CHARACTERIZING DATA BASE DEPENDENCIES

J.A. Makowsky[*]
Faculty of Computer Science
Technion - Israel Institute of Technology
Haifa, Israel

ABSTRACT: We characterize the class of Full Implicational Dependencies introduced by Fagin as the maximal class of dependencies which are safe, admit finite models in a strong sense and admit Armstrong Relations. We also give other more modeltheoretic characterizations of Full and Embedded Implicational Dependencies.

1. INTRODUCTION

Data base dependencies are first order formlae that can be specified to hold for a given relational data base. They form a basis for organizing semantic knowledge about data bases. They are also invaluable for structuring a data base so as to avoid various problems, or anomalies, having to do with inserting and deleting data (cf.[U]). In recent years several proposals have been made for reasonable sets of dependencies [ABU, BV, D F77, GJ, N, P, PP, PY, R78, SU, SW, TKY78 and U] and attempts have been made to unify all these notions. Two classes of particular interest emerged in the literature: The set of Embedded Implicational Dependencies (EID) as the largest class so far considered [F80, PY, CLM] including the Template Dependencies [US] and the Full Implicational Dependencies (FID) [F80].

In [CLM] some evidence is presented that the EID form too large a class: the inference problem for EID is undecidable and there are EIDs which admit only infinite relations; in contrast this is not so for the FIDs. The main result of this paper is a characterization of the FIDs in terms of properties which emerge from data base theory, i.e. safety, securability and admitting Armstrong relations. The idea behind such characterizations is to delimit the use of full first order logic in data base theory and to show that these limitations are inherent to the subject. Previously [CLM] model theoretic characterizations of both the EIDs and FIDs have been given, but we feel

---

that the characterizing properties had too much of an algebraic flavour and too little to do with data base theory.

The properties involved here are well established: Domain invariance or safety stems from the fact that we do not consider general models, but rather relations as such. So, if we want to use model theory, we have to justify that our dependencies only speak about the relation. The existence of Armstrong relations can be considered fundamental to data base theory, especially if they are finite. They allow the designer of the data base model to gain a thorough overview of what he already assumes to be true. The notion of securability is new to data base theory, though it seems implicite in many considerations: it is a property that not only guarantees the existence of finite relations satisfying a given finite set of dependencies, but it guarantees it continuously: the property holds or fails in a given set of data provided some finite configuration of data is present. The notion was introduced by Tharp in a different context [Th] and further studied in [Ma]. Now the FIDs satisfy all of these demands and the main theorem tells us that no extension of the FIDs will do so. Extending the FIDs may still be reasonable, but we have to pay a price: Undecidability of the satisfaction and consequence relations is likely by [CLM] and violation of one of the above demands sure.

The plan of the paper is as follows: In Section 2 we review and present new model theoretic characterizations of the EIDs and FIDs and present a new characterization of safe dependencies. Other such characterizations of safety may be found in [PY and Co]. In Section 3 we present the concept of securability and in Section 4 we discuss sets of dependencies having the Armstrong property. All this is collected in Section 5 to prove our main result. Detailed proofs will appear elsewhere.

2. SOME MODEL THEORY

Let R be a relation symbol of arity n. A model $M = (D,R)$ for R consists of a set D (the domain) together with a relation $R \subset D^n$. Throughout, we assume that first order formulae have just one non-logical symbol R and may contain equality. An atomic formula is a formula of the form $R(x_{i_1}, \ldots, x_{i_n})$ or $x_i = x_j$ where the

x_i's are variables. An (untyped or <u>general</u>) <u>embedded implicational dependency</u>, or
EID for short, is a sentence of the form

$$\forall x_1,\ldots,x_p((A_1 \wedge \ldots \wedge A_q) \rightarrow \exists y_1,\ldots,y_r(B_1 \wedge \ldots \wedge B_s))$$

where each A_i, B_j is an atomic formula. We will assume that each x_i appears in
at least one of the A_j's. W.l.o.g., it may be assumed that no A_j is an equality
and equaligy in the B_i's only uses x_i's. An EID is said to be <u>full</u> if it has no
existential quantifier. We abbreviate full implicational dependencies by FID.

An EID (FID) is said to be <u>typed</u> if no variable occurs in two different columns
(places) in <u>R</u>, and equality occurs only between variables appearing in the same
column. Models for types EIDs could be replaced by many-sorted models $M = (D_1,\ldots,D_n;R)$
with $R \subseteq D_1 \times D_2 \times \ldots \times D_n$. A model is trivial if $|D| = 1$ and $R = D^n$ (in the
typed case this amounts to each D_i being a singleton $\{d_i\}$ and $R = \{(d_1,d_2,\ldots,d_n)\}$).
Given a set of formulae Σ and a formula φ, we say $\Sigma = \varphi$ (φ is a consequence of Σ)
if φ holds in all (finite and infinite) models of Σ. We write $\Sigma \models_f \varphi$ if φ
holds in all finite models of Σ.

2.1 Domain Invariance (Safety, Permissibility)

A first order formula φ is said to be <u>Domain independent</u> [F80] provided φ holds
in a model (D,R) iff it holds in a model (D',R') with $R = R'$. Domain independence is
called <u>safety</u> in [U, p.113] and safe formulae are called permissible in [Co].
Note that "domain-independent" is used in [P4] for what we call "typed".

In [CLM] we defined a <u>dependency</u> to be a domain independent formula that holds in the
trivial structure.

Note that the FIDs and EIDs are indeed dependencies.

Let φ be a first order formula in prenex normal form with matrix in disjunctive

normal form $\bigvee\limits_i \bigwedge\limits_j \varphi_{ij}$.

We define a variable x to be constrained following [Co]:

(i)   A <u>direct constraint</u> on x is an atomic formula $R(\ldots x \ldots)$ in which x occurs.

(ii) An <u>indirect constraint</u> on x is a formula $x = y$ where y is directly
      constrained.

(iii)   If x is existentially quantified in φ then x is constrained iff in every

disjunct $\wedge\varphi_{ij}$ in which x occurs some φ_{ij} is a direct or indirect cons-

traint on x.

(iv)   If x is universally quantified in φ then x is constrained iff in some

disjunct $\wedge\varphi_{ij}$ in which x occurs, every φ_{ij} in which x occurs is the

negation of a direct or indirect constraint on x.

Now let $F_i(z)$ be the formula $\exists x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_n \underline{R}(x_1, \ldots x_{i-1}, z, x_{i+1} \ldots x_n)$ and

$F(z)$ be $v_i^n F_i(z)$. If φ is a first order formula we denote by φF the formula

which is obtained from φ by inductively replacing each bound variable occurrence

$\exists x$ by $\exists x(F(x) \wedge ..)$ and $\forall x$ by $\forall x(F(x) \rightarrow ...)$.

The following is an easy observation (cf. [Co], [Va]).

Proposition 1:   Let φ be a first order formula which holds in the trivial model.

Then the following are equivalent:

(i) φ is a dependency;

(ii) φ is equivalent to φF;

(iii) φ is equivalent to a formula φ_o in which each bound variable is constrained.

This can be read as a characterization of safe formulae in syntactic terms. Since

φ may contain arbitrary complex "superfluous" parts, we cannot replace "φ is

equivalent to..." by   "φ is of the form...". Proposition 1 states precisely and

improves the observation in [Co]. We predict that this will be a "folk theorem" in

the sense of [Ha].

Proposition 1 is a special case of an easy theorem in model theory characterizing

sentences preserved under relativizations. Its origin goes back to the 1950 period.

A more refined version involving theories was proved using ultraproducts by Keisler

[Ke] and [CK, problem 5.2.20]. Note that Vardi observed [Va] that dependencies are

recursive enumerable but not recursive. See also section 6.

2.2 Model Theoretic Characterizations of FID and EID

In [CLM] several characterization theorems for FID and EID were exhibited. We restate

these results here and add some new ones. Undefined model theoretic and algebraic

concepts may be found in the standard text book [CK]. We denote by ΠM_i and $\lim_{i \in I} M_i$
the product and direct limit (for homomorphisms) of a family M_i $i \in I$ of structures.
We say that a dependency (formula) φ is:

(i)   a <u>substructure dependency (formula)</u> if $M \models \varphi$ and $M_o \subset M$ is a substructure
      then $M_o \models \varphi$.

(ii) a <u>product dependency (formula)</u> if $M_i \models \varphi$ for $i \in I$ then $\prod_{i \in I} M_i \models \varphi$.

(iii) a <u>faithful dependency (formula)</u> if $M_i \models \varphi$ for $i \in I$ iff $\prod_{i \in I} M_i \models \varphi$.

(iv) a <u>limit dependency (formula)</u> if $M_i \models \varphi$ for $i \in I$ then $\lim_{i \in I} M_i \models \varphi$.

Here $\lim_{i \in I} M_i$ denotes the direct limit of the directed system M_i, $i \in I$, with respect
to some partial order on I such that for any two structures M_i, M_j, i and j in I,
there is a k above both i,j in I such that M_i, M_j are mapped in M_k .

If $N = \prod_{i \in I} M_i$ we call the M_i's the factors of N. We call a **dependency** (formula) a

(v)   a <u>factor-dependency (formula)</u> if $\prod_{i \in I} M_i \models \varphi$ then $M_i \models \varphi$ for each $i \in I$.

(vi) a <u>chain-dependency (formula)</u> if M_i, $i \in I$ is a substructure chain and **for**
each $i \in I$ $M_i \models \varphi$ then $\bigcup_{i \in I} M_i \models \varphi$.

Now we can extend Fagin's observation on preservation properties of various
dependencies [F80] :

<u>Proposition 2</u>: (i) The EIDs are product-, **chain-** and limit dependencies.

(ii) The typed EIDs are in addition also factor dependencies ,hence faithful.

(iii) The FIDs are also substructure dependencies.

<u>Proof:</u> Consult [CK,exercices to chapter 5] .

The model theorist's interest, having its sources in research done in the 1950s,
is in converse theorems, giving rise to syntactical characterizations of various
semantically defined classes of formulas.E.g. the substructure formulas are exactly
the formulas logically equivalent to universal formulas (i,e. formulas in prenex form
with only universal quantifiers). The best source for this in general is[CK,chapter 5[

With a bit more work we even get a converse:

Theorem 3: Let φ be a first order formula true in the trivial structure. Then the following are equivalent:

(i) φ is logically to a FID.

(ii) φ is both a product and a substructure dependency.

Outline of Proof: (i) ⇒ (ii) follows since FIDs are universal Horn-sentences, hence by [CK] both factor and substructure dependencies. To prove (ii) ⇒ (i) we note first that φ is equivalent to some universal Horn-sentence ψ by [CK] and by proposition 1 ψ is equivalent to ψF. Now, it is easy to see that if ψ is a universal Horn-sentence, so ψF is equivalent to an FID. In a way this is again "folk-theorem" in the sense of [Ha].

Theorem 4: Let φ be a first order formula true in the trivial structure. Then the following are equivalent:

(i) φ is logically equivalent to an EID.

(ii) φ is both a product and limit dependency.

Outline of Proof: Note first [CK, exercise 5.2.24] that every limit formula is of the form $\bigwedge_i (\forall \bar{x}(\psi_i \rightarrow \exists \bar{y} \theta_i))$ with ψ_i, θ_i quantifier free and positive. Now we use the product property to eliminate disjunctions in θ_i and propositional logic to eliminate them in ψ_i. Then we apply proposition 1.

2.3 Critique

We were rather casual about the proofs of Theorems 3 and 4 because we think that they miss the point. Fagin proved that the (typed) EIDs are faithful dependencies in order to show that they always admit Armstrong relations. We shall show that no such converse theorem can hold in a later section. But our main point here is that algebraic operations alone is not what is needed in data base theory, faithfulness is not a priori interesting and in fact products should be avoided in data base constructions, even for the construction of Armstrong relations.

What we would like are analogues to Theorems 3 and 4 where the algebraic properties are replaced by more practice oriented properties stemming from experience with data

base dependencies. What we have singled out as desirable properties will be discussed in detail in the next two sections.

3. SECURABILITY

The following concept has been studied in a different context in [Th, M]. A first order formula φ is called n-securable, for n a natural number, if for every structure M (for φ) there is substructure M_o of cardinality less or equal to n such that for every structure M_1 with $M_o \subset M_1 \subset M$ $M_1 \models \varphi$ iff $M \models \varphi$. A formula is securable if we replace less than n by "finite". Let Sec be the set of securable formulae.

The following is immediate from the definition:

Lemma 5: Sec is closed under Boolean operations.

Lemma 6: (i) $\models$ and $\models_f$ coincide on Sec.

(ii) The consequence relation is decidable on Sec (provided we know which formulae are in Sec.

This leaves us with the problem of identifying a securable formula.

Lemma 7: (i) The FIDs are in Sec.

(ii) There is an EID which is not in Sec.

Proof: (i) If an FID is true in M then it is secured by the trivial substructure, otherwise its negation is secured by finitely many witnesses. (ii) By a result in [CLM] the consequence relation on EID is undecidable, contradicting Lemma 6, if we assume EID is in Sec.

In [Th and M] the following is proved:

Theorem 8: For a first order formula the following are equivalent:

(i) φ is in Sec.

(ii) φ is n-securable for some n.

(iii) φ is equivalent to a Boolean combination of universal sentences.

(iv) φ and $\neg\varphi$ are both chain sentences.

(iii) gives us a syntactic characterization of the securable sentences. Securability

is stronger than just the finite model property: it is in a sense a continuous finite model property. We claim that securability captures the finite model property inherent in data base theory.

By a classical theorem of model theory, due to Galvin [Ga], [CK], every universal sentence is a Boolean combination of universal Horn-sentences. If we combine this with proposition 1 we get easily

Theorem 9: A dependency φ is securable iff it is logically equivalent to a Boolean combination of FIDs.

Outline of Proof: By Theorem 8 φ is equivalent to a Boolean combination of universal sentences, hence by Galvin's theorem to a Boolean combination of universal Horn-sentences. Now a sentence of the form $\forall \bar{x}(\bigwedge_i A_i \rightarrow \underline{False})$ is equivalent to $\underline{False}$ by our assumption on the existence of trivial substructures, hence the result.

4. ARMSTRONG RELATIONS

Let S be a set of first order formulae, We say that a structure M is an Armstrong-relation for Σ a subset of S, if $M \models \Sigma$ but for each $\alpha \in S$ with α not a consequence of Σ, M does not satisfy α. We say that S is Armstrong if every subset of S, which has a model, has an Armstrong-relation. The aim of this section is to study sets of formulae or dependencies which are Armstrong. This approach seems to be new, though it is implicite in [F80].

The following are immediate from the definitions:

Lemma 10: (i) If S is Armstrong and S_0 is a subset of S, then S_0 is Armstrong.

(ii) If S is Armstrong and S_1 is such that every α in S_1 is equi-
valent to a finite (or infinite) conjunction of elements of S, so
S_1 is Armstrong.

(iii) If S is a set of pairwise inconsistent formulae so S is Armstrong.
Note that there are very complex S which are Armstrong, due to (iii). This substantiates our claim that Armstrong alone does not characterize any S.
In [F80] Fagin proved the following:

Theorem 11: For a set of first order formulae S the following are equivalent:

(i) S is Armstrong,

(ii)   Whenever S_o, S_1 are subsets of S and $\wedge S_o \models \vee S_1$ then there is α in S_1 such that $\wedge S_o \models \alpha$.

(iii)   There is an operation $\square$ that maps indexed families M_i of models into models $\square M_i$ such that for every α in S $\square M_i \models \alpha$ iff for every i $M_i \models \alpha$. (We say that S is $\square$-faithful.)

The existence of Armstrong relations seems to be quite fundamental to Data base theory, so it is natural to ask for the class of dependencies to be Armstrong. Fagin proved

Theorem 11 for this purpose for it follows that the typed EIDs are Armstrong since they are faithful by proposition 2. But lemma 10 (iii) shows very clearly

that there is no hope to improve Theorem 11 so that Armstrong implies faithfulness. We end this section with a crucial lemma:

Lemma 12: Let S be a set of first order formulae and S_o be a subset. Let us assume that S is Armstrong and every α in S is equivalent to a Boolean combination of formulae of S_o. Then α is equivalent to a finite conjunction of formulae in S_o.

Proof: We first observe that two structures M_o, M_1 satisfy the same sentences of S iff they satisfy the same sentences from S_o. Now suppose α is in S and is neither valid nor inconsistent. We claim that there is an α_o in S_o which is not valid and $\alpha \models \alpha_o$.

By Theorem 11 (ii) it is enough to show that:

$\alpha \models \vee \{\beta \in S_o | \beta \text{ not valid}\}$. Call this set S_1. Since S is Armstrong, if this is not the case, there are Armstrong relations M_o, M_1 where all the members of S_1 are false but $M_o \models \alpha$, $M_1 \models -\alpha$, which contradicts the fact that M_o and M_1 satisfy the same sentences from S_o. Now let Σ be $\{\beta \in S_o | \alpha \models \beta\}$. Again using that S is Armstrong we conclude that α is equivalent to the conjunction Σ. But now using compactness we conclude that there is a finite subset Σ_o of Σ such that α is equivalent to the conjunction of Σ_o.   Q.E.D.

5. THE CHARACTERIZATION THEOREM FOR FIDs

Note that by our definition the FIDs are closed under finite conjunctions. So our main theorem reads:

Theorem 13: Let S be a set of dependencies such that:

(i) Each FID is in S

(ii) Each dependency in S is securable

(iii) S is Armstrong.

Then S consists, up to logical equivalence, exactly of the FIDs.

Proof: This is now a direct consequence of Theorem 9 and Lemma 12.

Note that the EIDs satisfy (i) and (iii) and the class of Boolean combinations of FIDs satisfies (i) and (ii); so Theorem 13 is best possible.

If we define that a dependency φ is negatively securable if for every M which does not satisfy φ there is a finite M_o such that for every M_1 $M_o \subset M_1 \subset M$ M_1 does not satisfy φ; then we can have an analogue theorem for EIDs.

Theorem 14: Let S be a set of dependencies such that:

(i) each EID is in S

(ii) each dependency in S is negatively securable

(iii) S is Armstrong

(iv) each model a subset of S has a trivial submodel.

Then S consists, up to logical equivalence, exactly of the EIDs.

The significance of Theorem 13 is not in possible applications but in its limitative character: It seems to explain why the FIDs are a good class of dependencies and why it cannot be expanded without giving up some of its nice properties. We are fully aware that securability is a rather strong property. It would be interesting to know if we could replace it by a weaker hypothesis such as (a) the consequence relation on S is decidable or (b) each finite subset of S which has a model has a finite model? But it is obvious that then we have to assume stronger closure properties on S, such as closure under subformulas. Since these properties will, by Theorem 13, imply securability, Theorem 13 is probably the neatest such characterization.

Theorem 14 shows how we can relax the securability condition.

6. UNDECIDABILITY OF SEMANTIC PROPERTIES

We often are in a situation that we have a set Σ of first order formulas which is recursively enumerable and such that every formula having some property P is equivalent to a formula in Σ .This is the case in theorem 8,9,13,14. In many cases(here theorem 8,9,13) is such that the set of valid sentences in Σ , call it Val(Σ), is recursive. In our context this is mostly so because of the finite model property of the sets in question. It is now natural to ask, whether to have property P is decidable for first order formulas. Vardi observed the following nice theorem:

Theorem 15: (Vardi [Va 2]) Let Σ be a set of first order formulas such that Σ is recursively enumerable and Val(Σ) is not empty and recursive. Then the set Eq(Σ) = { θ:there is a $\sigma \in \Sigma$ such that $\models \sigma \leftrightarrow \theta$} is not recursive.

Proof: We show that,under the above hypotheses, Eq(Σ) recursive implies that Val(Σ) is not recursive. Let φ be any sentence. If φ is not in Eq(Σ) φ is not valid, since Val(Σ) is not empty. But if φ is in Eq(Σ) we can find θ in Σsuch that φ andθ are logically equivalent since Eq(Σ) is recursive by our assumption. But then φ is valid iff θ is in Val(Σ).Now, if we assume Val(Σ) recursive, this would give us a decision-procedure for all valid formulas, which contradicts Godel's theorem.

The following corollary is immediate from theorem 8,9,13:

Corollary 16:   The following are not recursive:

(i) Eq(FID)

(ii)   Sec

(iii) For any recursive subset $\Sigma_0$of FIDs or Sec the set Eq(Σ_0).

7. ACKNOWLEDGEMENT

[ABU]   Aho, A.V., Beeri, C. and Ullman, J.D.: The theory of joins in relational databases, 19th FOCS (1977), pp. 107-113.

[BV]   Beeri, C. and Vardi, M.Y.: On the properties of total join dependencies, Proc. XP1, Workshop on Relational Data Base Theory, Stoney Brook, N.Y. (June 1980).

[Co]   Cooper, E.C.: On the expressive power of query languages for relational databases, TR-14-80, Center for Research in Computing Technology, Harvard University (1980).

[CLM]   Chandra, A.K., Lewis, H.R. and Makowsky, J.A.: Embedded implicational dependencies and their inference problem, Proc. XP1, Workshop on Relational Data Base Theory, Stoney Brook, N.Y. (June 1980), SIGACT 1981.

[CK]   Chang, C.C. and Keisler, H.J.: Model Theory, Studies in Logic and Foundations of Math., Vol. 73 (1973).

[D]   Delobel, C.: Normalization and hierarchical dependencies in the relational data model, ACM Trans. on Database Systems 3, 3 (1978).

[F77]   Fagin, R.:   Multivalued dependencies and a new normal form for relational data bases, ACM Trans. on Database Systems 2, 3 (1977).

[F80]   Fagin, R.:   Horn clauses and database dependencies, ACM Symp. on Theory of Comp., Los Angeles, CA (April 1980), pp. 123-134.

[Ga]   Galvin, F.: Horn sentences, Ann. Math. Logic 1, (1970), pp. 389-4-22.

[GJ]   Grant, J. and Jacobs, B.E.: On generalized dependencies, to appear.

[Ha]   Harel, D.: on Folk theorems, Comm. ACM, July 1980.

[Ke]   Keisler, H.J.: Some applications of infinitely long formulas, J. Symbolic Logic 30, (1965), pp. 339-349.

[M]   Makowky, J.A.:   Securable quantifiers, k-unions and admissible sets, Logic Colloquium '73 (1975), pp. 409-428.

[N]   Nicolas, J.M.: First order logic formalization for functional, multivalued and mutual dependencies, ACM-SIGMOD Conf. (1978), pp. 40-46.

[P]   Paredaens, J.:   Transitive dependencies in a data base scheme, R 387 MBLE, Brussels (1979).

[PP]   Parker, D.S., and Parsaye-Ghomi, K.: Inference involving EMVDs and transitive dependencies, TR UCLA.

[PY]   Papadimitriou, C. and Yannakakis, M.: Algebraic dependencies, 21st FOCS, Syracuse, N.Y. (October 1980), pp. 328-332.

[R78]   Rissanen, J.: Theory of relations for databases - a tutorial survey 7th Symp. MFCS, Lecture Notes on Comp. Sc. 64 (1978), pp. 536-551.

[SU]   Sadri, P. and Ullman, J.D.: A complete axiomatization for a large class of dependencies in relational databases, 12th SIGACT (April 1980).

[SW]   Sagiv, Y. and Walecka, S.: Subset dependencies as an alternative to embedded multivalued dependencies, TR U1UCDCS-R-79-980, Dept. CS, University of Illinois (1979).

[TKY]   Tanaka, K.Y., Kambayashi, Y. and Yajima, S.: Properties of embedded multi-valued dependencies in a relational database, TR ER78-03, Dept. of Inf. Sc., University of Kyoto (1978).

[U]   Ullman, J.D., Principles of Database Systems, Computer Science Press (1980).

[Th]   Tharp, L.: Continuity and elementary logic, JSL 39 (1974), pp. 700-716.

[Va]   Vardi, M.: The decision problem of database dependencies, preprint June 1980, Weizmann Institute, Israel.

[Va1]   Vardi,M.   This volume

[Va2]   Vardi,M.   Oral communication