) = \rho'(s(T))$ . (2) Let  $t \in f_T(R)$  and let  $\rho$  be a valuation of the symbols of  $T$  such that  $\rho(u) \in R$  for all  $u \in T$  and  $t = \rho(s(T))$ . Extend  $\rho$  to a valuation  $\rho'$  of all the symbols of  $T_1$  by mapping a symbol in a deleted copy of an attribute  $A$  to the image of the symbol that appears in the same row at the copy of  $A$  that we kept. This is possible since columns in  $T_1$  that are copies of the same attribute are renaming of each other. Clearly,  $\rho'(w) \in \pi_X(\bar{R})$  for each  $w \in T_1$  and  $t = \rho(s(T)) = \rho'(s(T_1))$ .  $\square$

Another interesting consequence of Lemma 3.3 concerns the complexity of the inference problem for data dependencies. An *embedded join dependency* (EJD) is a statement of the form  $\pi_X[\pi_{X_1}(R) \bowtie \pi_{X_2}(R) \bowtie \dots \bowtie \pi_{X_n}(R)] \subseteq \pi_X(R)$ . That is, an EJD is a statement  $\phi(R) \subseteq \pi_{a(\phi)}(R)$  where  $\phi$  is a shallow expression. From Lemma 3.3 every expression over extended relations has an equivalent shallow expression. We say that a set of statements (or dependencies)  $\Sigma$  *logically implies* another statement  $\sigma$ , denoted as  $\Sigma = \sigma$ , where  $\Sigma$  and  $\sigma$  are statements about a single relation  $R$ , if every relation  $R$  satisfying  $\Sigma$  satisfies also  $\sigma$ . The *inference problem* for a class of dependencies is to decide whether a set  $\Sigma$  of such dependencies implies another dependency  $\sigma$  in the class. We will show that the inference problem for dependencies of the form  $\phi(R) \subseteq \pi_{a(\phi)}(R)$  is polynomially reducible to (and thus not significantly harder than) the inference problem for EJD's (on an ordinary relation, not an extension of one).



At first we must extend the definition of the chase to dependencies of the form  $\phi(R) \subseteq \pi_X(R)$ , where  $\phi$  is a project-join expression. Let  $T$  be a tableau and  $\sigma$  a dependency  $\phi(R) \subseteq \pi_X(R)$ . Let  $T_1$  be a tableau obtained from  $\phi(T)$  by adding one column for each attribute in  $a(R) - X$  with all entries distinct new symbols (i.e. not appearing in  $T$ ). An application of the rule for  $\sigma$  to  $T$  is the replacement of  $T$  by  $T \cup T_1$ . If  $\Sigma$  is a set of dependencies and  $T$  a tableau, the chase of  $T$  under  $\Sigma$ ,  $\text{chase}_\Sigma(T)$ , is the result of repeated applications of the  $\sigma$ -rules for each  $\sigma \in \Sigma$  as far as possible. Note that  $\text{chase}_\Sigma(T)$ , which might be an infinite relation, satisfies all dependencies in  $\Sigma$ . The chase is a procedure that searches for a counterexample to an implication  $\Sigma \models \sigma$ . Let  $\sigma = \phi(R) \subseteq \pi_X(R)$  and let  $T$  be the tableau of  $\phi$ . Then  $\Sigma \models \sigma$  iff  $s(T) \in \pi_X(\text{chase}_\Sigma(T))$  [Maier et al. 1979, Sadri and Ullman 1980]. If  $s(T) \notin \pi_X(\text{chase}_\Sigma(T))$  then  $\text{chase}_\Sigma(T)$  does not satisfy  $\sigma$ , and thus it provides a counterexample to the implication  $\Sigma \models \sigma$ . We express this fact as follows.

*Proposition 3.1* (The Chase Partial Decision Procedure, [Maier et al. 1979])

Let  $\Sigma$  and  $\sigma$  be as above. Then  $\Sigma \models \sigma$  if and only if  $s(T) \in \text{chase}_\Sigma(T)$ .  $\square$

In the next section we shall give an interesting *dual* interpretation of the chase.

To extend the theory of deductions to dependencies of the form  $\phi_i(\bar{R}) \subseteq \pi_X(\bar{R})$ , we must somehow capture the semantics of the copies of the attributes of extended relation. Our next Lemma says that this can be done by a set of functional dependencies. In fact, multivalued dependencies are enough (Lemma 3.6)

*Lemma 3.4* Let  $\Sigma = \{\phi_1(\bar{R}) \subseteq \pi_{X_1}(\bar{R}), \dots, \phi_n(\bar{R}) \subseteq \pi_{X_n}(\bar{R})\}$  and  $\sigma = \phi_{n+1}(\bar{R}) \subseteq \pi_{X_{n+1}}(\bar{R})$ . Let  $\Sigma'$  and  $\sigma'$  be as  $\Sigma$  and  $\sigma$  with  $\subseteq$  replaced by  $\subseteq$  (i.e. with the expressions regarded as applied to an ordinary relation). Let  $F$  be the set of functional dependencies  $A_i - A_j$  for distinct copies of the same attribute that appear in some expression in  $\Sigma \cup \{\sigma\}$ . Then  $\Sigma \models \sigma$  if and only if  $\Sigma' \cup F \models \sigma'$ .

*Proof* Let  $X$  be the set of attributes that appear in some  $\phi_i$ .

(if).  $\Sigma' \cup \{F\} \models \sigma'$  means that every relation  $I$  over  $X$  that satisfies the functional dependencies  $F$  and  $\phi_i(I) \subseteq \pi_{X_i}(I)$  for  $i = 1, \dots, n$  satisfies also  $\phi_{n+1}(I) \subseteq \pi_{X_{n+1}}(I)$ . Let  $R$  be any relation over  $U$  satisfying  $\Sigma$ . Then  $I = \pi_X(\bar{R})$  satisfies  $F$  and  $\Sigma'$  and therefore  $\phi_{n+1}(\bar{R}) = \phi_{n+1}(I) \subseteq \pi_{X_{n+1}}(I) = \pi_{X_{n+1}}(\bar{R})$ .

(only if) Suppose that  $\Sigma' \cup F \not\models \sigma'$ . Then there exists a relation  $I$  over  $X$  which satisfies  $\Sigma'$  and  $F$  but not  $\sigma'$ . Since  $I$  satisfies  $F$ , any two columns of  $I$  that correspond to attributes  $A_i, A_j$ , copies of the same attribute  $A$  of  $U$ , are renamings of each other. Let  $R$  be a relation over  $U$  formed by keeping from  $I$  one copy of each attribute of  $U$ . Then  $R$  satisfies  $\Sigma$  but not  $\sigma$ .  $\square$

*Lemma 3.5* Let  $T$  be an extended tableau (viewed as a relation) with set of columns  $X$ . Let  $T'$  be obtained from  $T$  by adding one new copy of each attribute in  $U$  with any symbols as entries, and let  $Y$  be the set of columns of  $T'$ . Let  $F$  be the set of FD's  $A_i - A_j$  for distinct copies  $A_i, A_j \in X$  of the same attribute and  $M$  be the set of MVD's  $A_i \twoheadrightarrow A_j$  for  $A_i, A_j \in Y$  copies of the same attribute. Then  $\text{chase}_F(T) \subseteq \pi_X(\text{chase}_M(T'))$ .

*Proof* It suffices to show how an application of an FD-rule in  $T$  can be simulated by MVD-rules in  $T'$ . Suppose that the FD-rule  $A_i - A_j$  is applied to two rows  $u, v$  of  $T$ . Let  $u_{A_i} = v_{A_i} = a_i, u_{A_j} = a_j, v_{A_j} = a_j'$ . Suppose that all occurrences of  $a_j'$  are replaced by  $a_j$  ( $a_j$  could be a distinguished symbol). Let  $w$  be another row of  $T$  with  $w_{A_j} = a_j'$ . The row  $w$  will be replaced by another row  $w'$  which has  $a_j$  in the  $A_j$  column and agrees with  $w$  in the rest of the columns. We must show how to generate from  $T'$  using the MVD-rules a row whose projection on  $X$  agrees with  $w'$ . Let  $A_k$  be the copy of attribute  $A$  in  $Y - X$ . Applying the MVD-rule for  $A_j \twoheadrightarrow A_k$  to tuples  $w$  and  $v$  of  $T'$  we get a row  $w_1$  that agrees with  $v$  in  $A_k$  and with  $w$  in  $Y - A_k$ . Applying the MVD-rule for  $A_i \twoheadrightarrow A_j$  to tuples  $u$  and  $v$  we get a tuple  $v_1$  that agrees with  $v$  in  $Y - A_j$  and has  $a_j$  in column  $A_j$ . Now,  $w_1$  agrees with  $v_1$  in column  $A_k$ . Applying the MVD-rule for  $A_k \twoheadrightarrow A_j$  to  $w_1$  and  $v_1$  we get a tuple  $w_2$  that agrees with  $w_1$  in  $Y - A_j$  and with  $v_1$  in column  $A_j$ . Thus,  $w_2$  agrees with  $w$  in  $Y - A_j A_k$  and has  $a_j$  in column  $A_j$ . Therefore, its projection to  $X$  is  $w'$ .  $\square$



As a corollary of Lemmas 3.4 and 3.5 we have

**Lemma 3.6** Let  $\Sigma = \{\phi_1(\bar{R}) \underset{\subseteq}{\subseteq} \pi_{X_1}(\bar{R}), \dots, \phi_n(\bar{R}) \underset{\subseteq}{\subseteq} \pi_{X_n}(\bar{R})\}$  and  $\sigma = \phi_{n+1}(\bar{R}) \underset{\subseteq}{\subseteq} \pi_{X_{n+1}}(\bar{R})$ . Let  $\Sigma'$  and  $\sigma'$  be as  $\Sigma$  and  $\sigma$  with  $\underset{\subseteq}{\subseteq}$  replaced by  $\subseteq$  (i.e. with the expressions regarded as applied to an ordinary relation). Let  $Y$  contain the attributes appearing in some  $\phi_i$  and in addition one new copy of each attribute in  $U$ . Let  $M$  be the set of multivalued dependencies (on  $Y$ )  $A_i \twoheadrightarrow A_j$  with  $A_i, A_j \in Y$  distinct copies of the same attribute of  $U$ . Then  $\Sigma \models \sigma$  if and only if  $\Sigma' \cup M \models \sigma'$ .

*Proof* From Lemma 3.4,  $\Sigma \models \sigma$  iff  $\Sigma' \cup F \models \sigma$  where  $F$  are the functional dependencies  $A_i \rightarrow A_j$  with  $A_i, A_j \in X$ , the set of attributes that appear in the  $\phi_i$ 's.

(1) Suppose that  $\Sigma' \cup M \models \sigma'$ . Let  $I$  be a relation on  $X$  satisfying  $\Sigma' \cup F$ . Let  $I'$  be obtained from  $I$  by adding one column for each attribute in  $Y-X$  which is a renamed copy of a column in  $X$  that corresponds to the same attribute of  $U$ . (Since  $I$  satisfies  $F$ , all such columns are renamings of each other). Clearly,  $I'$  satisfies  $\Sigma' \cup F'$  where  $F'$  is the functional form of the MVD's in  $M$ . Therefore,  $I'$  satisfies also  $\Sigma' \cup M$  and  $\sigma'$ . Thus,  $I$  satisfies also  $\sigma'$ . Consequently,  $\Sigma' \cup F \models \sigma'$  and  $\Sigma \models \sigma$ .

(2) Suppose that  $\Sigma' \cup F \models \sigma'$ . Let  $T$  be the tableau of  $\phi_{n+1}$  with set of columns  $X$ , and  $T'$  with set of columns  $Y$ . The tableaux  $T$  and  $T'$  satisfy the assumptions of Lemma 3.5. Since  $\Sigma' \cup F \models \sigma'$ ,  $\text{chase}_{\Sigma' \cup F}(T)$  contains a row  $w$  whose projection to  $X_{n+1}$  is the summary  $s(T)$ . Since  $a(\phi_i) \subseteq X$  for each  $i$ , if the rule for  $\sigma_i' = \phi_i(\bar{R}) \subseteq \pi_{X_i}(\bar{R})$  can be applied to a set of rows of  $T$  to produce a new row  $u$ , then the rule can be applied also to the same set of rows of  $T'$  to produce a new row  $u'$  that agrees with  $u$  on  $X$ . Combining this observation with Lemma 3.5 we conclude that using the rules for  $\Sigma'$  and the MVDs in  $M$  we can generate from  $T'$  a row  $w'$  which agrees with  $u$  on  $X$ . Thus,  $s(T') = s(T) \in \pi_{X_{n+1}}(\text{chase}_{\Sigma' \cup M}(T'))$ , and consequently  $\Sigma' \cup M \models \sigma'$ .  $\square$

Combining now Lemma 3.6 with Lemma 3.3 we have

**Theorem 3.3** Let  $\Sigma = \{\phi_1(\bar{R}) \underset{\subseteq}{\subseteq} \pi_{X_1}(\bar{R}), \dots, \phi_n(\bar{R}) \underset{\subseteq}{\subseteq} \pi_{X_n}(\bar{R})\}$  and  $\sigma = \phi_{n+1}(\bar{R}) \underset{\subseteq}{\subseteq} \pi_{X_{n+1}}(\bar{R})$ . Then we can find a set  $\Gamma$  of embedded join dependencies and another EJD  $\gamma$  (over some set of attributes  $Y$ ) such that  $\Sigma \models \sigma$  if and only if  $\Gamma \models \gamma$ .

*Proof* From Lemma 3.3 we can construct for each  $\phi_i$  an equivalent shallow expression  $\phi_i'$ . Let  $Y$  contain the attributes that appear in all  $\phi_i'$ 's and in addition one more copy of each attribute of  $U$ . The set  $\Gamma$  consists of the set  $M$  of MVD's  $A_i \twoheadrightarrow A_j$  with  $A_i, A_j \in Y$  copies of the same attribute of  $U$ , and the set of EJD's  $\phi_i'(I) \subseteq \pi_{X_i}(I)$ , for  $i = 1, \dots, n$  where  $I$  ranges over relations on  $Y$ . The EJD  $\gamma$  is  $\phi_{n+1}'(I) \subseteq \pi_{X_{n+1}}(I)$ .  $\square$

Note that the transformation of the proof of Theorem 3.2 is polynomial. Thus, the inference problem for dependencies of the form  $\phi(\bar{R}) \underset{\subseteq}{\subseteq} \pi_X(\bar{R})$  is within a polynomial factor of the inference problem for EJD's. At present, however, it is not known whether this inference problem is even decidable.

Let us return now to the axiomatization of identities. We showed in the previous discussion that the functional dependencies in  $F$  (which capture the semantics of extended relations by Lemma 3.2) can be replaced by the corresponding MVD's, at least as far as inference of identities  $\phi(\bar{R}) \underset{\subseteq}{\subseteq} \pi_X(\bar{R})$  is concerned. Let us therefore introduce the following axiom.

A8: for all  $X \subseteq a(\bar{R})$ , and  $A_i, A_j$  copies of the same attribute of  $U$ ,  
 $\pi_{A_i A_j}(\bar{R}) \bowtie \pi_{A_i X}(\bar{R}) = \pi_{A_i A_j X}(\bar{R})$ .

Note that this axiom is the embedded multivalued dependency  $A_j \twoheadrightarrow A_i | X$ .

Let us see how some obvious (and useful) facts about expressions on extended relations can be derived from A8 combined with the axioms of Section 2. If  $\phi$  is an expression we denote by  $p(\phi)$  the set of attributes that appear in  $\phi$ ;  $a(\phi) \subseteq p(\phi)$ . Let  $\phi_{A/X}$  denote the expression obtained from  $\phi$  by replacing all occurrences of  $A$  in  $\phi$  by  $X$ .



- Lemma 3.7* (1) If  $A_i \in a(\phi)$ ,  $A_j \notin p(\phi)$ , then  $A_i A_j \bowtie \phi = \phi_{A_i/A_j}$  can be proved from A1-A8.  
 (2) If  $A_i \in p(\phi)$ ,  $A_j \notin p(\phi)$ , then  $\phi = \pi_X(\phi_{A_i/A_j})$ , where  $X = a(\phi)$ , can be proved from A1-A8.  
 (3) If  $A_i \in p(\phi)$ ,  $A_j \in p(\phi)$ , then  $\pi_X \phi = \pi_X(\phi_{A_i/A_j})$ , where  $X = a(\phi) - A_i$ , can be proved from A1-A8.

*Proof* We prove (1) and (2) by simultaneous induction on the structure of  $\phi$ .

(1) The basis ( $\phi = \pi_{X_A}(\bar{R})$ ) is Axiom A8. For the induction step suppose that  $\phi = \pi_X \sigma$ . Since  $A_i \in a(\phi)$ ,  $A_j \notin p(\phi)$ , we have  $A_i \in X$ , and therefore  $A_i \in a(\sigma)$ ; also  $A_j \notin p(\sigma)$ . Thus, from the induction hypothesis for part (1),  $A_i A_j \bowtie \sigma = \sigma_{A_i/A_j}$ . Since  $a(\sigma) \cap \{A_i, A_j\} = \{A_i\} = X \cap \{A_i, A_j\}$ , from A7b we have:  $A_i A_j \bowtie \pi_X \sigma = \pi_{A_i X}(A_i A_j \bowtie \pi_X \sigma) = \pi_{A_i X}(A_i A_j \bowtie \sigma) = \pi_{A_i X}(\sigma_{A_i/A_j}) = \phi_{A_i/A_j}$ .

Suppose now that  $\phi = \sigma \bowtie \tau$ . We have  $A_i \in a(\sigma) \cup a(\tau)$ ,  $A_j \notin p(\tau), p(\sigma)$ . If  $A_i \in a(\sigma) \cap a(\tau)$ , then  $A_i A_j \bowtie (\sigma \bowtie \tau) = A_i A_j \bowtie A_i A_j \bowtie (\sigma \bowtie \tau) = (A_i A_j \bowtie \sigma) \bowtie (A_i A_j \bowtie \tau) = \sigma_{A_i/A_j} \bowtie \tau_{A_i/A_j}$  (from the induction hypothesis for (1)) =  $\phi_{A_i/A_j}$ .

If  $A_i \in a(\sigma) - a(\tau)$ , then from the induction hypothesis for part (2),  $\tau = \pi_X(\tau_{A_i/A_j}) = \tau_{A_i/A_j}$ , since  $A_i \notin X = a(\tau)$ . Thus,  $A_i A_j \bowtie (\sigma \bowtie \tau) = (A_i A_j \bowtie \sigma) \bowtie \tau = \sigma_{A_i/A_j} \bowtie \tau_{A_i/A_j} = \phi_{A_i/A_j}$ .

(2) The basis ( $\phi = \pi_Y(\bar{R})$ ) follows from A1. For the induction step, the case  $\phi = \pi_Y \sigma$  follows also from A1. Suppose therefore that  $\phi = \sigma \bowtie \tau$ . If  $A_i \notin a(\phi)$ , then from the induction hypothesis for (2) and A1b we have  $\sigma = \sigma_{A_i/A_j}$  and  $\tau = \tau_{A_i/A_j}$ . Thus,  $\pi_X \phi = \phi = \sigma \bowtie \tau = \sigma_{A_i/A_j} \bowtie \tau_{A_i/A_j} = \phi_{A_i/A_j} = \pi_X(\phi_{A_i/A_j})$ .

If  $A_i \in a(\sigma) \cap a(\tau)$  (the case that  $A_i \in a(\sigma) - a(\tau)$  is similar), we have from the induction hypothesis for (1):  $\phi_{A_i/A_j} = \sigma_{A_i/A_j} \bowtie \tau_{A_i/A_j} = A_i A_j \bowtie \sigma \bowtie \tau = A_i A_j \bowtie \phi$ . Thus,  $\pi_X(\phi_{A_i/A_j}) = \pi_X(A_i A_j \bowtie \phi) = \pi_X(A_i \bowtie \phi)$  (by A7b since  $A_j \notin a(\phi), X$ ) =  $\pi_X \phi$ , (by A2 since  $A_i \in a(\phi)$ ) =  $\phi$ , since  $X = a(\phi)$ .

(3) From (2) and A1 we have  $\pi_X \phi = \pi_X(\phi_{A_i/A_j})$ . Let  $\psi = \phi_{A_i/A_j}$ . Now  $A_j \in p(\psi)$ ,  $A_i \notin p(\psi)$ . Thus, again from (2),  $\pi_X(\psi) = \pi_X(\psi_{A_i/A_j})$ . But  $\phi_{A_i/A_j} = \psi_{A_i/A_j}$ . Thus,  $\pi_X \phi = \pi_X(\phi_{A_i/A_j})$ .  $\square$

*Lemma 3.8* Let  $\phi$  be a project-join expression, and let  $\phi'$  be the canonical shallow expression for  $T_\phi$ . Then  $\phi(\bar{R}) = \phi'(\bar{R})$  can be proved by A1-A8.

*Proof* We proceed in the tree for  $\phi$  from the leaves to the root changing names of attributes so that the tableau of the resulting expression is shallow. Let  $v$  be a node of the tree of  $\phi$  and suppose that we have already modified the subtrees that are rooted in descendants of  $v$  so that for each such subtree representing expression  $\sigma$ , the tableau  $T_\sigma$  is shallow, and  $p(\sigma) - a(\sigma)$  does not contain any attributes appearing outside  $\sigma$ . Let  $\psi$  be the expression for the subtree rooted at  $v$ .

*Case 1*  $\psi = \pi_X \sigma$ , where  $\sigma$  is the (modified) shallow expression of the subtree rooted at the son of  $v$ .

For each attribute  $A_i$  in  $a(\sigma) - X$  that appears in another node that is not a descendant of  $v$  we introduce a new attribute  $A_j$  that appears nowhere else in the tree. Using part (3) of Lemma 3.7 we transform  $\psi$  into  $\psi' = \pi_X(\sigma_{A_i/A_j})$ . Note that  $T_{\psi'}$  is shallow, and  $p(\psi') - a(\psi')$  does not contain any attributes appearing in the rest of the tree.

*Case 2*  $\psi = \sigma \bowtie \tau$  where  $\sigma$  and  $\tau$  are the (modified) shallow expressions corresponding to the subtrees rooted at the sons of  $v$ .

Since  $p(\sigma) - a(\sigma)$  (resp.  $p(\tau) - a(\tau)$ ) does not contain any attributes appearing outside  $\sigma$  (resp.  $\tau$ ), we have  $p(\sigma) \cap p(\tau) = a(\sigma) \cap a(\tau)$ , and therefore,  $T_\psi$  is shallow. Also  $p(\psi) - a(\psi) = [p(\sigma) - a(\sigma)] \cup [p(\tau) - a(\tau)]$  does not contain any attributes appearing outside  $\psi$ .

Proceeding bottom-up in this way we transform  $\phi$  to  $\phi_1$  with a shallow tableaux. If we move now all projections in  $\phi_1$  that create repeated nondistinguished symbols to the root, move all



projections that create nonrepeated nondistinguished symbols to the leaves - possible using A7b since  $T_{\phi_1}$  is shallow - and use associativity of join to collect all joins in one node we will get a shallow expression  $\phi_2$  for  $T_{\phi_1}$ .

Let  $\phi_2 = \pi_a(\bowtie_k \pi_{X_k}(\bar{R}))$  where  $a = a(\phi_2) = a(\phi)$ . Each column of  $T = T_{\phi_2}$  has either a distinguished symbol or one repeated nondistinguished symbol. If two columns  $A_i, A_j$ , copies of  $A$  have a repeated symbol in exactly the same rows (i.e.  $A_i$  is a renaming of  $A_j$ ) then  $A_i$  appears in exactly the same  $X_k$ 's as  $A_j$ . If  $A_j \notin a$  then we can delete  $A_j$  using part (2) of Lemma 3.7; i.e.  $\phi_2$  is  $\pi_a(\tau_{A_i/A_j})$  for some  $\tau$ . If two columns  $A_i, A_j$  with  $A_j \in a$  have both a repeated symbol in some row  $w$ , then we can "merge"  $A_i$  and  $A_j$  as follows. Let  $S_1, S_2$  be the set of rows that have a repeated symbol respectively in  $A_i, A_j$ . We have  $A_i \in X_k$  for  $k \in S_1, A_j \in X_k$  for  $k \in S_2, A_i, A_j \in X_k$  for  $k \in S_1, S_2$ . Since  $w \in S_1 \cap S_2$  we have  $\pi_{X_k}(\bar{R}) = \pi_{X_k}(\bar{R}) \bowtie_{A_i A_j}$  (by A2 and Example 2.1). Thus,

$$\begin{aligned} \phi_2 &= \pi_a \left[ \left( \bowtie_{k \in S_1} \pi_{X_k}(\bar{R}) \right) \bowtie \left( \bowtie_{k \in S_2} \pi_{X_k}(\bar{R}) \right) \bowtie_{A_i A_j} \left( \bowtie_{k \in S_1, S_2} \pi_{X_k}(\bar{R}) \right) \right] \\ &= \pi_a \left[ \left( \bowtie_{k \in S_1 - S_2} (A_i A_j \bowtie \pi_{X_k}(\bar{R})) \right) \bowtie \left( \bowtie_{k \in S_2 - S_1} (A_i A_j \bowtie \pi_{X_k}(\bar{R})) \right) \bowtie \left( \bowtie_{k \in S_1 \cap S_2} \pi_{X_k}(\bar{R}) \right) \bowtie \left( \bowtie_{k \in S_1, S_2} \pi_{X_k}(\bar{R}) \right) \right] \\ &= (\text{by A8}) \pi_a \left[ \left( \bowtie_{k \in S_1 - S_2} \pi_{A_i X_k}(\bar{R}) \right) \bowtie \left( \bowtie_{k \in S_2 - S_1} \pi_{A_j X_k}(\bar{R}) \right) \bowtie \left( \bowtie_{k \in S_1 \cap S_2} \pi_{X_k}(\bar{R}) \right) \bowtie \left( \bowtie_{k \in S_1, S_2} \pi_{X_k}(\bar{R}) \right) \right]. \end{aligned}$$

In the tableau of the last expression the columns  $A_i A_j$  have become identical up to renaming of symbols. Thus, we can delete  $A_j$  as above. Continuing this procedure we end up with a shallow expression  $\psi_2$  such that the sets of rows in which any two columns  $A_i A_j$  with  $A_j \in a$  have a repeated symbol are disjoint. The expression  $\psi_2$  is the canonical shallow expression  $\phi'$  of  $\phi$  up to changing the names of some  $A_i \in a$  (which can be done using (3) of Lemma 3.7.)  $\square$

Using Lemma 3.8 we can show:

**Theorem 3.4** All valid identities  $\phi_1(\bar{R}) \subseteq \phi_2(\bar{R})$  are provable from A1-A8.

*Proof* From Lemma 3.8 we can transform  $\phi_1$  and  $\phi_2$  to their canonical shallow expressions  $\psi_1'$  and  $\psi_2'$ . Thus, it suffices to prove the theorem for canonical shallow expressions. Let  $\phi_1, \phi_2$  be two such expressions and  $T_1, T_2$  their tableaux. We can assume that  $[p(\phi_1) - a(\phi_1)] \cap [p(\phi_2) - a(\phi_2)] = \emptyset$ ; if not change the names of the attributes in  $p(\phi_1) - a(\phi_1)$  using (3) of Lemma 3.7.

Let  $A$  be an attribute of  $U$  and  $A_1, A_2, \dots, A_l$  its copies in  $p(\phi_1), A_1', A_2', \dots, A_l'$  its copies in  $p(\phi_2)$ . Suppose that  $A_1 = A_1', \dots, A_m = A_m'$  are the copies that are in  $a = a(\phi_1) = a(\phi_2)$ . Let  $N_i$  (resp.  $M_j$ ) be the set of rows of  $T_1$  (resp.  $T_2$ ) that have a distinguished or repeated nondistinguished symbol in column  $A_i$  (resp.  $A_j'$ ). Since  $\phi_1$  and  $\phi_2$  are canonical shallow expressions we have  $N_i \cap N_j = \emptyset$ , unless  $A_i, A_j \in a$  and  $N_i = N_j$  - similarly for the  $M_j$ 's. Let  $T_1', T_2'$  be  $T_1$  and  $T_2$  padded out with new columns of distinct nondistinguished symbols to  $X = p(\phi_1) \cup p(\phi_2)$ , and let  $\bar{T}_1 = \text{chase}_F(T_1'), \bar{T}_2 = \text{chase}_F(T_2')$ . From Lemmas 3.2 and 2.1 there is a homomorphism  $h$  from  $\bar{T}_2$  to  $\bar{T}_1$ . We will identify a row  $u$  of  $T_i$  ( $i=1,2$ ) with the corresponding row of  $\bar{T}_i$  and leaf of  $\phi_i$ . A column  $A_i$  (copy of  $A$ ) of  $\bar{T}_1$  has one different repeated symbol for each  $N_i$  and distinct nondistinguished symbols in the other rows - similarly with  $\bar{T}_2$ . Therefore, for each  $i = 1, \dots, l$ , either  $h(M_i) \subseteq h(N_j)$  for some  $j$ , or  $h(M_i)$  has a single row.

We carry out the following procedure for all  $A \in U$ .

(1) At first we group together the leaves of  $\phi_2$  that belong to the same  $M_i$ ; i.e. we write  $\phi_2$  as  $\pi_a(\tau_0 \bowtie \tau_{m+1} \bowtie \dots \bowtie \tau_l)$  where  $\tau_i$  for  $i > m$  is the join of the leaves in  $M_i$  and  $\tau_0$  the join of the rest of the leaves. Since  $A_i'$  appears only in  $\tau_i$  we can insert the projection that deletes  $A_i'$  for  $i > m$  into



the join; i.e.  $\phi_2 = \pi_a(\tau_0 \bowtie \pi_{Z_{m+1}} \tau_{m+1} \bowtie \dots \bowtie \pi_{Z_i} \tau_i)$  by A7b, where  $Z_i = a(\tau_i) - A_i'$ . Suppose that  $M_i$  is mapped into  $N_j$ ; using (3) of Lemma 3.7 we change  $A_i'$  into  $A_j$  in  $\pi_{Z_i} \tau_i$ . After doing this for all  $M_i$  with  $h(M_i) \subseteq N_j$  for some  $j$ , we move the projections to  $Z_i$ 's back to the root and thereby shrink the expression by A7a. Let  $\phi_2'$  be the resulting expression.

(2) Suppose that  $h(M_i)$  is not in  $N_j$  for any  $j$ . Then  $h(M_i) = \{u\}$  for some leaf  $u$  of  $\phi_1$  whose projection does not contain any copy of  $A$ . We include  $A_i'$  in the projection at  $u$ ; since  $A_i'$  does not appear in any other leaf this preserves equivalence by A7b. Let  $\phi_1'$  be the expression that results by doing this for all  $M_i$  that are mapped by  $h$  into  $N_j$ .

We have  $\phi_2' \subseteq \phi_2$  and  $\phi_1' = \phi_1$  provable from the axioms. Every leaf  $u$  of  $\phi_2'$  (corresponding to a leaf of  $\phi_2$ ) is mapped by  $h$  to a leaf  $h(u)$  of  $\phi_1'$  that contains the copy (or copies if they are in  $a$ ) of  $A$  that  $u$  contains.

Let  $\psi_1$  and  $\psi_2$  be the expressions that are constructed from  $\phi_1$  and  $\phi_2$  by doing the previous procedure for all  $A \in U$ . We have  $\psi_2 \subseteq \phi_2$ ,  $\psi_1 = \phi_1$ , and every leaf  $\pi_{X_i}(R)$  of  $\psi_2$  is mapped to a leaf  $\pi_{Y_j}(\bar{R})$  of  $\psi_1$  with  $X_i \subseteq Y_j$ . We replace  $\pi_{X_i}$  by  $\pi_{Y_j}$  in each leaf of  $\psi_2$  to get an expression  $\psi_2'$  with  $\psi_2' \subseteq \psi_2$  by A7a. Then we replace identical leaves with one of them to get  $\psi_2'' = \pi_a \sigma = \psi_2' \subseteq \psi_2 \subseteq \phi_2$  (by A2). Every leaf of  $\psi_2''$  is now identical to a distinct leaf of  $\psi_1$ ; i.e.  $\psi_1 = \pi_a(\sigma \bowtie \tau)$  for some  $\tau$ . From A2 we have  $\pi_a(\sigma \bowtie \tau) \subseteq \pi_a(\sigma)$ , and thus  $\phi_1 = \psi_1 \subseteq \psi_2'' \subseteq \phi_2$ .  $\square$

#### 4. ALGEBRAIC DEPENDENCIES

*Definition* An algebraic dependency is an assertion of the form

$$\phi_1(\bar{R}) \underset{f}{\subseteq} \phi_2(\bar{R})$$

where  $\phi_1$  and  $\phi_2$  are project-join expressions.  $\square$

*Example 4.1* The multivalued dependencies are special cases of algebraic dependencies. In fact, so are the far more general embedded join dependencies since the EJD on  $X_1, \dots, X_k$  embedded in  $X$  can be expressed as

$$\pi_X \left( \pi_{X_1}(\bar{R}) \bowtie \dots \bowtie \pi_{X_k}(\bar{R}) \right) \underset{f}{\subseteq} \pi_X(\bar{R}).$$

$\square$

*Example 4.2* We have already seen that the functional dependencies are algebraic. For example,  $A \rightarrow B$  can be expressed as

$$\pi_{B, B_2} \left( \pi_{AB_1}(\bar{R}) \bowtie \pi_{AB_2}(\bar{R}) \right) \underset{f}{\subseteq} \pi_{B, B_2}(\bar{R}).$$

We can say, informally, that the language of algebraic dependencies possesses some form of equality.  $\square$

*Example 4.3* Transitive dependencies [Paradaens 1979] are algebraic. For example, the dependency  $Tr(X, Y, Z)$  can be expressed as

$$\pi_{XZ} \left( \pi_{XY}(\bar{R}) \bowtie \pi_{YZ}(\bar{R}) \right) \underset{f}{\subseteq} \pi_{XZ}(\bar{R}).$$

$\square$

*Example 4.4* Any template dependency [Sadri and Ullman 1980] can be rendered as an algebraic dependency. Let  $T$  be a tableau defining a template dependency. Let  $\phi$  be the corresponding canonical shallow expression. Then the equivalent algebraic dependency is

$$\phi(\bar{R}) \underset{f}{\subseteq} \pi_{\sigma(T)}(\bar{R}).$$

$\square$

Apparently, the algebraic dependencies are quite general. More importantly, we shall show that if  $\Sigma \cup \{\sigma\}$  is a set of algebraic dependencies, and furthermore  $\Sigma \models \sigma$  (that is, all relations satisfying  $\Sigma$  must also satisfy  $\sigma$ ) then  $\sigma$  is derivable from  $\Sigma$  by A1-8. This strongly suggests that the notion of algebraic dependency is the natural conclusion of the search for a general axiomatizable data dependency.

In order to show the completeness of A1-A8 for algebraic dependencies, we first revisit the chase (recall Proposition 3.1). The chase is essentially a combinatorial construction of a counterexample to an implication  $\Sigma \models \sigma$ .

*Example 4.5* Let us prove pseudotransitivity of MVD's (recall Example 2.2) by using the chase.  $k = 2$ ,  $\phi_1 = \pi_{XY}(R) \bowtie \pi_{X(U-XY)}(R)$ ,  $\phi_2 = \pi_{YZ}(R) \bowtie \pi_{Y(U-YZ)}(R)$ ,  $\phi_3 = \pi_{X(Z-Y)} \bowtie \pi_{XYW}$ , where  $W = U - XYZ$ .

$T_3$  is shown in Figure 8(a) - where we have, for simplicity, one attribute for each set of attributes  $X, Y, Z - Y$  and  $W$ , respectively labeled  $X, Y, Z$  and  $W$ .

If we apply  $\phi_1$  to  $T_3$  we obtain the relation (tableau) shown in Figure 8(b); if we apply  $\phi_2$  to that we get the relation, of Figure 8(c). Since  $(X, Y, Z, W) \in \phi_2(\phi_1(T_3))$  we conclude that  $(X, Y, Z, W) \in \text{chase}(T_3) \supseteq \phi_2(\phi_1(T_3))$ , and hence we have shown that  $\Sigma \models \sigma_3$ .  $\square$

We introduce below another Proposition, (cf. Proposition 3.1) which shows the chase under a different light: as an algebraic construction of a proof of the implication  $\Sigma \models \sigma_{k+1}$ . This point of



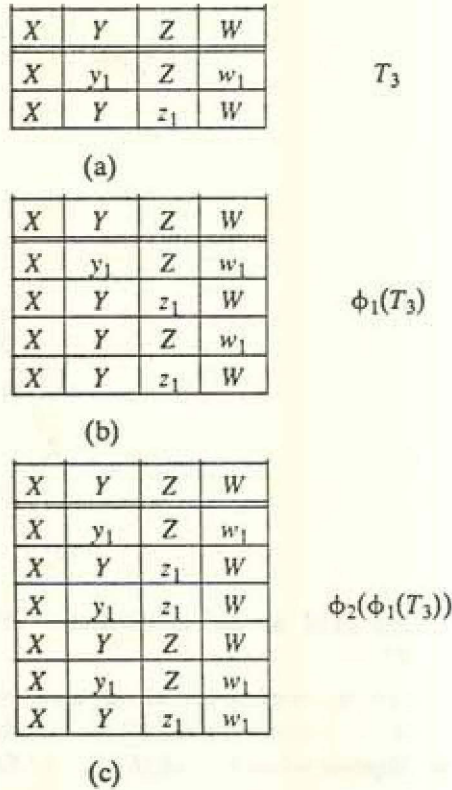


Figure 8

view is central in the proofs that follow.

We call  $\psi(R)$  a *substitution* of  $\phi_1, \dots, \phi_k$  if either  $\psi(R) = R$  or if  $\psi$  is some  $\phi_j$  applied (recursively) to other substitutions.

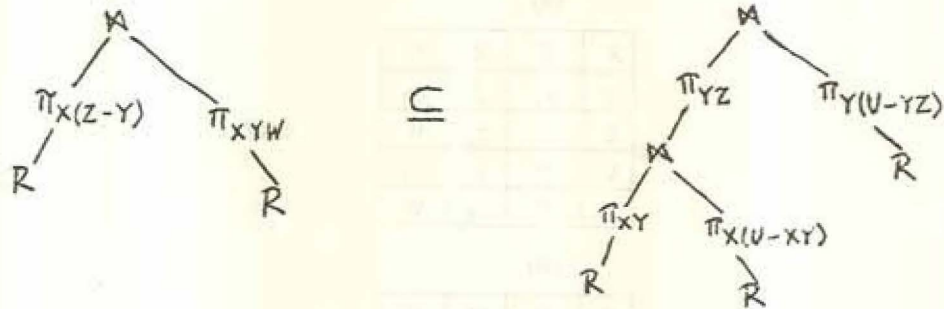
*Proposition 4.1* (The Dual Interpretation of the Chase)

Let  $\Sigma$  and  $\sigma_{k+1}$  be as above. Then,  $\Sigma \models \sigma_{k+1}$  iff there is a substitution  $\psi$  of  $\phi_1, \dots, \phi_k$  such that  $\phi_{k+1}(R) \subseteq \psi(R)$  is a tautology.

*Proof* If such a substitution exists, then,  $\Sigma \models \psi(R) \subseteq R$  by monotonicity; thus,  $\Sigma \models \phi_{k+1} \subseteq \psi(R) \subseteq R \Leftrightarrow \sigma_{k+1}$ .

For the other direction, suppose that  $\Sigma \models \sigma_{k+1}$ . By proposition 3.1, chase  $(T_{k+1})$  contains the tuple  $(A, B, \dots, Z)$ . We shall assign a substitution  $\psi_t$  to each tuple  $t$  of chase  $(T_{k+1})$ . If  $t \in T_{k+1}$ , then  $\psi_t = R$ . Otherwise,  $t$  was obtained by applying some  $\phi_j$  to tuples  $t_1, \dots, t_l$ . Then  $\psi_t$  is defined, recursively as  $\phi_j$  applied to  $\psi_{t_1}, \dots, \psi_{t_l}$ . Now let  $\psi$  be the substitution associated with  $(A, B, \dots, Z)$ . We claim that  $\phi_{k+1}(R) \subseteq \psi(R)$  is a tautology. But this follows from a result of [Aho et al. 1979], which states that  $\phi_{k+1}(R) \subseteq \psi(R)$  iff  $(A, B, \dots, Z) \in \psi(T_{k+1})$ ; and this is true by the construction of  $\psi$ .  $\square$

*Example 4.5* (continued) To show that  $\{\sigma_1, \sigma_2\} \models \sigma_3$  it would suffice to observe that the following inclusion is tautologically true.



The right-hand expression is recognized as the substitution - of  $\phi_1$  and  $R$  into  $\phi_2$  - which corresponds to the tuple  $(A, B, \dots, Z)$ .  $\square$

We now embark on our proof of completeness of our axiomatic system. Recall the set  $F$  of functional dependencies defined in the previous section:  $F = \{A_i \rightarrow A_j : A \in U, i, j = 1, 2, \dots\}$ . As in Lemma 3.4, a set of algebraic dependencies  $\Sigma = \{\phi_i(\bar{R}) \subseteq \psi_i(\bar{R}) \mid i = 1, \dots, n\}$  logically implies another dependency  $\sigma \Leftrightarrow \phi_{n+1}(\bar{R}) \subseteq \psi_{n+1}(\bar{R})$  if and only if  $\Sigma' \cup F \models \sigma'$ , where  $\Sigma'$  and  $\sigma'$  are as  $\Sigma$  and  $\sigma$  with  $\subseteq$  replaced by  $\subsetneq$ .

**Theorem 4.1** Let  $\sigma_i \Leftrightarrow \phi_i(\bar{R}) \subseteq \psi_i(\bar{R})$ ,  $i = 1, \dots, k+1$  be algebraic dependencies. Then any implication  $\Sigma = \{\sigma_1, \dots, \sigma_k\} \models \sigma_{k+1}$  can be proved by the Axioms A1-A8.

*Proof* Let  $X$  be the set of attributes that appear in the  $\sigma_i$ 's. From our previous discussion,  $\Sigma \models \sigma$  iff  $\Sigma' \cup F \models \sigma'$  where  $\sigma'$  and  $\Sigma'$  are dependencies of the form  $\phi(I) \subseteq \psi(I)$  where  $I$  ranges over all relations on  $X$ . The chase is a decision procedure that can be applied also for dependencies of this form. In this case, we start with the tableau  $T_0 = T_{\phi_{k+1}}$  corresponding to  $\phi_{k+1}$  (padded with distinct nondistinguished variables to the set  $X$  of attributes). The rule for FD's is as in Section 3. The rule for a dependency  $\phi_i(I) \subseteq \psi_i(I)$  is as follows. Suppose that there is a valuation  $\rho$  from  $T_{\phi_i}$  into the current tableau  $T$  -  $\rho$  can map distinguished to nondistinguished symbols, i.e. all we require is that  $\rho(s(T_{\phi_i})) \in \phi_i(T)$ . Extend  $\rho$  to the nondistinguished variables of  $T_{\psi_i}$  by assigning to each one of them a distinct nondistinguished variable that does not appear in  $T$ . An application of the rule for  $\phi_i(I) \subseteq \psi_i(I)$  is the addition to  $T$  of the rows of  $T_{\psi_i}$  with this valuation. The chase of  $T_0$  under a set of dependencies is the result of the repeated application of these rules to  $T_0$ ; note that the chase might be an infinite tableau. Now,  $\Sigma' \cup F \models \sigma'$  iff there is a valuation from  $T_{\psi_{k+1}}$  to  $\text{chase}_{\Sigma' \cup F}(T_{\phi_{k+1}})$  that maps distinguished symbols to distinguished symbols (a homomorphism), i.e. if and only if  $s(T_{\psi_{k+1}}) \in \psi_{k+1}(\text{chase}_{\Sigma' \cup F}(T_{\phi_{k+1}}))$ .

Suppose now that  $\Sigma' \cup F \models \sigma'$ . Then there is a finite  $n$  such that the tableau  $T'$  that results after the application of  $n$  rules contains the image of a valuation of  $T_{\psi_{k+1}}$  that preserves distinguished symbols (i.e. a homomorphism). Let us construct the chase by applying the FD's as far as possible between any two consecutive application of a rule for a dependency of  $\Sigma'$ . That is, we have a sequence of tableaux  $T_0 = T_{\phi_{k+1}}, T_0', T_1, T_1', \dots, T_n, T_n'$ , where  $T_i' = \text{chase}_F(T_i)$  and  $T_{i+1}$  is obtained from  $T_i'$  by a single application of a rule for some  $\sigma_j' \in \Sigma'$ . Let  $\chi_i$  be the canonical shallow expression for  $T_i$  (and  $T_i'$ ). From Lemma 3.8 we have  $\phi_{k+1} \subseteq \chi_0$  provable from A1-A8. Since



there is a homomorphism from  $T_{\phi_{k+1}}$  into  $T_{n'}$  we have  $\chi_n \subseteq_{\epsilon} \psi_{k+1}$  and from Theorem 3.4 this identity can be proved from A1-A8. Thus, it suffices to prove that  $\chi_i \subseteq_{\epsilon} \chi_{i+1}$  can be derived from  $\Sigma$  using A1-A8. Our proof uses essentially the ideas of Proposition 4.1 where the substitution must take into account that the dependencies are not full.

Suppose that  $T_{i+1}$  is obtained from  $T_i'$  by an application of the rule for dependency  $\sigma_j$ . For each attribute  $A_i$  in  $a = a(\phi_j) = a(\psi_j)$  we introduce one new attribute  $A_i'$  that does not appear in  $X$  or  $\chi_i$ . Let  $a'$  be the set of these new attributes. Let  $\rho$  be a valuation from  $T_{\phi_j}$  into  $T_i'$ . If  $\chi_i = \pi_{a(\phi_{i-1})} \left[ \bowtie_i \pi_{Y_i} \right]$ , let  $\theta_i$  be  $\pi_{a(\phi_{i-1})} \cup a' \left[ \bowtie_i \pi_{Z_i} \right]$ , where  $Z_i$  contains  $Y_i$  and those attributes  $A_i'$  for which a row of  $T_{\phi_j}$ , mapped by  $\rho$  into  $t$  has a distinguished symbol in  $A_i$ .

Let  $\bar{\phi}_j, \bar{\psi}_j$  be  $\phi_j$  and  $\psi_j$  with the attributes in  $a$  renamed into  $a'$ . From Lemma 3.7 we have  $\phi_j \bowtie_{A_i \in a} (\bowtie_{A_i' \in a'}) = (\phi_j)_{a|aa'}$ , and  $\psi_j \bowtie_{A_i \in a} (\bowtie_{A_i' \in a'}) = (\psi_j)_{a|aa'}$ . Thus, from  $\phi_j \subseteq_{\epsilon} \psi_j$  we can derive (using A1-A8) that  $(\phi_j)_{a|aa'} \subseteq_{\epsilon} (\psi_j)_{a|aa'}$ . But  $\bar{\phi}_j = \pi_{a'} \left[ (\phi_j)_{a|aa'} \right]$  and  $\bar{\psi}_j = \pi_{a'} \left[ (\psi_j)_{a|aa'} \right]$ . Thus, from Theorem 3.4 we can derive  $\bar{\phi}_j \subseteq_{\epsilon} \bar{\psi}_j$ . Let  $\bar{\theta}_i = \pi_{a'} \theta_i$ . The valuation  $\rho$  is now a homomorphism from  $T_{\bar{\phi}_j}$  into the chase under  $F$  of  $T_{\bar{\theta}_i}$ . Thus,  $\bar{\theta}_i \subseteq_{\epsilon} \bar{\phi}_j \subseteq_{\epsilon} \bar{\psi}_j$ . Therefore, from Axiom A2,  $\theta_i = \theta_i \bowtie_{\pi_{a'} \theta_i} \subseteq_{\epsilon} \theta_i \bowtie_{\bar{\psi}_j}$ , and  $\pi_{a(\phi_{i-1})}(\theta_i) \subseteq_{\epsilon} \pi_{a(\phi_{i-1})}(\theta_i \bowtie_{\bar{\psi}_j})$ . Since  $\rho$  was a valuation from  $T_{\phi_j}$  into  $T_i'$ , the canonical shallow expression for  $\pi_{a(\phi_{i-1})}(\theta_i)$  is  $\chi_i$  (i.e. the chase of  $\pi_{a(\phi_{i-1})}(\theta_i)$  under  $F$  will not identify any symbols of  $T_i'$ ). Thus,  $\chi_i = \pi_{a(\phi_{i-1})}(\theta_i)$ . On the other hand the rules for the FD's will copy the portion of  $T_{\phi_j}$  that is in the attributes  $a'$  in the tableau  $\bar{T}$  for  $\pi_{a(\phi_{i-1})}(\theta_i \bowtie_{\bar{\psi}_j})$  into the attributes  $a$ . Therefore, the chase $_F$  of  $\bar{T}$  (restricted to  $X$ ) will be exactly  $T_{i+1}'$ , and  $\chi_{i+1} = \pi_{a(\phi_{i-1})}(\theta_i \bowtie_{\bar{\psi}_j})$ . Thus,  $\chi_i \subseteq_{\epsilon} \chi_{i+1}$  can be derived from  $\sigma_j$  and the axioms.  $\square$

## 5. EXPRESSIVE POWER

In this Section we briefly examine algebraic and related dependencies from a model-theoretic viewpoint. In order to prove an interesting result, we are forced to expand our algebraic language to contain the operations of union and difference. The goal of this section is twofold. First, by exhibiting the power of the expanded language we further justify the usefulness of "equational" dependencies such as algebraic dependencies. Second, we point the way towards a host of interesting open model-theoretic questions concerning data dependencies.

Let  $P \subseteq 2^{D(A) \times D(B) \times \dots}$  be a predicate on finite relations. We say that  $P$  is *domain-independent* if, whenever  $R \in P$  and  $h$  is a set of permutations of  $D(A), D(B)$ , etc. then  $h(R) \in P$ . If  $P$  is domain-independent, its *index* is the number of equivalence classes in which  $P$  is divided if one considers  $R \equiv R'$  whenever  $R'$  is a "renaming"  $h(R)$  of  $R$ , as above.

*Theorem 5.1* Let  $P$  be any domain-independent predicate of finite index. Then there is an expression  $\phi_P$  over project, join, union and difference such that

$$R \in P \text{ iff } \phi_P(\bar{R}) = R.$$

*Proof* Let  $E_1, E_2, \dots, E_m$  be the equivalence classes of  $P$ . For each  $E_j$  we are going to construct an expression  $\epsilon_j$  such that for all relations  $R$

$$\epsilon_j(\bar{R}) = \begin{cases} R & \text{if } R \in E_j \\ \emptyset & \text{otherwise} \end{cases}$$

The Theorem would then follow, since  $\bigcup_{j=1}^m \epsilon_j(\bar{R})$  would be the required expression for  $P$ .

Consider therefore an equivalence class  $E_j$ . Intuitively, if  $R \in E_j$  then

- a.  $R$  has a fixed number  $k_j$  of tuples, and
- b.  $R$ 's tuples conform to a fixed "pattern".

Let us first construct an expression  $\phi_k$  such that

$$\phi_k(\bar{R}) = \begin{cases} R & \text{if } R \text{ has } k \text{ tuples} \\ \emptyset & \text{otherwise.} \end{cases}$$

Consider the expression

$$\phi_k'(\bar{R}) = \pi_{U_i} \left( R_1 \bowtie R_2 \dots \bowtie R_k - \bigcup_{1 \leq i < j \leq k} R_i R_j \bowtie (\bigwedge_{i \neq l \neq j} R_l) \right)$$

Here  $R_i$  means  $\pi_{U_i}(\bar{R})$ , the  $i^{\text{th}}$  copy of  $R$ . Then we have

$$\phi_k'(\bar{R}) = \begin{cases} R & \text{if } R \text{ has at least } k \text{ tuples} \\ \emptyset & \text{otherwise} \end{cases},$$

because if  $R$  has  $k$  tuples  $t_1, \dots, t_k$  then the join contains a tuple  $(t_1, t_2, \dots, t_k)$ , not contained in the union; similarly for the tuples  $(t_2, t_3, \dots, t_k, t_1)$ , etc. On the other hand, if  $R$  has fewer than  $k$  tuples, then the join is a subset of the union. Finally, we may define

$$\phi_k(\bar{R}) = \left( R - \phi_k'(\bar{R}) \right) - \left( R - \phi_{k-1}'(\bar{R}) \right).$$

Let now  $r = \{t_1, \dots, t_k\}$  be any relation in  $E_j$ .

Let the domain elements of  $A$  that appear in  $r$  be  $a_1, a_2$ , etc. We define for each  $i \leq k$ , the following subset of  $U$ :



$$\epsilon_j(\bar{R}) = \phi_k(\bar{R}) \times \left[ \bigcup_{i=1}^{k_j} \pi_{X_i} \left( \bigotimes_{i=1}^{k_j} \pi_{X_i}(\bar{R}) \bigotimes_{\substack{1 \leq p < q \leq k_j \\ A \in U}} (A_p \times A_q - A_p A_q) \right) \right]$$

The first part of  $\epsilon_j$  guarantees that  $R$  has  $k_j$  rows. The  $i^{\text{th}}$  argument of the union is either empty, or the relation consisting of the  $i^{\text{th}}$  row of  $r$ , in case that there is a mapping  $h$  from the domains of  $R$  to those of  $r$  that creates all rows of  $r$ . The second part of each argument prevents any two domain elements to be mapped by  $h$  to the same domain element of  $r$ , and thus  $h$  has to be a renaming. Since  $R$  has also  $k_j$  rows, it follows that  $\epsilon_j(\bar{R}) = R$  if and only if  $R$  is a renaming of  $r$ , and  $\epsilon_j(\bar{R})$  is empty otherwise. This completes the construction and the proof.  $\square$

*Example 5.1*

Let  $r = \{(a_1, b_1), (a_1, b_2), (a_2, b_3)\}$ .  $\epsilon_j$  is as shown.

$$\begin{aligned} \epsilon_j(\bar{R}) = \phi_3(\bar{R}) \times & \left[ \pi_{A_1 B_1} (A_1 B_1 \times A_1 B_2 \times A_2 B_3 \times \delta) \cup \right. \\ & \pi_{A_1 B_2} (A_1 B_1 \times A_1 B_2 \times A_2 B_3 \times \delta) \cup \\ & \left. \pi_{A_2 B_3} (A_1 B_1 \times A_1 B_2 \times A_2 B_3 \times \delta) \right], \end{aligned}$$

where  $\delta = (A_1 \times A_2 - A_1 A_2) \times (B_1 B_2 - B_1 B_3) \times (B_2 B_3 - B_2 B_3)$

and  $\phi_3(R) = [R - \pi_{U_1}(R_1 \times R_2 \times R_3 - (R_1 R_2 \times R_3) - (R_1 R_3 \times R_2))] -$   
 $- [R - \pi_{U_1}(R_1 \times R_2 - R_1 R_2)]. \square$

## 6. EMBEDDED IMPLICATIONAL DEPENDENCIES

An *embedded implicational dependency* (EID) [Fagin 1980] is a sentence of the form

$$(\forall x_1 \dots x_m)((A_1 \wedge A_2 \dots \wedge A_n) \rightarrow (y_1 \dots y_k)(B_1 \wedge \dots \wedge B_r)).$$

Each of the  $A_i$ 's and  $B_j$ 's is of the form either (a)  $z = w$  for some  $1 \leq j \leq r$  and  $z, w \in V_j$ , or (b)  $R(z_1, \dots, z_r)$  for some  $z_j \in V_j, j=1, \dots, r$ , where  $R$  is the only  $r$ -ary relation symbol and  $V_1, \dots, V_r$  are disjoint sets of variables.

Intuitively, an EID says that if certain tuples exist in the relation  $R$  then (a) certain pairs of domain elements must be identified and (b) some tuples must exist in  $R$ .

*Theorem 6.1* For every embedded implicational dependency there is an equivalent algebraic dependency, and vice-versa.

*Proof*

(1) Let  $\sigma$  be an embedded implicational dependency, and let  $C_1, \dots, C_r$  be the attributes of  $R$ . Let  $X$  be a set of attributes that contains  $|V_j|$  distinct copies of attribute  $C_j$  of  $R$ , one for each variable in  $V_j$ , and let  $Y$  be the subset of  $X$  which corresponds to variables that appear both in some  $A_i$  and some  $B_j$ . We shall construct two project-join expressions  $\phi, \psi$  on  $X$  with  $a(\phi) = a(\psi) = Y$  such that  $\sigma$  holds for a relation  $R$  if and only if  $\phi(\bar{R}) \subseteq \psi(\bar{R})$ . The expressions  $\phi$  and  $\psi$  are shallow of the form  $\phi = \pi_Y(\pi_{Z_1} \bowtie \dots \pi_{Z_n})$  and  $\psi = \pi_Y(\pi_{W_1} \bowtie \dots \pi_{W_r})$ . If  $A_i$  (resp.  $B_j$ ) is of the form  $z=w$  for  $z, w \in V_j$ , then  $Z_i$  (resp.  $W_j$ ) is  $C_j' C_j''$  where  $C_j', C_j''$  are the copies of  $C_j$  in  $X$  that correspond to  $z$  and  $w$ . If  $A_i$  (resp.  $B_j$ ) is of the form  $R(z_1, \dots, z_r)$ , then  $Z_i$  (resp.  $W_j$ ) is  $C_1' C_2' \dots C_r'$  where  $C_j'$  is the copy of  $C_j$  in  $X$  that corresponds to  $z_j$  for  $j=1, \dots, r$ .

Let  $t$  be an  $X$ -tuple. Its projection  $t_Y$  is in  $\phi(\bar{R})$  iff  $t_{Z_i}$  is in  $\pi_{Z_i}(\bar{R})$ . If  $Z_i = C_j' C_j''$  then we must have  $t_{C_j'} = t_{C_j''}$ . Thus, a  $Y$ -tuple  $u$  is in  $\phi(\bar{R})$  iff there is an assignment of values to the rest of the variables in the  $A_i$ 's so that the left-hand side of  $\sigma$  is satisfied by  $u$  and this assignment. Similarly with  $\psi$ . Therefore,  $\phi(\bar{R}) \subseteq \psi(\bar{R})$  if and only if  $\sigma$  holds in  $R$ .

(2) Let  $\phi(\bar{R}) \subseteq \psi(\bar{R})$  be an algebraic dependency. Let  $T_\phi, T_\psi$  be the tableaux of  $\phi$  and  $\psi$  and let us assume without loss of generality that the only common symbols are the distinguished ones. For each symbol of  $T_\phi$  we have a variable  $x_i$ ; and for each symbol of  $T_\psi$  that does not appear in  $T_\phi$  a variable  $y_j$ . The left-hand side of an EID  $\sigma$  over the  $x_i$ 's and  $y_j$ 's is constructed as follows. For every row  $t$  of  $T_\phi$ ,  $\sigma$  has one  $A_i$  of the form  $R(z_1, \dots, z_r)$ , where the  $z_i$ 's correspond to the symbols of  $t$  in the first copies of  $U$ , and additional  $A$ 's of the form  $z=w$  that equate variables that correspond to symbols of  $t$  in different copies of the same attribute. The right-hand side of  $\sigma$  is constructed similarly from  $\psi$ . It is easy to see then that  $\phi(\bar{R}) \subseteq \psi(\bar{R})$  iff  $\sigma$  holds of  $R$ .  $\square$

Fagin defined an operation on relations over the same set of attributes as follows. Let  $R_1, R_2, \dots$  be such relations. The *direct product* of  $R_1, R_2, \dots$ , denoted as  $\otimes \langle R_1, R_2, \dots \rangle$ , is the relation

$$\{ \langle \langle a_1, a_2, \dots \rangle \langle b_1, b_2, \dots \rangle, \dots, \langle d_1, d_2, \dots \rangle \rangle : \\ (a_j, b_j, \dots, d_j) \in R_j \text{ for } j=1, 2, \dots \}.$$

The direct product is essentially the Cartesian product, compressed to the same number of attributes as the original relations.

It is easy to see that  $\otimes$  commutes with  $\pi, \bowtie, -$  (extension of a relation), and thus for all algebraic expressions  $\phi$  over extended relations

$$\phi(\otimes \langle \bar{R}_1, \bar{R}_2, \dots \rangle) = \otimes \langle \phi(\bar{R}_1), \phi(\bar{R}_2), \dots \rangle.$$

Furthermore,  $\otimes$  is componentwise monotonic when applied to nonempty relations. That is, if  $R_1, R_2, \dots, R_1', R_2', \dots$  are not empty then



$$\begin{aligned} \otimes\langle R_1, R_2, \dots \rangle \subseteq \otimes\langle R_1', R_2', \dots \rangle \text{ iff} \\ R_1 \subseteq R_1', R_2 \subseteq R_2', \dots \end{aligned}$$

A predicate  $P$  on relations is called *faithful* with respect to direct product ([Fagin 1980]) if  $P$  holds of  $\otimes\langle R_1, R_2, \dots \rangle$  if and only if it holds of each  $R_i$  (whenever all  $R_i$ 's are nonempty). The next lemma follows now from the discussion above.

*Lemma 6.1* [Fagin 1980] Algebraic dependencies are faithful with respect to direct product.  $\square$

Let  $\Sigma$  be a set of predicates (of some class  $C$ ) on relations. An *Armstrong relation* of  $\Sigma$  (wrt  $C$ ) is a relation  $R$  such that, for all  $\sigma \in C$ ,  $R$  satisfies  $\sigma$  iff  $\Sigma \models \sigma$ .

*Corollary* [Fagin 1980]. Any set  $\Sigma$  of algebraic dependencies has an Armstrong relation.

*Proof* Let  $\sigma_1, \sigma_2, \dots$  be all algebraic dependencies that are not implied by  $\Sigma$ . Let  $R_i$  be a counterexample to the implication  $\Sigma \models \sigma_i$ , and let  $R = \otimes\langle R_1, R_2, \dots \rangle$ . Since the empty relation satisfies all algebraic dependencies, the  $R_i$ 's are nonempty. Thus, it follows from Lemma 6.1 that  $R$  is an Armstrong relation for  $\Sigma$ .  $\square$

## REFERENCES

- [Aho et al. 1979] A. V. Aho, Y. Sagiv, J. D. Ullman, "Equivalence Among Relational Expressions" *SIAM J. Comp.* 8:2, pp. 218-246.
- [Assu 1978] A. V. Aho, Y. Sagiv, T. Szymanski, J. D. Ullman, "Inferring a Tree from Lowest Common Ancestors With an Application to the Optimization of Relational Expressions", *SIAM J. Comput.*, to appear.
- [Armstrong 1974] W. W. Armstrong, "Dependency Structures of Data Base Relationships" *Proc. IFIP 1974*, pp. 580-583.
- [Beeri et al. 1977] C. Beeri, R. Fagin, J. H. Howard, "A Complete Axiomatization for Functional and Multivalued Dependencies: *Proc. 3rd SIGMOD Conference*, pp. 47-61, 1977.
- [Codd 1970] E. F. Codd "A Relational Model for Large Shared Data Bases" *CACM* 12:6, pp. 377-387.
- [Codd 1972] E. F. Codd "Relational Completeness of Data Base Sublanguages" in *Data Base Systems* (R. Rustin, ed.) Prentice-Hall, pp. 65-98.
- [Fagin 1977] R. Fagin, Multivalued Dependencies and a New Normal Form for Relational Databases" *ACM TODS* 2:3, pp. 262-278.
- [Fagin 1980] R. Fagin, "Horn Clauses and Database Dependencies", *Proc. 12th Annual ACM Symp. Theory of Computing*, pp. 123-124.
- [Maier et al. 1979] D. Maier, A. O. Mendelzon, Y. Sagiv, "Testing Implications of Data Dependencies, *Proc. ACM-SIGMOD Conference*.
- [Nicolas 1978] J. M. Nicolas, "Mutual Dependencies and Some Results on Undecomposable Relations" ONERA-CERT, Toulouse, France.
- [Paradaens 1979] J. Paradaens, "Transitive Dependencies in a Database Scheme", R387 Philips, Bruxelles, Belgium.
- [Rissanen 1977] J. Rissanen, "Independent Components of Relations", *ACM TODS*, 2:4, pp. 317-325.
- [Sadri and Ullman 1980] F. Sadri and J. D. Ullman, "A complete Axiomatization for a Large Class of Dependencies", *Proc. 12th Annual ACM Symp. Theory of Computing*, pp. 117-122.
- [Sagiv and Walecka 1979] Y. Sagiv, S. Walecka "Subset Dependencies as an Alternative to Embedded Multivalued Dependencies", U1UCDCS-R-79-980, Univ. of Illinois at Urbana.
- [Sciore 1979] E. Sciore "A Complete Axiomatization of Full Join Dependencies", TR #279, EECS Dept., Princeton University.
- [Ullman 1979] J. D. Ullman, *Principles of Database Systems*, Computer Science Press.