A NOTE ON LOSSLESS DATABASE DECOMPOSITIONS

Moshe Y. Vardi
IBM Research Laboratory
San Jose, CA 95193

Abstract

It is known that under a wide variety of assumptions a database decomposition is lossless if and only if the database scheme has a lossless join. Biskup, Dayal, and Bernstein have shown that when the given dependencies are functional then the database scheme has a lossless join if and only if one of the relation scheme is a key for the universal scheme. In this note we supply an alternative proof of that characterization: The proof uses tools from the theory of embedded join dependencies and the theory of tuple and equality generating dependencies, but is, nevertheless, much simpler than the previously published proof.

Keywords: database, decomposition, dependency, lossless join, relational model.

† The research reported here was done while this author was at Stanford University and supported by a Weizmann Post-Doctoral Fellowship and AFOSR grant 80-0212.
1. Introduction

A significant portion of research on relational database theory has been concerned with the properties of decompositions. The generic problem can be described as follows. Given a "universal" relation scheme presented as a set of attributes and a set of dependencies, what are the conditions under which it can be decomposed into a collection of relation schemes, each with its own sets of attributes and dependencies, having some desired properties. The properties considered were, at first, various normal forms. (see [Ma, Ul]).

A basic assumption underlying these ideas is that when a universal scheme is decomposed into smaller schemes, each of the universal relations associated with it is decomposed into smaller relations using the projection operation, i.e., each such relation is projected onto each one of the smaller schemes. For a decomposition to be useful, it should be lossless. In other words, it should be possible to reconstruct the universal relations from their projections. The desirability of this property is called in [BBG] the representation principle.

The property of losslessness has been studied in numerous papers [BR,MMSU,MUV,Va]. It has been shown there that, under a wide variety of assumptions, a decomposition is lossless if and only if the database scheme has a lossless join. When the given dependencies are functional the following characterization of database schemes that have lossless join is given in [BDB]: a database scheme has a lossless join if and only if one of the relation schemes is a key of the universal scheme.

The proof of the above characterization in [BDB] is quite involved and consists of a detailed study of the test for losslessness in [ABU]. In this note we describe an alternative proof for that characterization. The proof uses more advanced tools from dependency theory, but is, we believe, much simpler. Specifically, we use the theory of embedded join dependencies [BV1] and the theory of tuple and equality generating dependencies [BV2,BV3]. Our intention is to demonstrate how "higher-level" notions from dependency theory can be used to study its most fundamental questions. We believe that techniques used here could also be applicable to
other problems.

2. The Theorem

We assume familiarity with the terminology and concepts of relational database theory as presented in [Ma, UI]. The universe $U$ is a finite set of attributes. All attribute sets are subsets of $U$. An attribute set collection (asc) is a set $R = \{ R_1, \ldots, R_k \}$ of distinct attribute sets. We denote attribute sets by lightface letters and asc’s by boldface letters. A database scheme is an asc $R = \{ R_1, \ldots, R_k \}$ such that $\bigcup_{i=1}^{k} R_i = U$. A relation is a relation on $U$ unless explicitly specified otherwise. We use $I[X]$ to denote the projection of the relation $I$ on the attribute set $X$, and $^*I_j$ to denote the join of the set $\{ I_j \}$ of relations. We assume that the relations we are dealing with and, accordingly, the dependencies that refer to them are typed, that is, distinct attributes have disjoint domains. We also assume that all relations are finite.

A functional dependency (fd) is a statement $X \rightarrow Y$, where $X$ and $Y$ are attribute sets. It is satisfied by a relation $I$ if for all tuples $s$ and $t$ in $I$, if $s[X] = t[X]$ then $s[Y] = t[Y]$. A total join dependency (jd) is a statement $^*[R]$, where $R$ is a database scheme $\{ R_1, \ldots, R_k \}$. It is satisfied by a relation $I$ if $I = \bigcup_{i=1}^{k} I[R_i]$.

Let $\Sigma$ be a set of dependencies and let $\tau$ be a dependency. We say that $\Sigma$ implies $\tau$, denoted $\Sigma \models \tau$ if every relation on $U$ that satisfies all dependencies in $\Sigma$ satisfies also $\tau$. A database scheme $R$ has a lossless join with respect to $\Sigma$ if $\Sigma \models ^*[R]$. A set $\Sigma$ of fd’s is embeddable in a database scheme $R$ if for each fd $X \rightarrow Y$ in $\Sigma$ there is some $R$ in $R$ such that $XY \subseteq R$.

Theorem. [BDB] Let $\Sigma$ be a set of fd’s and let $R$ be a database scheme.

(1) If $R$ has a lossless join with respect to $\Sigma$, then there is some $R \in R$ such that $\Sigma \models R \rightarrow U$.

\footnote{Following convention we denote set union by juxtaposition.}
(2) If $\Sigma$ is embeddable in $R$ and there is some $R \in R$ such that $\Sigma \trianglerighteq R \rightarrow U$, then $R$ has a lossless join with respect to $\Sigma$. 

3. The Necessary Condition

3.1. Preliminary Definitions

We prove the necessary condition via an excursion through the theory of tuple and equality generating dependencies. We need first to review some definitions.

A valuation is a mapping on the domain of which tuples and relations are constructed. We can extend the definition of a valuation to tuples, in a component-wise manner, and to relations, in a tuple-wise manner.

Equality generating dependencies generalize fd's. An equality generating dependency (egd) says that if some tuples, fulfilling certain equalities, exist in the database, then some values in these tuples must be equal. Formally, an egd is a pair $\langle (a_1, a_2), I \rangle$, where $I$ is a relation and $a_1$ and $a_2$ occur in $I$. A relation $J$ satisfies $\langle (a_1, a_2), I \rangle$ if for any valuation $h$ such that $h(I) \subseteq J$ we have $h(a_1) = h(a_2)$. Note that if $a_1 = a_2$ then $\langle (a_1, a_2), I \rangle$ is trivially satisfied by every relation.

Lemma 1. [BV3] Let $h$ be a valuation and let $\langle (a_1, a_2), I \rangle$ be an egd. Then $\langle (a_1, a_2), I \rangle \trianglerighteq \langle (h(a_1), h(a_2)), h(I) \rangle$. 

Total tuple generating dependencies generalize tjd's. A total tuple generating dependency (ttgd) says that if some tuples, fulfilling certain equalities, exist in the database, then another tuple, whose values are taken from these tuples, must also exist in the database. Formally, a ttgd is a pair $\langle w, I \rangle$, where $I$ is a relation and $w$ is a tuple whose entries occur in $I$. A relation $J$ satisfies $\langle w, I \rangle$ if for any valuation $h$ such that $h(I) \subseteq J$, we have that $h(w) \subseteq J$. Note that if $w \subseteq I$, then $\langle w, I \rangle$ is trivially satisfied by every relation.

Fd's and tjd's can be viewed as a special case of egd's and ttgd's, respectively. Consider the fd $X \rightarrow A$, where $A \not\subseteq Y$. We define an egd $\tau_{X \rightarrow A} = \langle (a_0, a_1), J \rangle$ as follows: $J$ consists of
two tuples \( u \) and \( v \) such that \( u[X] = v[X], u[B] \neq v[B] \) for all \( B \in X, u[A] = a0, \) and \( v[A] = a1 \). We leave it to the reader to verify that a relation \( K \) satisfies \( X \rightarrow A \) if and only if it satisfies \( \tau_X \rightarrow A \). Consider the tgd \( \tau_R = \langle w, I \rangle \), where \( R = \{ R_1, \ldots, R_k \} \). We define a tgd \( \tau_R = \langle w, I \rangle \), where \( w \) is an arbitrary tuple, \( I = \{ w_{R_1}, \ldots, w_{R_k} \}, w_i[R_i] = w[R_i] \), and if \( A \in R_i \) then \( w_i[A] \) has a unique occurrence in \( I \). It is shown in [ASU] that a relation \( K \) satisfies \( \tau_R = \langle w, I \rangle \) if and only if it satisfies \( \tau_R \).

**Example 1.** Let \( U = ABCD \). Let \( I \) and \( J \) be the relations:

\[
I: \begin{array}{cccc}
A & B & C & D \\
\hline
a0 & b0 & c1 & d0 \\
a0 & b1 & c0 & d1
\end{array}
\quad J: \begin{array}{cccc}
A & B & C & D \\
\hline
a0 & b0 & c0 & d0 \\
a1 & b0 & c0 & d1
\end{array}
\]

Let \( u \) be the tuple:

\[
\begin{array}{cccc}
A & B & C & D \\
\hline
a0 & b0 & c0 & d0
\end{array}
\]

Let \( \tau_1 \) be the tgd \( \langle u, I \rangle \). Then \( \tau_1 \) is equivalent to the tjd \( \tau_1 = \langle ABD, AC \rangle \). Let \( \tau_2 \) be the egd \( \langle (a0, a1), J \rangle \). Then \( \tau_2 \) is equivalent to the fd \( BC \rightarrow A \).

An algorithm for testing implication of egd's and tgd's, called the *chase*, was presented in [BV2], following earlier algorithms for testing implications of fd's and tjd's [ABU, MMS]. We present here the special case where the implying dependencies are functional. Without loss of generality we assume that the given fd's have a single attribute on their right-hand side. A *chase* of a relation \( I \) by a set \( \Sigma \) of fd's is a maximal sequence of distinct relations \( I_0, I_1, \ldots \) such that \( I = I_0 \) and \( I_{n+1} \) is obtained from \( I_n \) by an application of a chase rule. To each fd in \( \Sigma \) there corresponds a chase rule:

**FD-rule** (for an fd \( X \rightarrow Y \) in \( \Sigma \)): \( I_{n+1} \) is obtained from \( I_n \) by identifying all occurrences of \( u[A] \) with all occurrences of \( v[A] \), for some tuples \( u \) and \( v \) in \( I_n \) such that \( u[X] = v[X] \).

To make the FD-rule unambiguous, we assume that the domain of values for each attribute is totally ordered and whenever two values are identified, the greater is identified with the smaller. Given \( \langle w, I \rangle \), we take \( w(A) \) as the smallest value in the domain for \( A \). (We can
always rename the values in \( w \) and \( I \) so that this is true). Similarly, given \( \langle a_1, a_2, I \rangle \), we take
\( a_1 \) as the smallest values in the domain for \( A \) and \( a_2 \) as the next smallest. Thus, the values in
\( w \) or \( a_1 \) do not change in the chase and \( a_2 \) can be identified only with \( a_1 \). It is shown in [BV2]
that all chases of \( I \) by \( \Sigma \) are finite and they have the same final relation, denoted \( \text{chase}_\Sigma(I) \).

To trace the tuples of \( I \) in the chase we adjoin to each tuple an ordinal number, that is
\( I = \{ w_1, \ldots, w_k \} \). The ordinal numbers do not change during the computation of the chase,
though the values in the tuples may change by the FD-rules. Thus, \( I_n \) consists of the tuples
\( w_1', \ldots, w_k' \) (not necessarily distinct), which correspond to the tuples \( w_1, \ldots, w_k \) of \( I \). We
denote by \( w_i' \) the tuple in \( \text{chase}_\Sigma(I) \) that correspond to \( w_i \).

Lemma 2. [BV2] Let \( \Sigma \) be a set of fd's and let \( I = \{ w_1, \ldots, w_k \} \) be a relation.

1. \( \Sigma \models \langle w, I \rangle \) if and only if \( w \in \text{chase}_\Sigma(I) \).

2. \( \Sigma \models \langle (w_i[A], w_j[A]), I \rangle \) if and only if \( w_i'[A] = w_j'[A] \).

3.2. The Proof

Suppose that \( R = \{ R_1, \ldots, R_k \} \) has a lossless join with respect to a set \( \Sigma \) of fd's. That
is, \( \Sigma \models ^* [R] \). If \( U \in R \), then the claim of the theorem is trivially satisfied. So assume that
\( U \notin R \). Let \( \tau_R = \langle w, I \rangle \). That is, \( I = \{ w_1, \ldots, w_k \} \), \( w_i[R_i] = w[R_i] \), and if \( A \notin R_i \) then \( w_i[A] \)
has a unique occurrence in \( I \). Since \( \Sigma \models ^* [R] \), also \( \Sigma \models \tau_R \). By Lemma 2, there is some \( i \) such
that \( w_i' = w \). That is, \( w_i'[A] = w[A] \) for all \( A \notin R_i \). By Lemma 2 we have that
\( \Sigma \models \langle (w_i[A], w[A]), I \rangle \), for all \( A \notin R_i \). Let \( \tau_{R_i \rightarrow A} = \langle (a_0, a_1), I \rangle \). That is, \( J \) consists of two
tuples \( u \) and \( v \) such that \( u[R_i] = v[R_i] \), \( u[B] \neq v[B] \) for all \( B \in R_i \), \( u[A] = a_0 \), and \( v[B] = a_1 \).
We define a valuation \( h \) such that \( h(w_i) = u \) and \( h(w_j) = v \) for \( j \neq i \). It is easy to see that \( h \) is
well defined, \( h(I) = J \), \( h(w_i[A]) = a_0 \) and \( h(w[A]) = a_1 \). By Lemma 1,
\( \langle (w_i[A], w[A]), I \rangle \models \tau_{R_i \rightarrow A} \). It follows that \( \Sigma \models R_i \rightarrow A \) for all \( A \in R_i \), and, consequently,
\( \Sigma \models R_i \rightarrow U \).
4. The Sufficient Condition

4.1. Preliminary Notions

We prove the sufficient condition via an excursion through the theory of join dependencies. We need first to review some definitions.

For an asc $R = \{R_1, \ldots, R_k\}$, the attribute set of $R$, denoted $\text{attr}(R)$, is $\bigcup_{i=1}^{k} R_i$. A join dependency (jd) is a statement $*[R]$, where $R$ is an asc $\{R_1, \ldots, R_k\}$. (We also use the notation $*[R_1, \ldots, R_k]$.) It is satisfied by a relation $l$ if $l[\text{attr}(R)] = \bigcap_{i=1}^{k} l[R_i]$. Note that a tjd is a special case of a jd. Jd's are called in [MMS] embedded jd's. Let $R$ and $S$ be asc's such that $\text{attr}(R) = \text{attr}(S)$. We say that $S$ covers $R$, denoted $R \subseteq S$, if for all $R \in R$ there is some $S \in S$ such that $R \subseteq S$.

**Lemma 3.** [BV] Let $X$, $Y$, and $Z$ be attribute sets, and let $R$ and $S$ be asc's. Then

1. $\emptyset \models *[X]$.
2. If $R \subseteq S$, then $*[R] \models *[S]$.
3. If $\text{attr}(S) \subseteq R$, then $*[R, *[S]] \models *[R \setminus \{\text{attr}(S)\} \cup S]$.
4. $X \rightarrow Y \models *[XY, XZ]$. 

4.2. The Proof

It is shown in [BB] that if $\Sigma \models R \rightarrow U$, then there is a sequence $X_1 \rightarrow Y_1, \ldots, X_n \rightarrow Y_n$ of fd's from $\Sigma$ with the following property: Let $R_0 = R$ and $R_i = R_{i-1}Y_i$, $1 \leq i \leq n$. Then $X_i \subseteq R_{i-1}$, for $1 \leq i \leq n$, and $R_n = U$. We show by induction on $i$ that $\Sigma \models *[R, X_1Y_1, \ldots, X_iY_i]$.

**Basis** ($i = 0$). By (1) of Lemma 3, we have $\Sigma \models *[R]$.

**Induction.** Suppose that $\Sigma \models *[R, X_1Y_1, \ldots, X_{i-1}Y_{i-1}]$. Since $X_i \subseteq R_{i-1}$, we have that $X_i \rightarrow Y_i \models *[X_iY_i, R_{i-1}]$ by (4) of Lemma 3. Since $R \cup \bigcup_{j=1}^{i-1} X_jY_j = R_{i-1}$,
\[ \Sigma \models^* [R, X_1 Y_1, \ldots, X_n Y_n] \] by (3) of Lemma 4.

In particular we have \( \Sigma \models^* [R, X_1 Y_1, \ldots, X_n Y_n] \). But by assumption \( R \in R \) and \( \Sigma \) is embeddable in \( R \), so \( \{R, X_1 Y_1, \ldots, X_n Y_n\} \leq R \). It follows that \( \Sigma \models^* [R] \) by (2) of Lemma 4.

References


