The Rise and Fall and Rise of of Dependency Theory

Part I: Rise and Fall

Moshe Y. Vardi

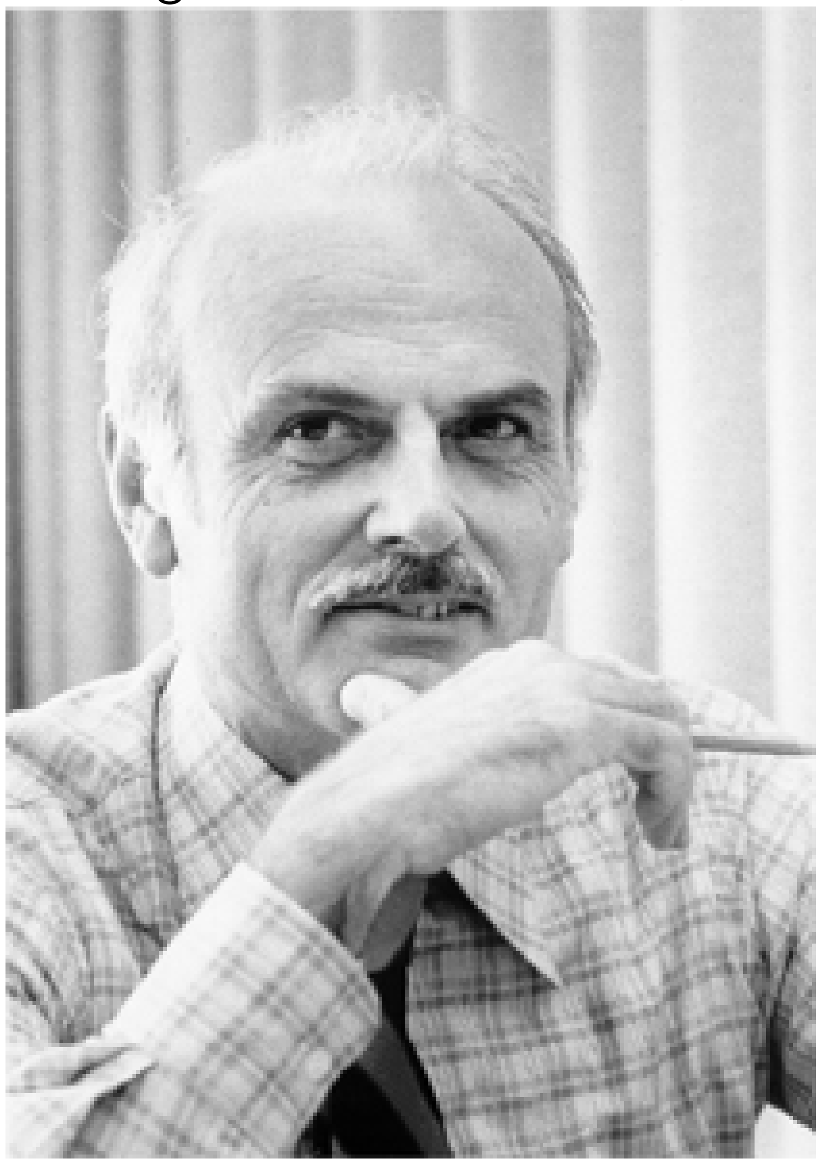
Rice University

Primary Keys

E.F. Codd, *A Relational Model of Data for Large Shared Data Banks*, CACM, June 1970

> "Future users of large data banks must be protected from having to know how the data is organized in the machine (the internal representation). ... In Section 1, a model based on n-ary relations, a normal form for data base relations, and the concept of a universal data sublanguage are introduced. ... Normally, one domain (or combination of domains) of a given relation has values which uniquely identify each element (n-tuple) of that relation. Such a domain (or combination) is called *primary key*."

Figure 1: Edgar Frank “Ted” Codd, 1923–2003



Functional Dependencies

E.F. Codd, *Further Normalization of The Data Base Relational Model, IBM Research Report, 1971*

Definition: A relation r on a scheme R satisfies the *functional dependency* $X \rightarrow Y$, where $X, Y \subseteq R$, if, for all $t_1, t_2 \in r$, we have that $t_1[X] = t_2[X]$ implies $t_1[Y] = t_2[Y]$;

Motivation: *normalization*

- *Scheme*: Emp, Dept, Mgr
- FDs: Emp $\longrightarrow$ Dept, Dept $\longrightarrow$ Mgr
- Violation of *3rd Normal Form*
- Decompose to Emp, Dept and Dept, Mgr.

Theory of Functional Dependencies

The Emergence of Database Theory:

- Codd, 1971: FDs and lossless joins
- Delobel and Casey, 1973: properties of FDs
- Armstrong, 1974: axiomatization of FDs
- Bernstein, 1976: synthesis of 3NF database scheme from FDs
- Fagin, 1976: FDs and propositional logic
- Beeri and Bernstein, 1979: linear-time algorithm for implication of FDs

Dependency Implication

Definition: Let R be a relation scheme. A set Σ of dependencies implies a dependency σ, if, for every relation r on R, if r satisfies all dependencies in Σ, then it also satisfies σ.

The Implication Problem: Given set Σ of dependencies and a dependency σ, does Σ implies σ?

Multivalued Dependencies

C. Delobel, *Contributions Theoretiques a la Conception d'un Systeme d'Informatiions*, PhD. Dissertation, U. Grenoble, 1973

C. Zaniolo, *Analysis and Design of Relational Schemata for Database Systems*, PhD. Dissertation, UCLA, 1976

R. Fagin, *Multivalued Dependencies and a New Normal Form for Relational Databases*, ACM TODS, 1977

Definition: Let R be a relation scheme, $X, Y \subseteq R$, and $Z = R - (X \cup Y)$. A relation r on R satisfies the *multivalued dependency* (MVD) $X \twoheadrightarrow Y$ if, for all $t_1, t_2 \in r$, we have that $t_1[X] = t_2[X]$ implies that there is $t \in R$ with $t[X] = t_1[X] = t_2[X]$, $t[Y] = t_1[Y]$, and $t[Z] = t_2[Z]$.

MVDs and Lossless Decomposition

Let R be a relation scheme, $X, Y \subseteq R$, and $Z = R - (X \cup Y)$. A relation r on R satisfies $X \twoheadrightarrow Y$ iff $r = \pi_{XY}(r) \bowtie \pi_{XZ}(r)$

Motivation: *normalization*

- *Scheme*: Emp, Project, Hobby
- MVD: Emp $\twoheadrightarrow$ Project
- Violation of *4th Normal Form* (Fagin)
- Decompose to Emp, Project and Emp, Hobby.

Beeri, Fagin &Howard, SIGMOD'77: axiomatization and polytime implication problem for FDs and MVDs.

Example: $X \rightarrow Y$ implies $X \twoheadrightarrow Y$

Join Dependencies

J. Rissannen, *Independent Components of Relations*, ACM TODS, 1977

Definition: Let R be a relation scheme, and $R_1, \ldots, R_k$ such that $R = R_1 \cup \ldots \cup R_k$. A relation r on R satisfies the *join dependency* $\star[R_1, \ldots, R_k]$ if $r = \pi_{R_1}(r) \bowtie \ldots \bowtie \pi_{R_k}(r)$.

Motivation: *normalization*

- 5th Normal Form [Fagin, SIGMOD'79]

The Chase

D. Maier, A.O. Mendelzon, Y. Sagiv: *Testing Implication of Data Dependencies*, ACM TODS, 1979

To check if Σ does *not* implies σ, search for a candidate counterexample:

- Propose a relation r that violates σ.
- "Massage" r using the dependencies in Σ to ensure that all dependencies in Σ are satisfied.
- Check that final r still violates σ.

Computational Complexity:

- In EXPTIME
- NP-hard [Beeri&Vardi, 1980]

Figure 2: Alberto O. Mendelzon, 1951–2005

Embedded Dependencies

Fagin, 1977, Delobel, 1978, Rissanen, 1977

Definition: Let R be a relation scheme, and $X, Y, Z \subseteq R$. A relation r on R satisfies the *embedded multivalued dependency* (EMVD) $X \twoheadrightarrow Y|Z$ if $\pi_{XYZ}(r) = \pi_{XY}(r) \bowtie \pi_{XZ}(r)$

Definition: Let R be a relation scheme, and $R_1, \ldots, R_k$ such that $R_1 \cup \ldots \cup R_k = R' \subseteq R$. A relation r on R satisfies the *embedded join dependency* (EJD) $\star[R_1, \ldots, R_k]$ if $\pi_{R'}(r) = \pi_{R_1}(r) \bowtie \ldots \bowtie \pi_{R_k}(r)$.

Major Difficulty: Chasing with embedded dependencies does not terminate.

Nov. 1978: Mr. Vardi Goes to Graduate School

- Vardi, June 1979: "Anything left to do in dependency theory?"
- Beeri: "Why don't you work on the implication problem for embedded dependencies?"
- C. Beeri and M.Y. Vardi, *On the Properties of Join Dependencies*, Workshop on Logic and Databases, December 1979
- Vardi, February 1980: "I learned today about undecidability. Perhaps the implication problem for EMVDs is undecidable."
- Beeri: "We are not doing computability theory. This is database theory. Everything ought to be decidable!"
- M.Y.Vardi, *Axiomatization of Functional and Join Dependencies in The Relational Model*, MS Thesis, April 1980

Dependencies and FOL

J.M. Nicolas, SIGMOD'78: Dependencies can be expressed in *first-order logic*.

Scheme: A,B,C,D

- FD: A $\rightarrow$ B
  $(\forall x, x_2, x_3, x_4, y_2, y_3, y_4)((r(x, x_2, x_3, x_4) \wedge r(x, y_2, y_3, y_4)) \longrightarrow x_2 = y_2)$
- MVD: A $\twoheadrightarrow$B
  $(\forall x, x_2, x_3, x_4, y_2, y_3, y_4)((r(x, x_2, x_3, x_4) \wedge r(x, y_2, y_3, y_4)) \longrightarrow r(x, x_2, y_3, y_4)$
- EMVD: A $\twoheadrightarrow$B—C
  $(\forall x, x_2, x_3, x_4, y_2, y_3, y_4)((r(x, x_2, x_3, x_4) \wedge r(x, y_2, y_3, y_4)) \longrightarrow (\exists z) r(x, x_2, y_3, z))$

Generalizing Dependencies

Fagin, STOC'80: *Embedded Implicational Dependencies*:
$(\forall x_1, x_2, \ldots)((A_1 \wedge A_2 \ldots) \rightarrow (\exists z_1, z_2, \ldots)((B_1 \wedge B_2 \ldots))$

- A_i's: relational formulas
- B_i's: atomic formulas (relational or equality)
- x_i's: *guarded* – occur on right only if they occur on left

Beeri&Vardi, 1980:

- *Tuple-generating dependencies*: B_i's – relational formulas
- *Equality-generating dependencies*: B_i's – equality formulas

Full and Embedded Dependencies

Full dependencies: no existential variables
Embedded dependencies: existential variables

Beeri&Vardi, 1980: chase for tgds and egds

- Generally, terminates only for full dependencies.
- Termination for embedded dependencies in special cases.

A Query-Based View of Dependencies

Conjunctive Queries: $(\exists z_1, z_2, \ldots)(B_1 \wedge B_2 \ldots)$ [Chandra&Merlin, STOC'76]

- Incredibly rich theory!

Definition: A database D satisfies a conjunctive-query-containment dependency (CQCD) $Q_1 \subseteq Q_2$, for conjunctive queries Q_1 and Q_2, if $Q_1(D) \subseteq Q_2(D)$.

Yannakakis&Papadimitriou, FOCS'80: EIDs are equivalent to CQCDs.

The Implication Problem

[Beeri&Vardi, ICALP'81, Chandra, Lewis &Makowsky, STOC'81]:

- *Implication of Embedded Dependencies*: undecidable
- Implication of Full Dependencies: EXPTIME-complete

Significance:

- First undecidable "database-theoretic" problem.
- First intractable "database-theoretic" problem.

But: Have we generalized too much?

Template Dependencies

Sadri&Ullman, STOC'80:
$(\forall x_1, x_2, \ldots)((A_1 \wedge A_2 \ldots) \rightarrow (\exists z_1, z_2, \ldots)B)$

- A_i's: relational formulas
- B: relational formulas
- x_i's: guarded
- *typedness*: a variable cannot occur in two distinct columns

Undecidability of implication:

- Gurevich&Lewis, PODS'82: Bounded number of atomic formulas
- Vardi, PODS'82: Bounded arity

"Practical" Dependencies

Definition: An *inclusion dependency* is a CQCD with respect to simple projective queries π_{X_1} and π_{X_2}. [Casanova, Fagin&Papadimitrou, PODS'82]

- **Significance**: Captures *referential integrity*

Implication:

- Implication of IDs: PSPACE-complete [Casanova, Fagin&Papadimitrou, PODS'82]
- Implication of IDs and FDs: undecidable [Chandra&Vardi, SICOMP'85, Mitchell, I&C'83]

Rise and fall

Brief History:

- Rise: pre-PODS (1971-1981)
- Heyday: early PODSes (e.g, 8 papers in PODS'82)
- Decline and disdain: late 1980s (e.g., 0 papers in PODS'88)

Speculative Explanation:

- Hard to build useful theory for undecidable problems
- Database design proved to be a modeling activity (ER, Chen, 1976)
- Attention shifted to query processing (bread and butter of databases)
- Lack of interest in integrity constraints by SIGMOD community (CHECK constraint added to SQL only in 1989)

But:

- Subsequent rise: Fagin
- Impact on Datalog research: Ullman

What about EMVDs?

Christian Herrmann, *On the Undecidability of Implications between Embedded Multivalued Database Dependencies*, I&C, 1995