Acta Informatica 36, 173–213 (1999)

Semantic foundations of 4NF in relational database design

Millist W. Vincent

School of Computer and Information Science, University of South Australia, Adelaide, Australia 5095 (e-mail: m.vincent@unisa.edu.au)

Received: 7 July 1995 / 28 December 1998

Abstract. The issue of providing a formal justification for the use of fourth normal form (4NF) in relational database design is investigated. The motivation and formal definitions for three goals of database design are presented. These goals are the elimination of: redundancy, key-based update anomalies and fact-based replacement anomalies. It is then shown that, depending on the type of constraints permitted, either Boyce-Codd normal form (BCNF) or 4NF are the exact conditions needed to ensure most of the design goals. However, it is also shown that the conditions required to ensure the absence of a particular class of key-based update anomaly are new normal forms which have not previously been identified. In particular, for the case where the only constraints are functional dependencies (FDs), it is shown that the required normal form is a new normal form that is stronger than third normal form (3NF) yet weaker than BCNF. Similarly, in the more general case where both FD and multivalued dependencies (MVDs) are present, the required normal form is a new normal form that is weaker than 4NF.

1 Introduction

Originating with the pioneering work of Codd [11], the theory of *normal forms* is one of the oldest topics in relational database theory. However the issue of understanding and justifying the use of normal forms from a semantic perspective is one that, although mentioned as an unsolved problem in database theory [31], has not been completed. In most works defining normal forms [11, 12, 15, 41], the emphasis has been on the syntactic properties of the normal forms rather than on their semantic justification. In the simplest case where the only constraints are *functional dependencies* (FDs), other researchers have addressed the problem of providing a semantic justification for BCNF and have shown that it is equivalent to certain desirable semantic properties [7, 8, 10, 39]. However, little research has addressed the same issue for *fourth normal form* (4NF) [15] in the more general case when *multivalued dependencies* [15] are present. The purpose of this paper is to address the issue by providing a comprehensive formal analysis of the relationship between 4NF and several desirable semantic properties of database design.

The desirable design properties investigated in this paper are the elimination of: *redundancy*, *key-based update anomalies* and *fact-based replacement anomalies*. While it's not claimed that this is an exhaustive list, it does include the main approaches that have been proposed in the literature during the last few decades. The motivation for each of these properties will now be briefly outlined (a more detailed presentation is contained in [35]).

The motivation for eliminating *redundancy* is based on the minimal principle which aims to store each unit of information only once in a database. Eliminating redundancy thus minimises storage usage and also avoids the associated difficulty in duplicated data of having to update all occurrences of a data item. In another paper [36] we proposed a formal definition of redundancy based on interpreting the set of attributes XY in an FD $X \to Y$ or MVD $X \to\to Y$ as the fundamental unit of information or *fact*. The difficulty with this approach is that it is dependent on the syntactic structure of FDs and MVDs and it is not clear how to generalise this definition to other types of relational dependencies, such as *join dependencies* [28], or to other data models. In this paper we propose a more fundamental definition of redundancy that corrects these deficiencies. We consider the occurrence of an attribute value in a relation to be redundant if it can be derived from the other data values in the relation and the set of dependencies which apply to the relation, i.e. the occurrence is 'fixed' by the other data in the relation and the set of dependencies. More precisely, a relation scheme is *redundant* if there exists a legal relation (satisfies the constraints) defined over the scheme containing an occurrence of a data value such that *any* change to the occurrence results in the violation of the dependencies. For example, consider the relation scheme $\{A, B, C, D\}$ and the set of dependencies $\{A \to B, A \to\to C\}$. Each of the occurrences of b_1 in the first four tuples of the relation of Fig. 1 is redundant since any change results in $A \to B$ being violated. Similarly, all occurrences of c_1 (and also c_2, d_1, d_2) in the first four tuples are redundant.

The second semantic aim of normalisation, that of avoiding *key-based update anomalies*, was introduced in [16]. A key-based update anomaly is

A	B	C	D
a_1	b_1	c_1	d_1
a_1	b_1	c_2	d_2
a_1	b_1	c_1	d_2
a_1	b_1	c_2	d_1
a_2	b_1	c_2	d_1

Fig. 1. A relation containing redundancy

defined to occur when an update[1] to a relation results in the new relation satisfying key uniqueness (no two tuples in the relation have the same value for a candidate key) but violating some other constraint on the relation. The reason for this being considered undesirable is that the enforcement of key uniqueness can be, and is [13], relatively easily enforced in relational database software whereas the same is not true for more general constraints such as FDs or MVDs. Thus if the satisfaction of all the constraints on a relation is a result of key uniqueness then the integrity of the relation after an update can be easily enforced, whereas the existence of a key-based update anomaly implies the converse.

In [16], only insertions and deletions were considered and in this paper we extend the approach to the modification of tuples as well as extending the results in [16] on deletion anomalies and insertion anomalies. Consistent with this key-based approach, we define a relation to have *modification anomaly* if the modification of a tuple in the relation results in the violation of the dependencies although key uniqueness is preserved. We also define some additional types of modification anomalies which satisfy the extra condition that the '*identity*' of the tuple be preserved by the modification. The motivation for this extra condition is based on the observation that in practice it is often undesirable to change the identity of a tuple because of the need to also update associated foreign key references as well as possible confusion as to which real world entity the tuple refers to. Most commercial relational systems recognise this need and allow the values of specified attribute sets to be immutable [13]. In the relational model, it is normal to equate the identity of a tuple with its value on a candidate key but, since there may be multiple candidate keys, there are several possibilities as to what could be interpreted by preserving the identity of a tuple. The three possibilities considered in this paper are: (i) at least one (arbitrary) candidate

[1] Update is here used in a general sense and means either the insertion, deletion or modification of a tuple in a relation.

key of the original tuple is unchanged by the modification; (ii) the primary (fixed) key of the original tuple is unchanged by the modification; (iii) all candidate keys of the original tuple are unchanged by the modification.

The final semantic property analysed is the elimination of *fact-based modification anomalies*. This approach to justifying normalisation is closest to the original intuitive justification of normal forms in [11]. In this approach, the set of attributes in an MVD or FD constraint, rather than all the attributes in a relation scheme, is interpreted as the as the fundamental unit of information or *fact* for retrieval and update. In essence, a fact-based update anomaly occurs when fact values cannot be independently updated without violating either the basic properties of the relational model or the dependencies. Then, for each of the operations of insertion, deletion, and modification (also called a *replacement*), fact-based update anomalies can be formally defined [10]. However, in this paper we restrict our attention to the case of replacements. The reasons for this are that while replacements can adequately handled without considering null values, a thorough treatment of fact-based insertion and deletion anomalies requires the consideration of nulls and is outside the scope of this paper.

We now outline the structure of this paper and summarise the main results obtained. Section 2 contains definitions of basic relational concepts. In Sect. 3, formal definitions are given for redundancy and RFNF, the associated normal form for relation schemes which ensures the absence of redundancy. We refer to RFNF, and the other normal forms to be discussed later in the context of other semantic properties, as *semantic normal forms*. The classical normal forms of 3NF, BCNF and 4NF will often be referred to as *syntactic normal forms*. The reason for emphasising the difference is the different nature of the two types of normal forms. Semantic normal forms encapsulate the desired semantic property by requiring that all relations defined over a scheme will have the specified property, while syntactic normal forms are expressed in terms of the syntactic structure of the constraints. The main result derived in Sect. 3 is that 4NF is equivalent to RFNF. In Sect. 4, formal definitions are given for the various type of key based update anomalies and the associated normal forms in which the anomalies are absent. In Sect. 5 we prove that 4NF is equivalent to the absence of key-based insertion anomalies. In Sect. 6 we show that the condition equivalent to the absence of key-based deletion anomalies is that the set of dependencies is equivalent to a set of FDs, a weaker condition than 4NF. In Sect. 7 we show 4NF is equivalent to the semantic normal forms which ensure the absence of three of the four types of modification anomalies. However, for the normal form which ensures the absence of a modification anomaly in which all key values are preserved, we show in Sect. 8 that the equivalent syntactic normal forms are new normal forms that have not appeared before in the literature.

In the case where only FDs are present, we show that the equivalent syntactic normal lies between 3NF and BCNF. Similarly, for the case where MVDs are also present, the equivalent syntactic normal form is a weaker condition than 4NF which requires that every attribute be a member of a candidate key. In Sect. 9 three types of fact-based modification are defined and 4NF is shown to be equivalent to their absence. A discussion of related work is contained in Sect. 9 and concluding comments are made in Sect. 10.

2 Basic concepts and notation

The notation used in this paper is the standard notation used in the database literature [22, 27]. A universe U is a finite set of attributes, each attribute having an associated domain of values. The *domain* of an attribute $A \in U$ is denoted by DOM(A) and in this paper is assumed to be *infinite*. As usual, the symbols $A, B, C, \ldots$ and their subscripts represent single attributes and $V, W, X, \ldots$ and their subscripts denote sets of attributes. The union of attribute sets X and Y is denoted by XY rather than $X \cup Y$. $X - Y$ denotes set difference.

A *relation scheme* R is a subset of U. Let the elements of a relation scheme be denoted by $R = \{A_1, \ldots, A_n\}$. A *tuple* over R is an element of DOM$(A_1) \times \ldots \times$ DOM(A_n) where × denotes the cartesian product. A *relation instance* (or simply a *relation*) r over R, denoted by $r(R)$, is a *finite* set of tuples defined over R. In this paper, all relations are defined over a single relation scheme R and so r(R) will be denoted simply by r. If t is a tuple over R and X is a subset of R, then $t[X]$ is the *restriction* of t to the attributes in X.

2.1 Functional and multivalued dependencies

A relation r *satisfies* the functional dependency (FD) $X \to Y$ on a relation scheme R if for all $t_1, t_2 \in r$, if $t_1[X] = t_2[X]$ then $t_1[Y] = t_2[Y]$, otherwise it *violates* the FD. A relation r satisfies the *multivalued dependency* (MVD) $X \to\to Y$ on R if for all $t_1, t_2 \in r$ with $t_1[X] = t_2[X]$, there exists a tuple $t_3 \in r$ such that $t_3[X] = t_1[X]$, $t_3[Y] = t_1[Y]$ and $t_3[R - XY] = t_2[R - XY]$. We shall assume that X and Y in any MVD $X \to\to Y$ are disjoint because of the result that $X \to\to Y$ is satisfied if and only if $X \to\to Y - X$ is satisfied [15]. A set of FDs and MVDs *apply* to a relation scheme R if the attributes in every dependency are members of R. Since we are considering only a single relation scheme, it will be assumed that a set of dependencies always applies to the relation scheme. The set of all relations which satisfy a set $\mathbf{\Sigma}$ of FDs and MVDs is denoted by SAT($\mathbf{\Sigma}$). An MVD or FD is *standard* if the lhs of the dependency is not empty. The set NS($\mathbf{\Sigma}$) is defined as the set of all sets of attributes which appear on the rhs of a nonstandard MVD or FD.

Given a set $\mathbf{\Sigma}$ of FDs and MVDs and an FD $Z \rightarrow W$ (or MVD $Z \rightarrow\rightarrow W$), $\mathbf{\Sigma}$ *implies* the FD $Z \rightarrow W$ (or MVD $Z \rightarrow\rightarrow W$) if every relation in SAT($\mathbf{\Sigma}$) also satisfies $Z \rightarrow W$ (or $Z \rightarrow\rightarrow W$). One can decide whether a set $\mathbf{\Sigma}$ of FDs and MVDs implies another FD or MVD by using proofs based on a finite application of rules from the following set of inference rules [4].

FD rules:

A1: If $Y \subseteq X$ then $X \rightarrow Y$
A2: If $X \rightarrow Z$ and $Y \subseteq R$ then $XY \rightarrow ZY$
A3: If $X \rightarrow Z$ and $Y \rightarrow Z$ then $X \rightarrow Y$

MVD rules:

A4: If $X \rightarrow\rightarrow Y$ then $X \rightarrow\rightarrow (R - XY)$
A5: If $X \rightarrow\rightarrow Y$ and $V \subseteq W$ then $WX \rightarrow\rightarrow VY$
A6: If $X \rightarrow\rightarrow Y$ and $Y \rightarrow\rightarrow Z$ then $X \rightarrow\rightarrow Z - Y$
A7: If $Y \subseteq X$ then $X \rightarrow\rightarrow Y$

Combined FD and MVD rules:

A8: If $X \rightarrow Y$ then $X \rightarrow\rightarrow Y$
A9: If $X \rightarrow\rightarrow Y, Z \subseteq Y, W \cap Y = \phi$ and $W \rightarrow Z$ then $X \rightarrow Z$

The following rules, although derivable from those above, are useful and will be used later.

A10: If $X \rightarrow YZ$ then $X \rightarrow Y$
A11: If $X \rightarrow\rightarrow Y$ and $X \rightarrow\rightarrow Z$ then $X \rightarrow\rightarrow YZ$

A dependency is *trivial* if it is satisfied by every relation. An FD $X \rightarrow Y$ is trivial if and only if $Y \subseteq X$ and an MVD $X \rightarrow\rightarrow Y$ is trivial if and only if $Y \subseteq X$ or $R = XY$ [22]. The *closure* of a set $\mathbf{\Sigma}$ of FDs and MVDs, denoted by $\mathbf{\Sigma}^+$, is the set of FDs and MVDs implied by $\mathbf{\Sigma}$. Two sets of dependencies, $\mathbf{\Sigma}$ and $\mathbf{\Psi}$, are *equivalent*, written as $\mathbf{\Sigma} \equiv \mathbf{\Psi}$, if $\mathbf{\Sigma}^+ = \mathbf{\Psi}^+$. If $\mathbf{\Sigma} \equiv \mathbf{\Psi}$, then $\mathbf{\Psi}$ is a *cover* for $\mathbf{\Sigma}$. The closure of a set of attributes X, denoted by X^+, is the set of attributes such that an attribute $A \in X^+$ if and only if $X \rightarrow A \in \mathbf{\Sigma}^+$.

The *dependency basis* for a set of attributes X, denoted by DEP(X), is a set of attributes sets which can be written as $\{X_1$, ..., X_p, X_l^+, ..., X_j^+, W_1, ..., $W_n\}$ with the following properties [3]:

(i) DEP(X) covers R, i.e. $R = \cup Z_i$ where $Z_i \in$ DEP(X);
(ii) The sets in DEP(X) are disjoint and nonempty;
(iii) $X \rightarrow\rightarrow Y \in \mathbf{\Sigma}^+$ if and only if $Y = \cup Z_i$ where $Z_i \in$ DEP(X);
(iv) $X_1, \ldots, X_p$ are single attribute sets such that $X = \overset{i=p}{\underset{i=1}{\cup}} X_i$;
(v) $X_l^+, \ldots, X_j^+$ are single attribute sets such that $X^+ - X = \overset{i=j}{\underset{i=1}{\cup}} X_i^+$.

The next concept required is that of a *reduced set* of FDs and MVDs.

Definition 2.1 *Let $\mathbf{\Sigma}$ be a set of FDs and MVDs. $\mathbf{\Sigma}$ is* reduced *if:*

(i) No dependency $d \in \mathbf{\Sigma}$ is redundant, i.e. for all $d \in \mathbf{\Sigma}$, $\mathbf{\Sigma} - \{d\}$ is not equivalent to $\mathbf{\Sigma}$;
(ii) Every dependency is left-reduced, i.e. for every MVD $X \to\to Y$ (or FD $X \to Y$) $\in \mathbf{\Sigma}$, there is no MVD $X' \to\to Y$ (or FD $X' \to Y$) $\in \mathbf{\Sigma}^+$ such that $X' \subset X$;
(iii) every dependency is right-reduced, i.e. for every MVD $X \to\to Y$ (or FD $X \to Y$) $\in \mathbf{\Sigma}$, there is no MVD $X \to\to Y'$ (or FD $X \to Y'$) $\in \mathbf{\Sigma}^+$ such that $\emptyset \subset Y' \subset Y$.

We note that this definition is weaker than the definition in [26] since we do not impose the condition that no set of attributes be able to be transferred from the lhs to the rhs of a dependency. Also, it can be easily verified that the following procedure terminates and generates a reduced cover for any set $\mathbf{\Sigma}$ of FDs and MVDs[2].

```
Input: A set $\mathbf{\Sigma}$ of FDs and MVDs
Output: A reduced cover for $\mathbf{\Sigma}$
Repeat
  For each dependency $d$ in $\mathbf{\Sigma}$ do
    If $d$ is redundant then $\mathbf{\Sigma} := \mathbf{\Sigma} - d$;
    If $d$ is not left-reduced then $\mathbf{\Sigma} := (\mathbf{\Sigma} - d) \cup d'$,
    where $d'$ is the dependency in $\mathbf{\Sigma}^+$ whose lhs
    is a proper subset of the lhs of $d$;
    If $d$ is not right-reduced then
      If $d$ is an FD $X \to Y$ and $X \to Y' \in \mathbf{\Sigma}^+$
        then $\mathbf{\Sigma} := (\mathbf{\Sigma} - \{X \to Y\}) \cup \{X \to Y', X \to Y - Y'\}$
      else ($d$ is an MVD $X \to\to Y$ and $X \to\to Y' \in \mathbf{\Sigma}^+$)
        $\mathbf{\Sigma} := (\mathbf{\Sigma} - \{X \to\to Y\}) \cup \{X \to\to Y', X \to\to Y - Y'\}$
Until no more changes can be made to $\mathbf{\Sigma}$.
```

Given a set $\mathbf{\Sigma}$ of FDs and MVDs, a set of attributes X is a *superkey* for a relation scheme R if the FD $X \to R \in \mathbf{\Sigma}^+$. X is a *candidate key* for R if it is a superkey and it has no proper subset X' such that $X' \to R \in \mathbf{\Sigma}^+$. The set of all superkeys in a scheme R is denoted by SK$(R, \mathbf{\Sigma})$ and the set of all candidate keys by CK$(R, \mathbf{\Sigma})$. The *primary key* of a relation scheme, denoted by K_p, is an arbitrarily chosen candidate key. An attribute is a *prime attribute* if it is a member of any candidate key. The set of *key constraints*, denoted by $\mathbf{\Sigma}_k$, is the set of all FDs in $\mathbf{\Sigma}^+$ of the form $K \to R$ where $K \in$ CK$(R, \mathbf{\Sigma})$. The set of all relations which satisfy $\mathbf{\Sigma}_k$ is denoted by SAT$(\mathbf{\Sigma}_k)$. Obviously, if a relation $r \in SAT(\mathbf{\Sigma})$ then $r \in$ SAT$(\mathbf{\Sigma}_k)$ but the

[2] We note that the rhs of every FD in a reduced set contains a single attribute.

converse is not true. Also, a relation $r \in \ \mathrm{SAT}(\mathbf{\Sigma}_k)$ if and only if no two tuples in the relation have the same value for a candidate key.

If $\mathbf{\Sigma}$ is a set of FDs and MVDs which apply to a relation scheme R, then $(R, \mathbf{\Sigma})$ is in *third normal form* $(3NF)$ if for every nontrivial FD $X \to A \in \mathbf{\Sigma}^+$, $X \in$ SK$(R, \mathbf{\Sigma})$ or A is prime [41]. $(R, \mathbf{\Sigma})$ is in *Boyce-Codd normal form* (*BCNF*) if for every nontrivial FD $X \to A \in \mathbf{\Sigma}^+$, $X \in$ SK$(R, \mathbf{\Sigma})$ [12] and is in *fourth normal form* (4NF) if for every nontrivial MVD $X \to\to Y \in \mathbf{\Sigma}^+$, $X \in$ SK$(R, \mathbf{\Sigma})$ [15].

2.2 Join dependencies

Let $R_1, \ldots, R_p$ be nonempty subsets of R. If there are tuples $t_1, \ldots, t_p$ (not necessarily distinct) in a relation r (R) such that $t_i[R_i \cap R_j] = t_j[R_i \cap R_j]$ for all i, j such that $1 \leq i \leq p$, $1 \leq j \leq p$, then $t_1, \ldots, t_p$ are said to *join completely* on $\{R_1, R_2, \ldots, R_p\}$. In this case, there exists a unique tuple t such that $t[R_i] = t_i[R_i]$, for all i such that $1 \leq i \leq p$, which is called the *join* of $t_1, \ldots, t_p$ on $\{R_1, \ldots, R_p\}$. A *join dependency* (JD) is a constraint denoted by $^*[R_1, \ldots, R_p]$. A relation r satisfies the join dependency $^*[R_1, \ldots, R_p]$ on R if for every set of tuples $t_1, \ldots, t_p \in r$ which join completely on $\{R_1, R_2, \ldots, R_p\}$, r also contains the join of $t_1, \ldots, t_p$ on $\{R_1, \ldots, R_p\}$. $R_1, \ldots, R_p$ are referred to as the *components* of the JD. The result [15] that any MVD $X \to\to Y$ is equivalent to the JD $^*[XY, XZ]$, where $Z = R - XY$, will be used often in this paper. A JD is trivial if it is satisfied by every relation r. It can be shown that a JD $^*[R_1, \ldots, R_p]$ is trivial iff there exists a component such that $R_i = R$.

A JD $^*[R_1, \ldots, R_p]$ is *total* (full) if $R = R_1 \ldots R_p$. Since, the only JDs considered in this paper are those equivalent to MVDs, it is easily seen that such a JDs is always full.

2.3 Tableau

A *tableau* is a matrix consisting of a set of rows [1, 23]. Each column in the tableau corresponds to an attribute in R. Each row consists of variables from a set V, which is the disjoint union of two sets V_d and V_n. V_d is the set of *distinguished variables* (dv's) and V_n is the set of *nondistinguished variables* (ndv's). Any variable can appear in at most one column, a dv must appear in each column and at most one dv can appear in a column.

A *valuation* is a function ρ that maps each variable to an element in DOM(A) where A is the column in which the variable appears. This is extended to a function from a tableau T to a relation over R in the obvious manner. Let $\mathbf{\Sigma}$ be a set of FDs and JDs (any MVD is treated as a JD). The *chase* is the result of applying the following transformations to a tableau T until no further changes can be made:

F-Rule: If $X \to A \in \mathbf{\Sigma}$ and T has rows ω_1 and ω_2 where $\omega_1[X] = \omega_2[X]$ and $\nu_1 = \omega_1[A]$ and $\nu_2 = \omega_2[A]$, then if either of ν_1 or ν_2 is a dv and the other is not, then every occurrence of the ndv is changed to the dv. If both are ndv's, then the one with the larger subscript is replaced by the one with the smaller subscript.

J-Rule: If $^*[R_1, \ldots, R_p] \in \mathbf{\Sigma}$ and there exists a row ω such that $\omega[R_1] \in T[R_1], \ldots, \omega[R_p] \in T[R_p]$, ω is added to T.

Let $chase_{\mathbf{\Sigma}}(T)$ be the tableau that results from applying the F-rules and J-rules until no more changes can be made to the tableau. It can then be shown [23] that the chase always terminates if all JDs are full (which is the case in this paper) and the resulting tableau is unique, independent of the sequence in which the rules are applied, up to a renaming of the ndv's. We will use the following results on the properties of the chase later in this paper [23]:

Lemma 2.1 *Any valuation ρ of $chase_{\mathbf{\Sigma}}(T)$ which is a one-to-one mapping satisfies $\mathbf{\Sigma}$.*

Lemma 2.2 *Let T_X be the tableau constructed as follows. It contains two rows, one row, denoted by ω_d, contains all dv's and the other, denoted by ω_X, contains dv's in the X-columns and ndv's elsewhere. If $T^* = chase_{\mathbf{\Sigma}}(T_X)$, then T^* contains the row ω_d and an FD $X \to Y \in \mathbf{\Sigma}^+$ iff the Y-columns in T^* contain only dv's.*

3 Redundancy and 4NF

In this section, we present a formal definition of the redundancy concept discussed in the Introduction and its associated semantic normal form which ensures the absence of redundancy. The main result derived in this section is that 4NF is a necessary and sufficient condition for the absence of redundancy.

Definition 3.1 *Let R be a relation scheme, A an attribute in R, $\mathbf{\Sigma}$ a set of dependencies, r a relation and t a tuple in r. The data value occurrence $t[A]$ is* redundant *(RED) if for* every *replacement of $t[A]$ by a value a' such that $t[A] \neq a'$ and resulting in a new relation r', then $r' \notin$ SAT($\mathbf{\Sigma}$).*

Based on this we define a semantic normal form in which redundancy is absent.

Definition 3.2 $(R, \mathbf{\Sigma})$ *is in* redundancy free normal form *(RFNF) if there does not exist $r \in$ $\mathrm{SAT}(\mathbf{\Sigma})$ which contains a data value occurrence that is RED.*

A	B	C
a_1 a_2	b_1 b_2	c_1 c_2

Fig. 2.

A	B	C	D	E	F	G
1	1	1	1	1	1	1
1	1	1	0	0	1	1

Fig. 3. An example illustrating Theorem 3.1

We emphasise that for a data value occurrence to be RED every change to the occurrence must result in a violation of the constraints or duplicate tuples. For example, consider the relation r in Fig. 2 which is defined over the attributes $\{A, B, C\}$ with the set of constraints being $\{A \rightarrow B, B \rightarrow C\}$.

Although changing b_1 to b_2 (or b_2 to b_1) in r results in $B \rightarrow C$ being violated, neither value is redundant according to our definition since all other changes to b_1 (or to b_2) do not result in constraint violation. Intuitively, this makes sense since neither b_1 nor b_2 is derivable from the values in the relation and the set of constraints. Also, we note that it is a specific occurrence of a data value in a tuple, rather than the value itself, which is redundant. For example, in Fig. 1 (see Sect. 1), the occurrence of b_1 in the first four tuples is RED but not the occurrence of b_1 in the last tuple.

Before establishing the first main result of this section, we present an important preliminary lemma that will be used extensively in this paper. The lemma can be derived from the results in [33].

Theorem 3.1 *Suppose $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$ and as before, denote the elements in DEP(X) by $\{X_1, \ldots, X_p, X_l^+, \ldots, X_j^+, W_1, \ldots, W_n\}$. Then every relation of two tuples for which the two tuples are different on every attribute in one of the W_i but equal on all other attributes is in SAT($\mathbf{\Sigma}$).*

The following example also illustrates this theorem.

Example 3.1 Let $R = \{A, B, C, D, E, F, G\}$ and $\mathbf{\Sigma} = \{AB \rightarrow\rightarrow DE, E \rightarrow\rightarrow F, E \rightarrow C\}$. Standard algorithms [3] can be used to show that if $X = AB$, then $X^+ = \{A, B, C\}$ and DEP(X) = $\{A, B, C, DE, F, G\}$. The relation shown in Fig. 3 satisfies $\mathbf{\Sigma}$.

We now derive the main result of this section.

Theorem 3.2 *$(R, \mathbf{\Sigma})$ is in RFNF iff it is in 4NF.*

Proof.
If: We shall show the contrapositive that if $(R, \mathbf{\Sigma})$ is not in RFNF then it is not in 4NF. If $(R, \mathbf{\Sigma})$ is not in RFNF then there exists $r \in \mathrm{SAT}(\mathbf{\Sigma})$, a tuple $t \in r$ and $A \in R$ such that every change to $t[A]$ results in the new relation violating $\mathbf{\Sigma}$. So if $t[A]$ is changed to a value a' such that $a' \notin r[A]$, resulting in a new tuple t' and a new relation r', then $r' \notin$ SAT($\mathbf{\Sigma}$). Suppose firstly that an FD $X \rightarrow Y$ is violated in r'. The violation must involve t', since $r \in \mathrm{SAT}(\mathbf{\Sigma})$ and t is the only tuple changed in r, and some other tuple t_1 such that $t'[X] = t_1[X]$ and $t'[Y] \neq t_1[Y]$ and t_1 is also in r, again because t is the only tuple changed in r. Since $a' \notin r[A]$ and $t'[X] = t_1[X]$, then $A \notin X$ and so $t'[X] = t[X]$ and thus $t_1[X] = t[X]$. Hence there are two tuples in r, t and t_1, which are identical on X and so $X \notin$ SK($R, \mathbf{\Sigma}$) and hence $(R, \mathbf{\Sigma})$ is not in 4NF.

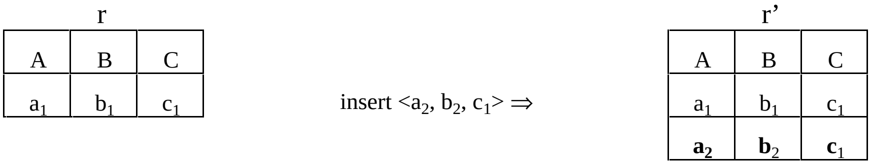
Alternatively, assume that an MVD $X \rightarrow\rightarrow Y$ is violated in r' and so there exists again t_1 where $t_1 \in r$ and $t_1 \in r'$ such that $t_1[X] = t'[X]$. So, since $a' \notin r[A]$, $A \in Y$ or $A \in Z$ where $Z = R - XY$. Again this implies $t_1[X] = t[X]$ in r and so contradicts the 4NF assumption.

Only If: The contrapositive that if $(R, \mathbf{\Sigma})$ is not in 4NF then it is not in RFNF will be shown. Because $(R, \mathbf{\Sigma})$ is not in 4NF there exists a nontrivial MVD $X \rightarrow\rightarrow Y \in \mathbf{\Sigma}^+$ where $X \notin$ SK($R, \mathbf{\Sigma}$) and so there exists $W_i \in$ DEP(X) such that $X^+ \cap W_i = \emptyset$. By Theorem 3.1, any relation r of two tuples which are identical on all attributes except those in W_i is in SAT($\mathbf{\Sigma}$). Firstly, if there exists an $X_j^+ \in \mathrm{DEP}(X)$ then both values of X_j^+ are RED in r since changing either causes $X \rightarrow X_j^+$ to be violated. Alternatively, if $X = X^+$ then we claim that there are two sets W_i and W_j in DEP(X) disjoint from X. If there is only W_i then, since $X \cap Y = \emptyset$, property (iii) of DEP(X) implies $Y = W_i$ and so, by property (i), $XY = R$ contradicting the fact that $X \rightarrow\rightarrow Y$ is nontrivial. It then follows that every value in W_j is RED in r. □

Since 4NF reduces to BCNF if only FDs are present, this result also shows that BCNF is the exact condition required to avoid redundancy when only FDs are present. It also follows from this result and the results in [36] that the redundancy property defined in that paper is equivalent to the one defined in this section.

4 Key-based update anomalies

In this section we give formal definitions of the various types of key-based update anomalies and the associated semantic normal forms in which these



r

A	B	C
a_1	b_1	c_1

insert <a_2, b_2, c_1> ⇒

r’

A	B	C
a_1	b_1	c_1
$\mathbf{a_2}$	$\mathbf{b}_2$	$\mathbf{c}_1$

Fig. 4. An example of an insertion anomaly

anomalies are absent. The definitions of an insertion anomaly and a deletion anomaly are taken from [16] whereas the definitions of modification anomalies are new.

4.1 Key-based insertion anomaly

Definition 4.1 *Let R be a relation scheme, $\mathbf{\Sigma}$ a set of dependencies and r a relation. A tuple t^* is said to be* compatible *with r if $t^* \notin r$*[3] *and $r \cup \{t^*\} \in$ SAT($\mathbf{\Sigma}_k$).*

As mentioned in Sect. 2, a relation is in SAT($\mathbf{\Sigma}_k$) if and only if no two tuples in the relation have the same value for any candidate key and so, if r in SAT($\mathbf{\Sigma}$), then t^* is compatible with r if and only if $t^*[K] \notin r[K]$ for all $K \in \mathrm{CK}(R, \mathbf{\Sigma})$. We now use this concept to define an insertion anomaly and a corresponding normal form.

Definition 4.2 *A relation r has a* key-based insertion anomaly *(*KIA*) w.r.t. to a set of dependencies $\mathbf{\Sigma}$ if:*

(i) $r \in \mathrm{SAT}(\mathbf{\Sigma})$*;*
(ii) there exists a tuple t^* such that t^* is compatible with r but $r \cup \{t^*\} \notin$ SAT($\mathbf{\Sigma}$).

Definition 4.3 *$(R, \mathbf{\Sigma})$ is in* key-based insertion normal form *(*KINF*) if there does not exist $r \in \mathrm{SAT}(\mathbf{\Sigma})$ which has a KIA w.r.t. to $\mathbf{\Sigma}$.*

The following example illustrates the previous definitions.

Example 4.1 Let $R = \{A, B, C\}$ and $\mathbf{\Sigma} = \{AB \to C, C \to\to A\}$. The only candidate key is AB and the relation r shown in Fig. 4 is in SAT($\mathbf{\Sigma}$). However, $(R, \mathbf{\Sigma})$ is not in KINF because r has a KIA w.r.t. $\mathbf{\Sigma}$ when the tuple $\langle a_2, b_2, c_1 \rangle$ is inserted into it since the resulting relation, r', $\in$ SAT($\mathbf{\Sigma}_k$) but violates $A \to\to B$.

[3] This ensures that $r \cup \{t^*\}$ is a relation, i.e. there are no duplicate tuples in $r \cup \{t^*\}$.

r

A	B	C
a_1	b_1	c_1
a_1	b_2	c_2
a_1	b_1	c_2
a_1	**b_2**	**c_1**

delete <a_1, b_2, c_1> ⇒

r’

A	B	C
a_1	b_1	c_1
a_1	b_2	c_2
a_1	b_1	c_2

Fig. 5. An example of a deletion anomaly

4.2 Key-based deletion anomaly

In a similar fashion to an insertion anomaly and insertion normal form, a deletion anomaly and deletion normal form are defined as follows.

Definition 4.4 *A relation r has a* key-based deletion anomaly *(*KDA*) w.r.t. to a set of dependencies $\mathbf{\Sigma}$ if:*

(i) $r \in \text{SAT}(\mathbf{\Sigma})$*;*
(ii) there exists a tuple $t^* \in r$ such that $r - \{t^*\} \notin$ SAT($\mathbf{\Sigma}$)[4].

Definition 4.5 *$(R, \mathbf{\Sigma})$ is in* key-based deletion normal form *(*KDNF*) if there doesn’t exist $r \in \text{SAT}(\mathbf{\Sigma})$ which has a KDA w.r.t. $\mathbf{\Sigma}$.*

The following example illustrates these definitions.

Example 4.2 Let $R = \{A, B, C\}$ and $\mathbf{\Sigma} = \{A \rightarrow\rightarrow B\}$. Since $\mathbf{\Sigma}$ contains no FDs, the only candidate key is ABC. $(R, \mathbf{\Sigma})$ is not in KDNF because the relation r shown in Fig. 5 has a KDA when the tuple is $\langle a_1, b_2, c_1\rangle$ is deleted from it since $r \in \text{SAT}(\mathbf{\Sigma})$ but the resulting relation, r', $\in$ SAT($\mathbf{\Sigma}_k$) but violates $A \rightarrow\rightarrow B$.

In the case of the set of constraints containing only FDs, a relation can have no deletion anomaly because of the result that if a relation satisfies a set of FDs then so does any subset of the relation [22].

4.3 Key-based modification anomalies

In this section, we extend the key-based approach of [16] to the modification of tuples and define several classes of a new type of update anomaly, called a key-based modification anomaly. Essentially a key-based modification anomaly occurs when the modification of a tuple results in the violation of the constraints even though key-uniqueness is maintained. By modelling a modification as a deletion then an insertion, the formal definition follows.

[4] It follows from part (i) of the definition that $r - \{t^*\} \in \text{SAT}(\mathbf{\Sigma}_k)$.

Definition 4.6 *A relation r has a* key-based modification anomaly 1 *(KMA$_1$) w.r.t. to a set of dependencies $\mathbf{\Sigma}$ if there exists $t \in r$ and a tuple t^* defined over R such that:*

(i) $r \in \mathrm{SAT}(\mathbf{\Sigma})$*;*
(ii) t^* *is compatible with* $(r - \{t\})$*;*
(iii) $(r - \{t\}) \cup \{t^*\} \notin$ *SAT($\mathbf{\Sigma}$).*

In the case of FD constraints, if a relation r has a KMA$_1$ then $r - \{t\}$ has a KIA (see Definition 4.2) since $r - \{t\} \in \mathrm{SAT}(\mathbf{\Sigma})$ if $r \in \mathrm{SAT}(\mathbf{\Sigma})$, but this property does not extend in the presence of MVDs since then $r \in \mathrm{SAT}(\mathbf{\Sigma})$ does not imply $r - \{t\} \in \mathrm{SAT}(\mathbf{\Sigma})$.

The next type of modification anomaly is motivated by the observation, discussed more thoroughly earlier, that it is often undesirable to change a tuple's identity during a modification. However, in general there may be multiple candidate keys and so we propose three possible interpretations as to what is meant by leaving the identity of a tuple unchanged. In increasing restrictiveness, they are:

(i) the replacement tuple is identical to the original on *any (arbitrary) candidate key*;
(ii) the replacement tuple is identical to the original on *the primary (fixed) key*;
(iii) the replacement tuple is identical to the original on *every candidate key*.

For each of these alternatives, we now present a formal definition.

Definition 4.7 *A relation r has a* key-based modification anomaly 2 *(KMA$_2$) w.r.t. to a set of dependencies $\mathbf{\Sigma}$ if there exists $t \in r$ and a tuple t^* defined over R such that:*

(i) $r \in \mathrm{SAT}(\mathbf{\Sigma})$*;*
(ii) t^* *is compatible with* $(r - \{t\})$*;*
(iii) there exists $K \in \mathrm{CK}(R, \mathbf{\Sigma})$ *such that* $t[K] = t^*[K]$*;*
(iv) $(r - \{t\}) \cup \{t^*\} \notin$ *SAT($\mathbf{\Sigma}$).*

Definition 4.8 *A relation r has a* key-based modification anomaly 3 *(KMA$_3$) w.r.t. to a set of dependencies $\mathbf{\Sigma}$ if it satisfies all conditions of Definition 4.7 except that condition (iii) is changed to:*
(iii') $t[K_p] = t^*[K_p]$*;*

Definition 4.9 *A relation r has a* key-based modification anomaly 4 *(KMA$_4$) w.r.t. to a set of dependencies $\mathbf{\Sigma}$ if it satisfies all the conditions of Definition 4.7 except that condition (iii) is changed to:*
(iii") $t[K] = t^*[K]$ *for all* $K \in \mathrm{CK}(R, \mathbf{\Sigma})$*;*

r

A	B	C	D
a$_1$	b$_1$	c$_1$	d$_1$
a$_2$	b$_1$	c$_1$	d$_1$

r’

A	B	C	D
a$_1$	b$_1$	c$_1$	d$_1$
a$_2$	b$_1$	c$_1$	**d$_2$**

r"

A	B	C	D
a$_1$	b$_1$	c$_1$	d$_1$
a$_2$	b$_1$	**c$_2$**	d$_1$

Fig. 6. An example illustrating modification anomalies

The following example illustrates the previous definitions.

Example 4.3 Let $R = \{A, B, C, D\}$ and $\mathbf{\Sigma} = \{ABC \to D, D \to C, B \to\to A\}$. The candidate keys are ABC and ABD and the relation r in Fig. 6 is in SAT($\mathbf{\Sigma}$). If the tuple $t = \langle a_2, b_1, c_1, d_1\rangle$ is changed to $t^* = \langle a_2, b_1, c_1, d_2\rangle$, resulting in the relation r', then r has a KMA$_2$ (and thus also a KMA$_1$). To verify this, each of the conditions of a KMA$_2$ will be verified. Condition (i) holds since $r \in \text{SAT}(\mathbf{\Sigma})$. Condition (ii) follows because both tuples in r' are distinct on both candidate keys. Condition (iii) holds since $t[ABC] = t^*[ABC]$ and (iv) holds because r' violates $B \to\to A$.

If ABC is chosen as the primary key, then r also has a KMA$_3$ when t is replaced by t^* since $B \to\to A$ is still violated. If ABD is chosen as the primary key, then replacing t by t^* does not constitute a KMA$_3$ since $t[ABD] \neq t^*[ABD]$. However, if instead t is replaced by $\langle a_2, b_1, c_2, d_1\rangle$, resulting in the relation r'', then r has a KMA$_3$ since $r'' \in \text{SAT}(\mathbf{\Sigma}_k)$ but violates $B \to\to A$.

Also, neither r nor any other relation can have a KMA$_4$. This follows because every attribute is prime and so any modified tuple satisfying condition (iii”) must be identical to the original, but then (i) and (iv) cannot be satisfied simultaneously.

We now use these definitions of anomalies in relation instances to define the semantic normal forms which are free of these anomalies.

Definition 4.10 *$(R, \mathbf{\Sigma})$ is in* key-based modification normal form 1 *(KMNF$_1$) if there doesn’t exist a relation r which has a KMA$_1$ w.r.t. $\mathbf{\Sigma}$.*

Definition 4.11 *$(R, \mathbf{\Sigma})$ is in* key-based modification normal form 2 *(KMNF$_2$) if there doesn’t exist a relation r which has a KMA$_2$ w.r.t. $\mathbf{\Sigma}$.*

Definition 4.12 *$(R, \mathbf{\Sigma})$ is in*key-based modification normal form 3 *(KMNF$_3$) if there doesn’t exist a relation r which has a KMA$_3$ w.r.t. $\mathbf{\Sigma}$.*

Definition 4.13 *$(R, \mathbf{\Sigma})$ is in* key-based modification normal form 4 *(KMNF$_4$) if there doesn’t exist a relation r which has a KMA$_4$ w.r.t. $\mathbf{\Sigma}$.*

From these definitions, it is easily seen that the following implications hold: r has a KMA$_4$ $\Rightarrow$ r has a KMA$_3$ $\Rightarrow$ r has a KMA$_2$ $\Rightarrow$ r has a KMA$_1$;

and hence the following implications also hold for the corresponding normal forms: $\text{KMNF}_1 \Rightarrow \text{KMNF}_2 \Rightarrow \text{KMNF}_3 \Rightarrow \text{KMNF}_4$.

Since equivalent sets of dependencies embody exactly the same logical information, it is important that any normal form possess the property of being *cover insensitive*, i.e. the property is independent of which equivalent cover is chosen. The classical syntactic normal forms 3NF, BCNF and 4NF are cover insensitive since they are defined as a property of the closure of the set of dependencies which, by definition, is the same for all equivalent covers. We now show that all the semantic normal forms defined in this section are also satisfy the property.

Theorem 4.1 *The update anomalies KIA, KDA, KMA*$_1$*, KMA*$_2$*, KMA*$_3$*, KMA*$_4$ *normal forms KINF, KDNF, KMNF*$_1$*, KMNF*$_2$*, KMNF*$_3$ *and KMNF*$_4$ *are cover insensitive.*

Proof. Immediate since, by definition, a set of attributes is a candidate key w.r.t. a set of dependencies iff it is a candidate key w.r.t. any equivalent cover and a relation violates a set of dependencies iff it violates any equivalent set by the definition of equivalence. □

5 KINF and 4NF

In this section, we show that 4NF and KINF, the normal form defined in Sect. 4, are equivalent. This result also follows from the results of [16] but our proof is more direct and is based on a stronger preliminary lemma.

Lemma 5.1 *If* $(R, \mathbf{\Sigma})$ *is not in 4NF then every nonempty* $r \in \text{SAT}(\mathbf{\Sigma})$ *has a KIA w.r.t.* $\mathbf{\Sigma}$.

Proof. If $(R, \mathbf{\Sigma})$ is not in 4NF then there exists a nontrivial MVD $X \rightarrow\rightarrow Y \in \mathbf{\Sigma}^+$ where $X \notin \text{SK}(R, \mathbf{\Sigma})$. Let t be any tuple in r and let t^* be the tuple defined by $t^*[X] = t[X]$ and $t^*[A] \notin r[A]$ for all $A \in (R - X)$. Such a tuple always exists because the domains are infinite. The claim is that r has a KIA when t^* is inserted into it. This is immediate from the fact that for any $K \in \text{CK}(R, \mathbf{\Sigma})$, $K - X \neq \emptyset$ since $X \notin \text{SK}(R, \mathbf{\Sigma})$, and the definition of t^*. □

Theorem 5.1 $(R, \mathbf{\Sigma})$ *is in 4NF iff it is in KINF.*

Proof.
If: The contrapositive, that if $(R, \mathbf{\Sigma})$ is not in 4NF then it is not in KINF, follows from Lemma 5.1 and the fact any relation containing a single tuple is in SAT($\mathbf{\Sigma}$).
Only if: Let $(R, \mathbf{\Sigma})$ be in 4NF and suppose to the contrary that $(R, \mathbf{\Sigma})$ is not in KINF. So there exists $r \in \text{SAT}(\mathbf{\Sigma})$ and a tuple $t^* \notin r$ such that

$r \cup \{t^*\} \in \text{ SAT}(\mathbf{\Sigma}_k)$ but $r \cup \{t^*\}$ violates a nontrivial dependency $X \to Y$ or $X \to\to Y$ in $\mathbf{\Sigma}$. For this to occur, there has to be at least two distinct tuples in $r \cup \{t^*\}$ which are identical on X. But this implies $X \notin \text{SK}(R, \mathbf{\Sigma})$ since $r \cup \{t^*\} \in \text{ SAT}(\mathbf{\Sigma}_k)$ which contradicts the assumption that $(R, \mathbf{\Sigma})$ is in 4NF. □

6 KDNF and 4NF

In this section the relationship between KDNF, the normal form defined in Sect. 4, and 4NF is investigated. We prove that 4NF is a stronger condition than KDNF and that KDNF is equivalent to the condition that the set of constraints is equivalent to a set of FDs.

Lemma 6.1 *If $(R, \mathbf{\Sigma})$ is in 4NF then it is in KDNF.*

Proof. Similar argument to the one used in Theorem 5.1. □

It was noted earlier that no relation scheme can have a KDA if the set of constraints contains only FDs. The following example shows that even in the presence of both FDs and MVDs, a relation scheme may have no KDA even though it is not in 4NF. In other words, 4NF is not a necessary condition for KDNF.

Example 6.1 Let $R = \{A, B, C\}$ and $\mathbf{\Sigma} = \{A \to\to B, C \to B\}$. The only candidate key is AC and so both dependencies violate 4NF. Using inference rules A9 and A8, $\mathbf{\Sigma}$ is equivalent to the set of FDs $\mathbf{\Sigma}' = \{A \to B, C \to B\}$. However, as noted earlier, a KDA is cover insensitive and so $(R, \mathbf{\Sigma})$ is in KDNF because $\mathbf{\Sigma}'$ contains only FDs.

We now introduce a restriction on the MVDs in the set of dependencies which will ensure a necessary and sufficient condition for a relation scheme to be in KDNF [19].

Definition 6.1 *Let $\mathbf{\Sigma}$ be a set of FDs and MVDs. An MVD $X \to\to Y \in \mathbf{\Sigma}$ is* pure *if it is nontrivial and $X \to Y \notin \mathbf{\Sigma}^+$ and $X \to R - XY \notin \mathbf{\Sigma}^+$. $\mathbf{\Sigma}$ is* pure *if it contains only FDs or if every MVD in $\mathbf{\Sigma}$ is pure.*

The following example illustrates the definition.

Example 6.2 Let $R = \{A, B, C\}$ and $\mathbf{\Sigma} = \{A \to\to B, B \to C\}$. From the inference rules $A \to C \in \mathbf{\Sigma}^+$ and so $A \to\to B$ is not pure. The MVD $A \to\to B$ in the set $\{A \to\to B, C \to B\}$ is also not pure since $A \to B \in \mathbf{\Sigma}^+$.

The motivation for the definition is to distinguish between those MVDs which convey 'true' multivalued information and those that only represent

FD information. This notion is captured in the following lemma which shows that the existence of a pure MVD is both a necessary and sufficient condition for the set of constraints not to be equivalent to a set of only FDs. Similar definitions aimed at ensuring that the MVDs are not FD equivalents have also been proposed by others [40]. It is also clear, using rules A8 and A4, that any set of FDs or MVDs has a pure cover generated by replacing each nonpure $X \rightarrow\rightarrow Y$ by $X \rightarrow Y$ if $X \rightarrow Y$ in $\mathbf{\Sigma}^+$, or by $X \rightarrow R - XY$ if $X \rightarrow R - XY$ in $\mathbf{\Sigma}^+$.

Lemma 6.2 *A set $\mathbf{\Sigma}$ of FDs and MVDs contains at least one pure MVD iff $\mathbf{\Sigma}$ is not equivalent to a set of FDs.*

Proof.
If: Suppose to the contrary that $\mathbf{\Sigma}$ is not equivalent to a set of FDs but doesn't contain a pure MVD. Then, as noted previously, if every non pure MVD $X \rightarrow\rightarrow Y \in \mathbf{\Sigma}$ is replaced by either $X \rightarrow Y$ or $X \rightarrow R - XY$, an equivalent set of FDs is obtained. This is a contradiction.
Only If: Suppose to the contrary that $\mathbf{\Sigma}$ contains a pure MVD $X \rightarrow\rightarrow Y$ but $\mathbf{\Sigma} \equiv \mathbf{\Sigma_f}$ where $\mathbf{\Sigma_f}$ is a set of FDs. By definition, this means that $X \rightarrow\rightarrow Y$ is implied by $\mathbf{\Sigma_f}$. Then, by Theorem 7.2 in [22], there exists either $X \rightarrow Y$ or $X \rightarrow R - XY \in \mathbf{\Sigma_f^+}$ and, by definition of equivalence, these FDs are also in $\mathbf{\Sigma}^+$ thus contradicting the assumption that $X \rightarrow\rightarrow Y$ is pure. □

In order to establish the main result of this section, we firstly derive additional preliminary lemmas.

Lemma 6.3 *If $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$, then the tableau T^*, where $T^* = chase_{\mathbf{\Sigma}}$ (T_X), consists of two or more rows and all rows are identical on X.*

Proof. Let T_X, ω_d, ω_X, T^*, ω_d^* and ω_x^* be as defined in Sect. 2. Firstly, T^* must consist of more than one row since otherwise one derives from Lemma 2.2 the contradiction that $X \in \mathrm{SK}(R, \mathbf{\Sigma})$. Secondly, we claim that for each attribute $A \in X, T^*[A]$ consists of a single dv and so all rows in T^* are identical on X. This follows from an inductive argument. Initially, by definition of T_X, each column in X contains a single dv and so the property holds. Then, let T' represent the tableau at any stage of the chase and assume inductively that the property is true. If a J-rule is applied to T' to produce a new row ω', then by definition of the J-rule, for each attribute $B \in R$ there is a row ω in T' such that $\omega'[B] = \omega[B]$. So, by the induction hypothesis, for every attribute $A \in X, \omega'[A]$ will contain the same dv as $T'[A]$ and the hypothesis is again true. Alternatively, if an F-rule is applied then the dv in each of the columns in X will remain unchanged since the F-rule does not change dv's. □

We now use this lemma to derive a result which will be used later to construct a relation with a deletion anomaly.

T*

	X	Y	Z
	.	.	.
ω_1:	x	y_1	z_1
ω_2:	x	y_2	z_2
ω_3:	x	y_1	z_2
ω_4:	x	y_2	z_1
	.	.	.

Fig. 7. Structure of tableau T^*

Lemma 6.4 *If $X \to\to Y$ is a pure MVD in $\mathbf{\Sigma}$, there exists a relation $r \in \mathrm{SAT}(\mathbf{\Sigma})$ which contains at least 4 distinct tuples ω_1, ω_2, ω_3 and ω_4 such that $\omega_1[X] = \omega_2[X] = \omega_3[X] = \omega_4[X], \omega_1[Y] = \omega_2[Y], \omega_2[Y] = \omega_3[Y], \omega_3[Y] = \omega_4[Y], \omega_1[Z] = \omega_3[Z], \omega_3[Z] = \omega_2[Z], \omega_2[Z] = \omega_4[Z]$ (where $Z = R - XY$).*

Proof. Form the tableau T_X as described in Sect. 2 and let $T^* = chase_{\mathbf{\Sigma}}(T_X)$. The claim is that T^* satisfies the conditions of the theorem from which it follows trivially that so does $\rho(T^*)$ for any one-to-one valuation ρ. The desired property of T^* can perhaps be more easily illustrated in Fig. 7.

From Lemma 6.3, T^* consists of more than one row and every row is identical on X. From Lemma 2.2, one row in T^* is the row ω_d which contains only dv's. For notational convenience, relabel ω_d as ω_1. Next, $X \to Y \notin \mathbf{\Sigma}^+$ since $X \to\to Y$ is pure and so, by Lemma 2.2, there must be at least one row in T^*, which we label as ω_2, which contains a ndv in a Y-column and so $\omega_2[Y] \neq \omega_1[Y]$.

Suppose firstly that $\omega_1[Z] \neq \omega_2[Z]$. By Lemma 2.1, T^* satisfies $X \to\to Y$ and so there is a row ω_3 with $\omega_3[X] = \omega_1[X]$, $\omega_3[Y] = \omega_1[Y]$ and $\omega_3[Z] = \omega_2[Z]$ and a row ω_4 with $\omega_4[X] = \omega_1[X]$, $\omega_4[Y] = \omega_2[Y]$ and $\omega_4[Z] = \omega_1[Z]$. These conditions also imply that ω_1, ω_2, ω_3 and ω_4 are distinct and so satisfy the conditions of the theorem.

Alternatively, suppose that $\omega_2[Z] = \omega_1[Z]$. Because by Lemma 2.1 ω_1 contains only dv's, $\omega_2[Z]$ contains only dv's in the Z-columns. Then, since $X \to Z \notin \mathbf{\Sigma}^+$ because $X \to\to Y$ is pure, there is a row $\omega_3 \in T^*$ containing a ndv in a Z-column and so $\omega_3[Z] \neq \omega_2[Z]$ and $\omega_3[Z] \neq \omega_1[Z]$ and hence ω_3 must be distinct from ω_2 and ω_1. There are then three subcases to be considered.

(a) $\omega_3[Y] = \omega_1[Y]$. Since by Lemma 2.1 $T^* \in \mathrm{SAT}(\mathbf{\Sigma})$, there must be a row ω_4 in T^* with $\omega_4[Z] = \omega_3[Z]$ and $\omega_4[Y] = \omega_2[Y]$. These conditions

also imply that ω_4 is distinct and so $\omega_1, \omega_4, \omega_3$ and ω_2 satisfy the requirements of the lemma.

(b) $\omega_3[Y] = \omega_2[Y]$. Again, since $T^* \in \mathrm{SAT}(\mathbf{\Sigma})$, there is a distinct tuple ω_4 with $\omega_4[Y] = \omega_1[Y]$ and $\omega_4[Z] = \omega_3[Z]$. Again these conditions imply that ω_4 is distinct and so ω_1, ω_3, ω_4 and ω_2 satisfy the requirements of the lemma.

(c) $\omega_3[Y] \neq \omega_2[Y]$ and $\omega_3[Y] \neq \omega_1[Y]$. Again, to satisfy $X \rightarrow\rightarrow Y$, there is a row ω_4 with $\omega_4[Y] = \omega_1[Y]$ and $\omega_4[Z] = \omega_3[Z]$ and a row ω_5 with $\omega_5[Y] = \omega_2[Y]$ and $\omega_5[Z] = \omega_3[Z]$. Then ω_1, ω_5, ω_4 and ω_2 satisfy the requirements of the lemma. □

This lemma is now used to derive the main result of this section which shows that the condition that the set of dependencies contain only FDs is equivalent to KDNF.

Theorem 6.1 *$(R, \mathbf{\Sigma})$ is in KDNF iff $\mathbf{\Sigma}$ is equivalent to a set of FDs.*

Proof.
If: Follows immediately from the facts that a KDA is cover insensitive and that there can be no KDA when the only dependencies are FDs.
Only if: We shall show the contrapositive that if $\mathbf{\Sigma}$ is not equivalent to a set of FDs then there exists r with a KDA. By Lemma 6.2, there is a pure MVD $X \rightarrow\rightarrow Y$ in $\mathbf{\Sigma}$ and by Lemma 6.4, there exists $r \in \mathrm{SAT}(\mathbf{\Sigma})$ of at least four tuples with the properties specified. Relation r has a KDA since deleting any of the four specified tuples results in $X \rightarrow\rightarrow Y$ being violated, but the new relation is in SAT($\mathbf{\Sigma}_k$) since $r \in \mathrm{SAT}(\mathbf{\Sigma})$. □

7 KMNF$_1$, KMNF$_2$ and KMNF$_3$ and 4NF

In this section we derive results on the relationship between 4NF and the key-based semantic normal forms KMNF$_1$, KMNF$_2$ and KMNF$_3$. The two cases of whether the constraints contain only FDs or both FDs and MVDs are treated separately.

7.1 The FD case

The main result we establish is that BCNF is equivalent to KMNF$_1$, KMNF$_2$ and KMNF$_3$. Firstly, we state some elementary lemmas whose proofs are omitted since they involve only simple applications of the inference rules.

Lemma 7.1 *If $X \rightarrow A \in \mathbf{\Sigma}$ and $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$ then $XA \notin \mathrm{SK}(R, \mathbf{\Sigma})$.*

Lemma 7.2 *If $K \in \mathrm{CK}(R, \mathbf{\Sigma})$ then there is no nontrivial $X \rightarrow A \in \mathbf{\Sigma}^+$ such that $XA \subseteq K$.*

Lemma 7.3 *If* $X \to A \in \mathbf{\Sigma}$ *and* $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$ *then for all* $K \in \mathrm{CK}(R, \mathbf{\Sigma})$*,* $K - X^+ \neq \emptyset$*.*

The next lemma is needed for the construction of counter-examples in the proofs of the main theorem concerning KMNF$_3$ and BCNF.

Lemma 7.4 *Let* $\mathbf{\Sigma}$ *be a reduced set of FDs and suppose* $X \to A \in \mathbf{\Sigma}$ *and* $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$*. Also, let* V *and* Y *be subsets of* X *such that* $X = V \cup Y$*,* $V \cap Y = \emptyset$*,* $V \neq \emptyset$*,* $Y \neq \emptyset$*. Construct a relation* r *of two tuples,* t_1 *and* t_2*, such that* t_1 *and* t_2 *are identical on* V^+ *and different elsewhere. Then* r *has the following properties:*

(a) $r \in \mathrm{SAT}(\mathbf{\Sigma})$*;*
(b) $t_1[V] = t_2[V]$*;*
(c) $t_1[Y] \neq t_2[Y]$*;*
(d) $t_1[A] \neq t_2[A]$.

Proof. Properties (a) and (b) follow from the fact that $V \notin \mathrm{SK}(R, \mathbf{\Sigma})$ (since $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$) and a result on two tuple relations (Theorem 4.1 in [22]). To establish (c), assume to the contrary that $t_1[Y] = t_2[Y]$. By definition of r, $V \to Y \in \mathbf{\Sigma}^+$ and applying the inference rules shows that $V \to A \in \mathbf{\Sigma}^+$ contradicting the fact that $\mathbf{\Sigma}$ is reduced since $V \subset X$. Similarly, (d) holds since otherwise the reduced assumption is again violated by replacing $X \to A$ by $V \to A$. □

These results are now used to establish the first main theorem of this section that BCNF is equivalent to KMNF$_3$. The proof is rather technical, so we provide firstly a brief sketch of the 'if' part of the proof, i.e. that KMNF$_3$ implies BCNF (the 'only if' part is immediate). We do this by showing the contrapositive that if $(R, \mathbf{\Sigma})$ is not in BCNF then one can always construct a two tuple relation which has a KMA$_3$ and so $(R, \mathbf{\Sigma})$ is not in KMNF$_3$. The construction uses a result that if $(R, \mathbf{\Sigma})$ is not in BCNF then there exists an FD $X \to A \in \mathbf{\Sigma}$ such that X is not a superkey and then, depending on the possible inclusion relationships between XA and the primary key K_p, assigns particular values to X and A in a two tuple relation such that the resulting relation has a KMA$_3$.

Theorem 7.1 *If* $\mathbf{\Sigma}$ *contains only FDs then* $(R, \mathbf{\Sigma})$ *is in BCNF iff it is in KMNF*$_3$*.*

Proof.
Only if: As for Theorem 5.1.
If: We will establish the contrapositive that if $(R, \mathbf{\Sigma})$ is not in BCNF then there exists r with a KMA$_3$. As $(R, \mathbf{\Sigma})$ is not in BCNF there is a nontrivial $X \to A \in \mathbf{\Sigma}$ where $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$ (Theorem 12.7 in [38]). However, every set of FDs has a reduced cover and so without loss of generality $\mathbf{\Sigma}$ is

assumed to be reduced and $X \rightarrow A$ an FD in $\mathbf{\Sigma}$ which violates BCNF. The proof is divided into three exhaustive cases.

(a) $A \notin K_p$ where K_p is the primary key of R. By Lemma 7.1, $XA \notin \mathrm{SK}(R, \mathbf{\Sigma})$ since $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$ and using the result mentioned in the previous proof, a relation r of two tuples, t_1 and t_2, which are identical on $(XA)^+$ and different elsewhere is in SAT($\mathbf{\Sigma}$). Let t^* be the tuple obtained by modifying $t_2[A]$ so that $t^*[A] \neq t_1[A]$. The claim is that r has a KMA$_3$ when t_2 is updated to t^*. Condition (i) is satisfied because $r \in \mathrm{SAT}(\mathbf{\Sigma})$, (ii) holds since $t_2[K_p] = t^*[K_p]$ (as $A \notin K_p$) and $t^*[K_p] \neq t_1[A]$ because $r \in \mathrm{SAT}(\mathbf{\Sigma})$, (iii') holds because $A \notin K_p$ and (iv) is satisfied because $t_1[X] = t^*[X]$ and $t_1[A] \neq t^*[A]$.

(b) $A \in K_p$ and $X \cap K_p = \emptyset$. Firstly, $X \neq \emptyset$ since otherwise applying the inference rules shows that $K_p - A \in$ SK($R, \mathbf{\Sigma}$) contradicting the fact that $K_p \in \mathrm{CK}(R, \mathbf{\Sigma})$. Next we claim that if $Y \rightarrow B \in \mathbf{\Sigma}$ and $Y \neq \emptyset$ then $Y \cap$ NS($\mathbf{\Sigma}$) $= \emptyset$. To verify this, suppose to the contrary $\emptyset \rightarrow C \in \mathbf{\Sigma}$ and $C \in Y$. An application of the inference rules shows that $Y \rightarrow B$ can be replaced by $Y - C \rightarrow B$ while maintaining equivalence and so contradicts the assumption that $\mathbf{\Sigma}$ is reduced. Construct then a relation r of two tuples, t_1 and t_2, as follows. For all $C \in$ NS($\mathbf{\Sigma}$), set $t_1[C] = t_2[C]$ and for all other $C \in R$ set $t_1[C] \neq t_2[C]$. We note that since $X \neq \emptyset$, $X \cap$ NS($\mathbf{\Sigma}$) $= \emptyset$ and so r is actually a relation, i.e. t_1 and t_2 are not duplicates. Let t^* be the tuple such that $t^*[X] = t_1[X]$ and $t^*[R - X] = t_2[R - X]$. The claim is that r has a KMA$_3$ when t_2 is changed to t^*. To satisfy (i) of a KMA$_3$, we show that $r \in \mathrm{SAT}(\mathbf{\Sigma})$. Any FD of the form $\emptyset \rightarrow C$ is satisfied since $t_1[C] = t_2[C]$ and any FD $Y \rightarrow C$ where $Y \neq \emptyset$ is satisfied since $t_1[Y] \neq t_2[Y]$ because $Y \cap$ NS($\mathbf{\Sigma}$) $= \emptyset$. To verify (ii), firstly $K \cap$ NS($\mathbf{\Sigma}$) $= \emptyset$ for all $K \in \mathrm{CK}(R, \mathbf{\Sigma})$ since otherwise an application of the inference rules derives the contradiction that $K - B \in$ SK($R, \mathbf{\Sigma}$) where $B \in K \cap$ NS($\mathbf{\Sigma}$). By Lemma 7.3, $K - X \neq \emptyset$ for all $K \in \mathrm{CK}(R, \mathbf{\Sigma})$ since $X \notin$ SK($R, \mathbf{\Sigma}$) and combining this with $K \cap$ NS($\mathbf{\Sigma}$) $= \emptyset$ shows that $K \cap R - X -$ NS($\mathbf{\Sigma}$) $\neq \emptyset$ and so (ii) holds by construction of r and t^*. Condition (iii') follows by definition of t^* and the fact that $X \cap K_p = \emptyset$. Condition (iv) follows by definition of t^* and the fact that $A \notin$ NS($\mathbf{\Sigma}$) since otherwise the reduced assumption is violated by replacing $X \rightarrow A$ by $\emptyset \rightarrow A$.

(c) $A \in K_p$ and $X \cap K_p \neq \emptyset$. Let $V = X \cap K_p$ and $Y = X - K_p$ and so $X = V \cup Y$, $V \cap Y = \emptyset$. Also, $V \neq \emptyset$ by assumption and $Y \neq \emptyset$ since otherwise $XA \subseteq K_p$, since $A \in K_p$, and so Lemma 7.2 implies a contradiction. By Lemma 7.4, the relation r of two tuples, t_1 and t_2, which are identical on V^+ and different elsewhere has the properties given by the lemma. Let t^* be the tuple defined by $t^*[Y] = t_1[Y]$ and $t^*[R - Y] = t_2[R - Y]$. The claim is that r has a KMA$_3$ when t_2 is replaced by t^*.

Condition (i) follows from (a) of Lemma 7.4. To verify (ii), $V^+YA \subseteq X^+$ from the properties of the closure and then, by Lemma 7.3, $K - V^+YA \neq \emptyset$ for all $K \in \text{CK}(R, \mathbf{\Sigma})$ and so (ii) holds by definition of t_1, t_2 and t^*. Condition (iii') holds since t^* and t_2 differ only on attributes in $X - K_p$ and (iv) follows from Lemma 7.4 and the construction of t^*. □

A corollary of the previous result is the following which shows that BCNF, KMNF$_1$ and KMNF$_2$ are also equivalent.

Corollary 7.1 *If $\mathbf{\Sigma}$ contains only FDs, then BCNF, KMNF$_1$, KMNF$_2$ and KMNF$_3$ are equivalent.*

Proof.
$KMNF_1 \Rightarrow KMNF_2 \Rightarrow KMNF_3$. Immediate from the definitions.
$KMNF_3 \Rightarrow BCNF$: Immediate from the theorem.
$BCNF \Rightarrow KMNF_1$: As for Theorem 5.1. □

7.2 The FD and MVD case

In this section we generalise the results of the previous section by showing that KMNF$_1$, KMNF$_2$, KMNF$_3$ and 4NF are again equivalent when MVDs are present in the set of constraints provided that there is also at least one FD. First, a preliminary lemma that will be extensively used in later sections is established.

Lemma 7.5 *Let $X \notin \text{SK}(R, \mathbf{\Sigma})$. If $W \in \text{DEP}(X)$ and $W \cap X^+ = \emptyset$ then $K \cap W \neq \emptyset$ for all $K \in \text{CK}(R, \mathbf{\Sigma})$.*

Proof. By Theorem 3.1, any two tuple relation r for which the tuples are the same except for those attributes in W is in SAT($\mathbf{\Sigma}$). Hence $r \in \text{SAT}(\mathbf{\Sigma}_k)$ and so $K \cap W \neq \emptyset$ because the two tuples must be distinct on every candidate key. □

We now present the main result of this section which shows that 4NF and KMNF$_3$ are equivalent. As before, we first briefly outline the main idea of the 'If' part of the proof for the benefit of the reader (the 'Only If' part is again immediate). The 'If' part is established by showing the contrapositive that if ($R, \mathbf{\Sigma}$) is not in 4NF then one can construct a relation which has a KMA$_3$ and so ($R, \mathbf{\Sigma}$) is not in KMNF$_3$. The construction is based on the result that there exists an MVD or FD in $\mathbf{\Sigma}$ where the lhs is not a superkey if ($R, \mathbf{\Sigma}$) is not in 4NF and then constructing a relation with specific values for the attributes in the dependency so that resulting relation has a KMA$_3$. The technical complexity and length of the proof arises from the fact that different techniques are needed for constructing the relation depending on whether the dependencies are standard or not and on the relationship between the attributes in the lhs of the dependency and the primary key

Theorem 7.2 *If $\mathbf{\Sigma}$ is a set of FDs and MVDs containing at least one nontrivial FD, then $(R, \mathbf{\Sigma})$ is in 4NF iff it is KMNF$_3$.*

Proof.
Only if: As for Theorem 5.1.
If: We shall show the contrapositive that if $(R, \mathbf{\Sigma})$ is not in 4NF then it is not in KMNF$_3$. Because 4NF is cover insensitive, $\mathbf{\Sigma}$ is assumed to be pure and reduced without loss of generality. Then since $(R, \mathbf{\Sigma})$ is not in 4NF, it follows from Lemma 4.3 in [36] that there exists a dependency in $\mathbf{\Sigma}$ which violates 4NF. In the proof, the complementary rule (rule A4) is needed and so $X \rightarrow\rightarrow Y$ will often be written as $X \rightarrow\rightarrow Y|Z$ where $Z = R - XY$. We shall now consider several cases separately.

(a) There exist nonstandard dependencies in $\mathbf{\Sigma}$. The FD and MVD cases are considered separately.

(a.1) There exists a nonstandard FD $\emptyset \rightarrow B \in \mathbf{\Sigma}$. B cannot be a superkey or else an application of the inference rules shows that $\emptyset$ is a superkey and so the lhs of every dependency in $\mathbf{\Sigma}$ is a superkey thus contradicting the assumption that $(R, \mathbf{\Sigma})$ is not in 4NF. Also, $B \notin K$ for all $K \in \text{CK}(R, \mathbf{\Sigma})$ since otherwise the inference rules show the contradiction that $K - B \in \text{SK}(R, \mathbf{\Sigma})$. Construct then the tableau T_B as in Lemma 2.2, let $T^* = chase_{\mathbf{\Sigma}}(T_B)$ and let r be any one-to-one valuation of T_B. Then by Lemma 6.3, r contains at least two tuples and all tuples are identical on B. We then claim that r has a KMA$_3$ when some B value in r is changed to a new value. Condition (i) follows from Lemma 2.1; (ii) holds because $r \in \text{SAT}(\mathbf{\Sigma}_k)$ and B is not prime; (iii) holds because B is not prime and (iv) holds because the new relation violates $\emptyset \rightarrow B$.

(a.2) There exists a nonstandard MVD $\emptyset \rightarrow\rightarrow Y|Z \in \mathbf{\Sigma}$: Let K_p be the primary key of R. Since there is at least one FD in $\mathbf{\Sigma}$, $R \notin \text{CK}(R, \mathbf{\Sigma})$ and so there is at least one attribute B disjoint from K_p. We then claim that B cannot be a superkey. To verify this, suppose to the contrary that B is a superkey and that $B \in Y$ (by symmetry the same argument holds if $B \in Z$). Then since B is a superkey, $B \rightarrow Z \in \mathbf{\Sigma}^+$ and so, by inference rule A9, $\emptyset \rightarrow Z \in \mathbf{\Sigma}^+$ contradicting the assumption that $\mathbf{\Sigma}$ is pure.

As before, construct the tableau T_B as in Lemma 2.2, let $T^* = chase_{\mathbf{\Sigma}}$ (T_B) and let r be any one-to-one valuation of T_B. Then by Lemma 6.3, r contains at least two tuples and all tuples are identical on B. Next we claim that there are two tuples, t_1 and t_2, such that $t_1[Z] \neq t_2[Z]$. This follows since if all rows are equal on Z, then they must be equal to ω_d on Z since dv's are not changed during the chase and so, by Lemma 2.2, $B \rightarrow Z \in \mathbf{\Sigma}^+$ which implies, by inference rule A9, that $\emptyset \rightarrow Z$ which again contradicts the pure assumption. We then claim that r has a KMA$_3$ when $t_2[B]$ is changed to a value b^* disjoint from the values in r. Condition (i) follows from Lemma 2.1; (ii) holds because for any $K \in \text{CK}(R, \mathbf{\Sigma})$, if $B \notin K$ then the new relation is in SAT($\mathbf{\Sigma}_k$) because $r \in \mathrm{SAT}(\mathbf{\Sigma}_k)$ and if $B \in K$, then the new relation is in SAT($\mathbf{\Sigma}_k$) because b^* is disjoint from the values in r; (iii) holds because $B \notin K_p$ and (iv) holds since $t_1[Z] \neq t_2[Z]$ and the property of b^*.

(b) All dependencies in $\mathbf{\Sigma}$ *are standard.* If the only violations of 4NF are caused by FDs, there can be no MVDs in $\mathbf{\Sigma}$ because if there is, the lhs of the MVD must be a superkey contradicting the fact that $\mathbf{\Sigma}$ is pure and Theorem 7.1 then shows that $(R, \mathbf{\Sigma})$ is not in KMNF$_3$. Alternatively, suppose there is $X \rightarrow\!\!\!\rightarrow Y \in \mathbf{\Sigma}$ violating 4NF, i.e. the MVD is nontrivial and $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$. We shall now show that there exists a relation which has a KMA$_3$. Split each of the sets X, Y and Z into a set which intersects with K_p and a set which is disjoint from K_p and so $X \rightarrow\!\!\!\rightarrow Y|Z$ is written as $X'X_k \rightarrow\!\!\!\rightarrow Y'Y_k|Z'Z_k$ where $X = X'X_k$, $Y = Y'Y_k$, $Z = Z'Z_k$ and $X'Y'Z' \cap K_p = \emptyset$. Again, several subcases are considered.

(b.1) $X_k = \emptyset$. Define a relation r of two tuples, t_1 and t_2, such that t_1 and t_2 are different on every attribute. Obviously, $r \in \mathrm{SAT}(\mathbf{\Sigma})$ since $\mathbf{\Sigma}$ contains only standard dependencies. Then define the tuple $t^* by t^*[X] = t_1[X]$, $t^*[R - X] = t_2[R - X]$. The claim is that r has a KMA$_3$ when t_2 is replaced by t^*. Condition (i) of a KMA$_3$ follows from the definition of r and (ii) holds follows from Lemma 7.3 and the fact that $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$. Also, $t^*[K_p] = t_2[K_p]$ since $X_k = \emptyset$ and so (iii') holds. Condition (iv) follows from $Z \neq \emptyset$ (since $X \rightarrow\!\!\!\rightarrow Y$ is nontrivial) and the definitions of r and t^*.

(b.2) $X_k \neq \emptyset$. We now break the proof up into several subcases.

(b.2.1.1) There doesn't exist a dependency in $\mathbf{\Sigma}$ *with a subset of* X_k *on the lhs*. It follows from this assumption that $X' \neq \emptyset$ since otherwise the MVD $X'X_k \rightarrow\!\!\!\rightarrow Y'Y_k|Z'Z$ could be written as $X_k \rightarrow\!\!\!\rightarrow Y'Y_k|Z'Z$. Construct r of two tuples, t_1 and t_2, such that $t_1[X_k] = t_2[X_k]$ and $t_1[B] \neq t_2[B]$ for all $B \in R - X_k$. Define t^* by $t^*[X] = t_1[X]$ and $t^*[R - X] = t_2[R - X]$. We claim that r has a KMA$_3$ when t_2 is replaced by t^*. Firstly, $r \in \mathrm{SAT}(\mathbf{\Sigma})$ because $t_1[B] \neq t_2[B]$ for all $B \in R - X_k$ and so the only dependency that r could violate is one with a subset of X_k on the lhs and by (b.2.1.1) this cannot occur. Next, by Lemma 7.3, $K - X \neq \emptyset$ for all $K \in \mathrm{CK}(R, \mathbf{\Sigma})$ and so by definition of r and t^*, $t^*[K] \neq t_1[K]$ and thus (ii) holds. Condition (iii') holds because t^* and t_2 differ only on X' and (iv) holds because t^* and t_1 agree on X yet differ on Y and Z.

(b.2.2) There exists a dependency in $\mathbf{\Sigma}$ *with a subset of* X_k *as the lhs*. In other words, there exists either $X'_k \rightarrow\!\!\!\rightarrow V$ or $X'_k \rightarrow A \in \mathbf{\Sigma}$ with $X'_k \subseteq X_k$. Consider firstly the MVD case.

(b.2.2.1) $X'_k \rightarrow\!\!\!\rightarrow V|U \in \mathbf{\Sigma}$. Write $X'_k \rightarrow\!\!\!\rightarrow V|U$ as $X'_k \rightarrow\!\!\!\rightarrow V'V_k|U'U_k$ where, as before, V_k and W_k are the intersection of K_p with V and U. Firstly, by property (ii) of DEP, U is equal to a union of elements of DEP(X'_k) and at least one of these, W, must be disjoint from $(X'_k)^+$ since otherwise $U \subseteq (X'_k)^+$ contradicting the assumption that $\mathbf{\Sigma}$ is pure.

We now construct a relation which has a KMA$_3$. Since, as noted earlier, $R \neq K_p$ and since $R = X'_k V' V_k U' U_k$, either $V' \neq \emptyset$ or $U' \neq \emptyset$. Assume that $V' \neq \emptyset$. By symmetry, the same argument applies if $U' \neq \emptyset$. Construct r of two tuples, t_1 and t_2, which are identical on all attributes except those in W and modify r by replacing t_2 by the tuple t^* defined by $t^*[V'] \neq t_1[V']$ and t^* and t_2 are identical elsewhere. The claim is that r has a KMA$_3$ when t_2 is replaced by t^*. Condition (i) follows from Theorem 3.1, (ii) holds because of Lemma 7.5 and the fact that only the V' values in t_2 are modified and (iii') holds for the same reason. Finally, (iv) is valid since the tuples t^* and t_1 agree on X'_k but differ on U and V.

(b.2.2.2) $X'_k \to A \in \mathbf{\Sigma}$. Firstly, by Lemma 7.2, $A \notin K_p$ because $X'_k \subseteq K_p$. Then $X'_k \notin \mathrm{SK}(R, \mathbf{\Sigma})$ since $X'_k \subseteq K_p$ and so there exists $W \in \mathrm{DEP}(X'_k)$ such that $W \cap (X'_k)^+ = \emptyset$. Choose any such W and construct a relation r of two tuples, t_1 and t_2, where $t_1[R-W] = t_2[R-W]$ and $t_1[B] \neq t_2[B]$ for all $B \in W$. Modify r by replacing t_2 with the tuple t^* defined by $t^*[A] \neq t_1[A]$ and $t^*[R-A] = t_1[R-A]$. The claim is that r has a KMA$_3$ when t_2 is replaced by t^*. Condition (i) follows from Theorem 3.1. The compatibility condition (ii) holds since by Lemma 7.5, t_1 and t_2 differ on $K' \cap W$ for all $K' \in \mathrm{CK}(R, \mathbf{\Sigma})$ and so t^* and t_1 differ on K' since $t^*[W] = t_2[W]$. Condition (iii') holds because $t^*[A] \neq t_1[A]$ and $A \notin K$, while (iv) holds because t^* and t_1 agree on X'_k yet differ on A. □

A simple corollary of the previous theorem is the following important result that 4NF is also equivalent to KMNF$_1$ and KMNF$_2$.

Theorem 7.3 *If* $\mathbf{\Sigma}$ *contains at least one FD then 4NF is equivalent to KMNF*$_1$*, KMNF*$_2$ *and KMNF*$_3$*.*

Proof. As for Corollary 7.1. □

We note that the requirement in Theorem 7.3 that the set of dependencies contain at least one FD is necessary for the equivalence of 4NF and KMNF$_3$. To verify this, if there are only MVDs in the set of dependencies then the only candidate key is R since only trivial FDs are implied by a set of MVDs [22]. So every attribute in R is prime and it follows from the definitions of modification anomalies that $(R, \mathbf{\Sigma})$ is in KMNF$_3$. However, any nontrivial MVD violates the 4NF condition since R is the only candidate key and so every relation scheme with only MVDs in the set of dependencies is in KMNF$_3$ yet not in 4NF.

8 KMNF$_4$ and 4NF

8.1 The FD case

In this section we show that the syntactic normal form equivalent to KMNF$_4$ is a new normal form, what we refer to as *prime attribute normal form*

(*PANF*). This new normal form is stronger than 3NF yet weaker than BCNF. We also show that PANF is not equivalent to EKNF, another normal form that lies between 3NF and BCNF, and that the simplest version of the well known synthesis algorithm generates schemes which are in PANF.

Definition 8.1 *Let $\mathbf{\Sigma}$ be a reduced set of FDs. $(R, \mathbf{\Sigma})$ is in* prime attribute normal form *(PANF) if for every FD $X \to A \in \mathbf{\Sigma}$, either $X \in \mathrm{CK}(R, \mathbf{\Sigma})$ or XA contains only prime attributes.*

It is easily seen from this definition that PANF lies between 3NF and BCNF and the following examples show that this inclusion is strict.

Example 8.1 Let $R = \{A, B, C, D\}$ and $\mathbf{\Sigma} = \{AB \to C, CD \to AB, BD \to A\}$. The candidate keys are CD and BD and so $(R, \mathbf{\Sigma})$ is in 3NF since C is a prime attribute, but $(R, \mathbf{\Sigma})$ is not in PANF since A is not a prime attribute.

Example 8.2 Let $R = \{A, B, C, D\}$ and $\mathbf{\Sigma} = \{AB \to C, AB \to D, C \to B\}$. It can be easily verified that the only candidate keys are AB and AC and so $(R, \mathbf{\Sigma})$ is in PANF since the lhs of the first two FDs is a candidate key and, in the FD $C \to B$, both C and B are prime. However, $C \to B$ violates BCNF.

We now show the equivalence of PANF and KMNF$_4$.

Theorem 8.1 *If $\mathbf{\Sigma}$ contains only FDs, then $(R, \mathbf{\Sigma})$ is in KMNF$_4$ iff it is in PANF.*

Proof.
If: Suppose to the contrary that $(R, \mathbf{\Sigma})$ is in PANF yet is not in KMNF$_4$. Then there exists r and $t \in r$ such that an FD $X \to A$ is violated when t is modified to t^*. By definition of a KMA$_4$ the new relation is in SAT($\mathbf{\Sigma}_k$) and so $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$ and hence, by the definition of PANF, XA contains only prime attributes. This is a contradiction since, by definition of a KMA$_4$, $t[XA]$ is unchanged during the update and so $X \to A$ cannot be violated.

Only if: The contrapositive that if $(R, \mathbf{\Sigma})$ is not in PANF then it is not in KMNF$_4$ will be established. If $(R, \mathbf{\Sigma})$ is not in PANF there exists $X \to A \in \mathbf{\Sigma}$ such that $X \notin \mathrm{SK}(R, \mathbf{\Sigma})$ and XA contains a nonprime attribute. If A is nonprime then the same construction used in (a) of the proof of Theorem 7.1 shows that $(R, \mathbf{\Sigma})$ is not in KMNF$_4$. Alternatively if A is prime then X must contain nonprime attributes. If X contains only nonprime attributes then the same construction used in (b) of Theorem 7.1 shows that $(R, \mathbf{\Sigma})$ is not in KMNF$_4$. Alternatively, suppose X contains both prime and nonprime attributes where V is the set of prime attributes and Y is the set of nonprime attributes. The conditions of Lemma 7.4 are satisfied and the same construction used in (c) of Theorem 7.1 shows that $(R, \mathbf{\Sigma})$ is not in KMNF$_4$. □

We note that since KMNF_4 is cover insensitive, then a corollary of this theorem is that PANF is also cover insensitive, a result that is not immediate from the definition of PANF.

Another normal form, namely *elementary key normal form* (*EKNF*) [41], also lies between 3NF and BCNF. We now demonstrate that EKNF and PANF are not comparable. Firstly we recall the definition of EKNF.

Definition 8.2 *Suppose that* $\mathbf{\Sigma}$ *contains only FDs and let* $X \rightarrow A$ *an FD in* $\mathbf{\Sigma}$*. Then* $X \rightarrow A$ *is* elementary *if there doesn't exist a nontrivial FD* $X' \rightarrow A \in \mathbf{\Sigma}^+$ *such that* $X' \subset X$.

We note that it follows from this definition that every FD in a reduced set of FDs is elementary.

Definition 8.3 *Let* R *be a relation scheme,* $\mathbf{\Sigma}$ *a set of FDs and* $K \in \text{CK}(R, \mathbf{\Sigma})$*. Then* K *is an* elementary key *if for some attribute* $A, K \rightarrow A$ *is an elementary FD. An attribute which belongs to some elementary key is called an* elementary key attribute.

Definition 8.4 *Let* R *be a relation scheme and* $\mathbf{\Sigma}$ *a set of FDs.* $(R, \mathbf{\Sigma})$ *is in* elementary key normal form *(EKNF) if for every nontrivial FD* $X \rightarrow A \in \mathbf{\Sigma}$*, either* $X \in \text{CK}(R, \mathbf{\Sigma})$ *or* A *is an elementary key attribute.*

The first example, taken from [41], shows that a scheme can be in PANF but not in EKNF, while the second example demonstrates the converse.

Example 8.3 Let $R = \{A, B, C\}$ and $\mathbf{\Sigma} = \{A \rightarrow B, B \rightarrow A\}$. The candidate keys are AC and BC. Neither candidate key is elementary because neither $AC \rightarrow B$ nor $BC \rightarrow A$ is an elementary FD and so there are no elementary key attributes. Hence $(R, \mathbf{\Sigma})$ is not in EKNF since if one considers the FD $A \rightarrow B$ then $A \notin \text{CK}(R, \mathbf{\Sigma})$ and B is not an elementary key attribute. However, $(R, \mathbf{\Sigma})$ is in PANF since every attribute in R is prime.

Example 8.4 As in Example 8.1, let $R = \{A, B, C, D\}$ and $\mathbf{\Sigma} = \{AB \rightarrow C, CD \rightarrow AB, BD \rightarrow A\}$. Both the candidate keys BD and CD are elementary because of the FDs $CD \rightarrow AB$, $BD \rightarrow A$ and so B, C and D are elementary key attributes. Thus $(R, \mathbf{\Sigma})$ is EKNF since both CD and BC are superkeys and C is an elementary key attribute. However, as noted previously, $(R, \mathbf{\Sigma})$ is not in PANF because A is not prime.

We now address the problem of generating relation schemes which are in PANF. We start with the simplest version [32] of the synthesis algorithm [6] for generating 3NF schemes. Other versions of the algorithm combine schemes which result from dependencies having the same, or equivalent, left-hand sides. We conjecture that Theorem 8.2 is also valid for these alternatives versions.

```
ALGORITHM 8.1. A SYNTHESIS ALGORITHM FOR
ACHIEVING 3NF.
Input: A relational scheme $R$ and a reduced
set $\mathbf{\Sigma}$ of FDs.
Output: A dependency preserving, lossless
decomposition of $R$ into 3NF.
Method:
For each FD $X \to A \in \mathbf{\Sigma}$, create the scheme $XA$.
If there is no scheme which contains a
candidate key $K$ then create an extra scheme
which contains $K$ alone.
```

We now show that the relation schemes generated by Algorithm 8.1 are in PANF.

Theorem 8.2 *Each of the relation schemes generated by Algorithm 8.1 is in PANF.*

Proof. If the relation scheme is a candidate key then the result is immediate, so alternatively assume that it is the scheme $R' = XA$ which corresponds to the FD $X \to A \in \mathbf{\Sigma}$. From the properties of projected FDs [22, 32] any FD which holds in R' must also hold in R. We show firstly that X is a candidate key in R'. $X \in \mathrm{SK}(R', \mathbf{\Sigma})$ because $R' = XA$. Also, $X \in \mathrm{CK}(R', \mathbf{\Sigma})$ since if not there exists $K \in \mathrm{CK}(R', \mathbf{\Sigma})$ with $K \subset X$, and so $K \to A \in \mathbf{\Sigma}^+$ which contradicts the assumption that $X \to A$ is reduced. We now show that any other FD $Y \to B$ which holds in R' satisfies PANF. We divide the proof into the two cases $B = A$ and $B \neq A$.

(a) $B = A$. For this case, it follows that $Y = X$ since otherwise $Y \subset X$ and this violates the property that $X \to A$ is reduced and so $Y \to B$ satisfies PANF.

(b) $B \neq A$. Firstly, since $R' = XA$ then $B \in X$ and so, from Lemma 7.2, $Y \not\subset X$ and thus $Y = AX'$ where $X' \subset X$. X' is prime in R' since $X \in \mathrm{CK}(R', \mathbf{\Sigma})$ and so it remains to show that A is prime. Assume it is not. Since $AX' \to B$ and $R' = AX, AX - B \in \mathrm{SK}(R', \mathbf{\Sigma})$ and so there exists $K \in \mathrm{CK}(R', \mathbf{\Sigma})$ such that $K \subseteq AX - B$. Then since A is not prime in R', $K \subseteq X - B$ and so $K \subset X$ because $B \in X$ which contradicts the fact that $X \to A$ is reduced using the same argument as in the previous paragraph. □

8.2 The FD and MVD case

We now address the problem of deriving necessary and sufficient conditions for a relation scheme to be in KMNF_4 when both FDs and MVDs are present. We proved in the previous section that in the case where the set of constraints contains only FDs, KMNF_4 is a weaker condition than BCNF. The following example shows that similarly, in the case where the set of constraints includes both MVDs and FDs, 4NF is a stronger condition than is required for a relation scheme to be in KMNF_4.

Example 8.5 Let $R = \{G, S, T, H\}$ and $\boldsymbol{\Sigma} = \{G \rightarrow T, T \rightarrow G, G \rightarrow\rightarrow H\}$. It can be verified that the candidate keys are GSH and HST and that the MVD $G \rightarrow\rightarrow H$ is pure. $(R, \boldsymbol{\Sigma})$ is not in 4NF because none of the lhs is a superkey yet, as will be seen in the next theorem, it is not in KMNF_4 because every attribute is prime.

We now present the main theorem of this section which gives a necessary and sufficient condition for a relation scheme to have no KMA_4.

Theorem 8.3 *If $\boldsymbol{\Sigma}$ contains at least one pure MVD, then $(R, \boldsymbol{\Sigma})$ is in KMNF_4 iff every attribute in R is prime.*

Proof.
If: Consider any $r \in \text{SAT}(\boldsymbol{\Sigma})$. Since every attribute in R is prime, any modification which leaves the prime attributes of a tuple unchanged doesn't change the tuple and so r has no KMA_4 violation.
Only if: We shall establish the result by showing the contrapositive that if there is a nonprime attribute then there exists r with a KMA_4 anomaly. Firstly, since a KMA_4 is cover insensitive, then without loss of generality $\boldsymbol{\Sigma}$ is assumed to be pure and reduced. Also, if the original set of dependencies $\boldsymbol{\Sigma}$ contains a pure MVD then, by Lemma 6.2, a pure reduced cover for $\boldsymbol{\Sigma}$ must contain at least one MVD.

Firstly, if $\boldsymbol{\Sigma}$ contains a nonstandard dependency then the same constructions used in (a) of Theorem 7.3 shows that $(R, \boldsymbol{\Sigma})$ is not in KMNF_4 so we will assume that all dependencies are standard. Consider any $X \rightarrow\rightarrow Y|Z \in \boldsymbol{\Sigma}$. Because $\boldsymbol{\Sigma}$ is pure, $X \notin \text{SK}(R, \boldsymbol{\Sigma})$ since otherwise $X \rightarrow Y \in \boldsymbol{\Sigma}^+$ contradicting the pure assumption. Since $XYZ = R$ and by assumption R contains a nonprime attribute, either YZ contains a nonprime attribute or only X contains a nonprime attribute. We now consider each case in turn.

YZ contains a nonprime attribute. In this case, the same argument used in (b.2.2) of Theorem 7.3 shows that $(R, \boldsymbol{\Sigma})$ is not in KMNF_4.

Only X contains a nonprime attribute: Write X as $X'X_p$ where X' contains only nonprime attributes and X_p contains only prime attributes. We firstly show that we can restrict attention to the case where X' is a subset of the lhs of every MVD in $\boldsymbol{\Sigma}$. Otherwise, since X' contains all the nonprime attributes in R, this implies that either the rhs of an MVD, or its complement, contains nonprime attributes and so the previous argument shows that $(R, \boldsymbol{\Sigma})$ is not in KMNF_4.

We now want to show that there exists $W \in \text{DEP}(X_p)$ such that $W = X'YZ$. The proof of this assertion is presented as a separate lemma (Lemma

8.1). Construct then a relation r of two tuples, t_1 and t_2, such that $t_1[R - W] = t_2[R - W]$ and $t_1[B] \neq t_2[B]$ for all $B \in W$. Replace t_2 by t^* defined by $t^*[X'] = t_1[X']$ and $t^*[R - X'] = t_2[R - X']$. The claim is that r has a KMA$_4$ when t_2 is replaced by t^*. Condition (i) follows from Theorem 3.1. Condition (ii) holds because $r \in \mathrm{SAT}(\mathbf{\Sigma}_k)$ and t^* and t_2 are equal on prime attributes, which also implies condition (iii”). Finally, (iv) holds because $W \cap Y \neq \emptyset$ and $W \cap Z \neq \emptyset$ and so by definition of r and t^*, t_1 and t^* agree on X but differ on both Y and Z. □

In order to derive the lemma needed for the completion of the theorem we need the following algorithm for calculating DEP(X) [3, 27]. Our version is a simplification of that in [27] since the set of constraints is reduced and so the rhs of an FD contains a single attribute.

```
**ALGORITHM 8.2.** An algorithm for generating
DEP($X$)
**Input:** A relation scheme $R$, a reduced set of
FDs and MVDs $\mathbf{\Sigma}$, and a set of attributes
$X$.
**Output:** DEP($X$)
**Method:**
  **var**: $U'\ U'', V'\ V'', W$: sets of attributes;
     OLD_DEP, NEW_DEP: sets of sets of attributes;
1:   NEW_DEP $:= \{\{A\}|A \in X\} \cup \{R - X\}$;
     **repeat**
       OLD_DEP := NEW_DEP;
2:     **for each** $U \rightarrow\rightarrow V$ or $U \rightarrow V \in \mathbf{\Sigma}$ **do**
3:        $U'' := \cup\{W|(W \in \mathrm{NEW\_DEP})$ and $(W \cap U \neq \emptyset)\}$;
4:        $V'' := V - U''$;
5:        **if** $V'' \neq \emptyset$
          **then**
             **for each** $W \in$ NEW_DEP **do**
6:              **if** $(W \cap V" \neq \emptyset)$ **and** $(W \cap V'' \neq W)$
                 **then**
7:                  NEW_DEP := (NEW_DEP$-\{W\})\cup$
                                 $\{W \cap V'', W - V''\}$;
            **od**
       **od**
     **until** (NEW_DEP = OLD_DEP);
```

We also note that in this algorithm, if a set Z is in NEW_DEP at any iteration then $X \rightarrow\rightarrow Z \in \mathbf{\Sigma}^+$. We now use the algorithm establish the following lemma.

Lemma 8.1 *Let X' be a nonempty set of attributes and let $\mathbf{\Sigma}$ be a reduced set of FDs and MVDs such that the lhs of every MVD in $\mathbf{\Sigma}$ contains X'. Then, for any MVD $X'X \rightarrow\rightarrow Y|Z \in \mathbf{\Sigma}$, $DEP(X) = \{X_1, \ldots, X_n, X'YZ\}$ where $X = X_1 \ldots X_n$.*

Proof. Initially, because $R = X'XYZ$, NEW_DEP $= \{X_1, \ldots, X_n, X'YZ\}$ and we shall now prove that NEW_DEP never changes. We consider separately the two cases of whether the dependency tested at line 2 of Algorithm 8.1 is an MVD or an FD.

(a) The MVD case. By the assumption of the lemma, any MVD $U \rightarrow\rightarrow V \in \mathbf{\Sigma}$ can be written as $X'U' \rightarrow\rightarrow V$ where $U = X'U'$. Consider $V'' = V - U''$ defined at line 4. Obviously $X'YZ \subseteq U''$ since $X' \cap U = X'$ and so $V'' \cap X'YZ = \emptyset$ and thus the test at line 6 fails when $W = X'YZ$. The only other elements in NEW_DEP are the attributes in X and the second test at line 6 fails for these and so NEW_DEP remains unchanged.

(b) The FD case. Consider the effect on NEW_DEP if instead the FD $U \rightarrow V$ is applied at line 2. Since $\mathbf{\Sigma}$ is reduced, V consists of a single attribute. We shall break the proof up into the following subcases.

(b.1) $V \in X$. NEW_DEP is unaltered since V is already in NEW_DEP.

(b.2) $V \in X'$. We shall show that NEW_DEP is unchanged by assuming the contrary and deriving a contradiction. Since V is a single attribute, V is added to NEW_DEP at line 7 and so $X \rightarrow\rightarrow V \in \mathbf{\Sigma}^+$ and since the lhs and rhs of every MVD in $\mathbf{\Sigma}$ are disjoint, a simple application of the inference rules shows that $(X' - V)X \rightarrow\rightarrow Y \in \mathbf{\Sigma}^+$. This contradicts the assumption that $\mathbf{\Sigma}$ is reduced and so NEW_DEP remains unchanged.

(b.3) $V \in Y$. As before, assume to the contrary that NEW_DEP changes and thus V is added to NEW_DEP and so $X \rightarrow\rightarrow V \in \mathbf{\Sigma}^+$. If $V = Y$ then the assumption that every MVD is left-reduced is contradicted since $X' \neq \emptyset$, and if $V \subset Y$ the inference rules show that $X'X \rightarrow\rightarrow V \in \mathbf{\Sigma}^+$ contradicting the assumption that every MVD is right-reduced.

(b.4) $V \in Z$. Same argument used in (b.3). □

We note that in the case that $\mathbf{\Sigma}$ does not contain a pure MVD (so $\mathbf{\Sigma}$ is equivalent to a set of FDs), then the necessary and sufficient condition reduces to the one given in Theorem 8.1. We also note that the condition in Theorem 8.3 is not equivalent to the normal form PANF defined in Sect. 8.1 since the relation scheme in Example 8.2 is in PANF yet contains a nonprime attribute.

9 Fact-based modification anomalies and 4NF

In this section we consider again the relationship between 4NF and the maintenance of database integrity when tuples are modified, but under different assumptions to those used in Sect. 4. Instead of the tuple being the basic unit of information, in this section we use the *fact-based approach* which assumes that only certain subsets of a tuple, called facts, are the basic information units for retrieval and update. The fact-based has been widely used in research relating to database design and database semantics [5, 10, 39].

If facts are the basic information units, then it is desirable that the integrity of a relation is maintained when facts are updated and so a fact-based update anomaly is considered to occur when the update of a fact results in a violation of the integrity constraints. In general, an update to a relation can be either an insertion, a deletion or a modification of a tuple but in this section attention is restricted to the modification of tuples. This is because we believe that this case can be adequately handled without using null values, whereas a complete analysis of the insertion and deletion of facts require the use of null values and thus has been left as a topic for future investigation.

The final issue to be addressed before formally defining a fact-based modification anomaly is to choose the attribute sets for the set of facts. The approach adopted here, and elsewhere, is to use the attribute sets corresponding to the FDs and MVDs. However even in this approach there are still several choices. The first is to use only the FDs and MVDs in $\mathbf{\Sigma}$. The second is to recognise the symmetrical nature of MVDs and also include the complementary MVDs (Rule A4). The last is to include derived dependencies and so use any nontrivial dependency in $\mathbf{\Sigma}^+$. We allow for each of these possibilities and formally define three types of fact-base modification anomalies.

Definition 9.1 *A relation r has a* fact-based modification anomaly 1 *(FMA$_1$) w.r.t. a* reduced set $\mathbf{\Sigma}$ *of FDs and MVDs if there exists a tuple $t \in r$ and a tuple t^* defined over R such that:*

(i) $r \in \mathrm{SAT}(\mathbf{\Sigma})$*;*
(ii) t^* *is compatible with* $(r - \{t\})$ *(See Sect. 4);*
(iii) The set of attributes on which t and t^* differ is a subset of ATT(d) where $d \in \mathbf{\Sigma}$ and ATT(d) are the attributes in d;
(iv) $\{r - \{t\}\} \cup \{t^*\}$ *(SAT($\mathbf{\Sigma}$).*

Some observations on the above definition are appropriate at this point. A reduced set of dependencies is used in this definition because a dependency in such a set cannot have their lhs or rhs decomposed and so are irreducible units of information [14]. Condition (ii) requires that the update satisfy key-uniqueness since, as mentioned in Sect. 1, this is a fundamental property of the relational model which is easily enforceable. We note, however, that there is no concept of the maintenance of a tuple's identity during a modification and so there's no equivalent to condition (iii) of Definition 4.7. We now use the definition to define a normal form for relation schemes which ensures that these violations can never occur.

Definition 9.2 *$(R, \mathbf{\Sigma})$ is in* fact modification normal form 1 *(FMNF$_1$) if there doesn't exist r which has an FMA$_1$.*

We now extend these definitions to allow for different sets of facts.

Definition 9.3 *A relation r has a* fact-based modification anomaly 2 *(FMA$_2$) if it satisfies all the conditions of Definition 9.1 except that condition (iii) is changed to:*

(iii') The set of attributes on which t and t^* differ is a subset of ATT(d) where $d \in \mathbf{\Sigma}'$ and $\mathbf{\Sigma}' = \mathbf{\Sigma} \cup \{X \rightarrow\rightarrow R - XY | X \rightarrow\rightarrow Y \in \mathbf{\Sigma}\}$.

Definition 9.4 *$(R, \mathbf{\Sigma})$ is in* fact modification normal form 2 *(FMNF$_2$) if there doesn't exist r which has an FMA$_2$.*

We note that in the case where the set of constraints contains only FDs, an FMA$_1$ and an FMA$_2$ are identical and thus so are FMNF$_1$ and FMNF$_2$.

Definition 9.5 *A relation r has a* fact-based modification anomaly 3 *(FMA$_3$) if it satisfies all the conditions of Definition 9.1 except that condition (iii) is changed to:*

(iii") The set of attributes on which t and t^* differ is a subset of ATT(d) where $d \in \mathbf{\Sigma}^+$.

Definition 9.6 *$(R, \mathbf{\Sigma})$ is in* fact modification normal form 3 *(FMNF$_3$) if there doesn't exist r which has an FMA$_3$.*

The following example illustrates the previous definitions.

Example 9.1 Let $R = \{A, B, C\}$, $\mathbf{\Sigma} = \{A \rightarrow B, B \rightarrow C\}$ and r is as shown in Fig. 8. Then r has a FMA$_1$, a FMA$_2$ and a FMA$_3$ when $\langle 2, 2, 1\rangle$ is updated to $\langle 2, \mathbf{1}, \mathbf{2}\rangle$ since the attributes $BC = ATT(B \rightarrow C)$ and the resulting relation, r', is in SAT($\mathbf{\Sigma}_k$) (since A is the only candidate key) but violates the FD $B \rightarrow C$.

It follows directly from the definitions of modification anomalies that the following relationships hold: a relation r has an FMA$_1$ $\Rightarrow$ r has an FMA$_2$ $\Rightarrow$ r has an FMA$_3$, and thus in relation schemes: a relation scheme $(R, \mathbf{\Sigma})$ is in FMNF$_3$ $\Rightarrow$ $(R, \mathbf{\Sigma})$ is in FMNF$_2$ $\Rightarrow$ $(R, \mathbf{\Sigma})$ is in FMNF$_1$.

We show the equivalence between 4NF and the modification normal forms.

Theorem 9.1 *The following conditions are equivalent:*

(i) $(R, \mathbf{\Sigma})$ is in 4NF;
(ii) $(R, \mathbf{\Sigma})$ is in FMNF$_3$;
(iii) $(R, \mathbf{\Sigma})$ is in FMNF$_2$;
(iv) $(R, \mathbf{\Sigma})$ is in FMNF$_1$.

r

A	B	C
1	1	1
2	2	1

replace <2, 2, 1> by <2, **1**,**2**> ⇒

r’

A	B	C
1	1	1
2	**1**	**2**

Fig. 8. An example illustrating modification anomalies

Proof.
$(i) \Rightarrow (ii)$: As for Theorem 5.1.
$(ii) \Rightarrow (iii)$: Follows directly from the definitions of the normal forms.
$(iii) \Rightarrow (iv)$: Follows directly from the definitions of the normal forms.
$(iv) \Rightarrow (i)$: We shall prove the contrapositive that if $(R, \mathbf{\Sigma})$ is not in 4NF then it is not in FMNF$_1$. If $(R, \mathbf{\Sigma})$ is not in 4NF then, by the results in [35, 36], there exists a nontrivial dependency $X \rightarrow Y$ or $X \rightarrow\rightarrow Y \in \mathbf{\Sigma}$ such that X is not a superkey. We now construct a relation r which has an FMA$_1$ by considering separately the cases where XY is not a superkey and XY is a superkey.

(*a*) XY *is not a superkey.* From the properties of DEP there exists $W \in$ DEP(XY) such that $W \cap (XY)^+ = \emptyset$. Construct a two tuple relation r for which the two tuples are identical on all attributes except those in W. If any value of an attribute $A \in Y$ is changed to a value that is not in r, then r has an FMA$_1$ from Theorem 3.1 and Lemma 7.5.

(*b*) XY *is a superkey:* We firstly claim that Z, where $Z = R - XY$, is nonempty. This is because if the dependency is an FD then Z is nonempty because X is not a superkey, and if the dependency is an MVD then Z is nonempty because the dependency is nontrivial. Then as XY is a superkey, $XY \rightarrow Z \in \mathbf{\Sigma}^+$ and combining with the fact that $X \rightarrow\rightarrow Y \in \mathbf{\Sigma}^+$ and inference rule A9 then $X \rightarrow Z \in \mathbf{\Sigma}^+$. Then from Lemma 2.2, there has to be a nontrivial FD $V \rightarrow A \in \mathbf{\Sigma}$ for every attribute $A \in Z$ since otherwise the ndv in row $\omega_X[A]$ could not be changed to a dv. Construct then a two tuple relation for which the two tuples agree on X^+ and disagree elsewhere. If the value of an attribute $A \in Z$ is now changed so that the two tuples disagree on A, then r has an FMA$_1$ since $V \rightarrow A \in \mathbf{\Sigma}$, the new relation is in SAT($\mathbf{\Sigma}_k$) (from Lemma 7.3 and the fact that $X \rightarrow Z \in \mathbf{\Sigma}^+$) but violates $X \rightarrow A$. □

10 Related work

The first paper to address the problem of providing a formal justification for the use of normal forms and formalising the concept of an update anomaly was by Bernstein and Goodman [7]. Based on the concept of an update ‘affecting’ some sets of attributes but not others, they provided formal definitions of three types of update anomalies (insertion anomalies, deletion anomalies and replacement anomalies) and proved that BCNF is a necessary and sufficient condition for the avoidance of each type of update anomaly. They also investigated the usefulness of BCNF in the context of multiple relations and showed that in this setting, having individual relation schemes in BCNF does not guarantee an absence of processing difficulties. However, these conclusions were based on the strong *universal instance assumption* (*UIA*) and it was later shown [20] that if one replaced the UIA assumption by the less restrictive and now widely accepted *weak instance approach* [2, 18, 30], then most of the problems encountered by Bernstein and Goodman in the context of multiple relations disappear. More recently, Vossen used a similar approach and derived similar results to Bernstein and Goodman in a slightly different fashion [39]. We have not included this approach in this paper as it was addressed in an allied paper where it was shown that 4NF is equivalent to an absence of insertion anomalies in the sense of Bernstein and Goodman [37].

The original work on the relationship between normal forms and key-based update anomalies is due to Fagin [16, 17]. In the earlier paper, Fagin showed that BCNF and 4NF were equivalent to the property that every relation in SAT($\mathbf{\Sigma}$) is also in SAT($\mathbf{\Sigma}_k$) and then used this property to define the normal forms PJ/NF for join dependencies. In the later paper the approach was generalised to consider include domain constraints and the normal form DK/NF and key-based update anomalies were defined. The main differences between this paper and of Fagin’s work is that we have extended his approach, where only insertions and deletions were considered, to the modification of tuples and also extended his results on insertion and deletions anomalies by deriving necessary and sufficient conditions for an absence of key-based deletion anomalies. Another difference is that we assume that attribute domains are infinite whereas Fagin considered the effect of finite domains.

Another approach to justifying the use of BCNF is due to Biskup [8]. Similar to the approach in Sect. 9, sets of attributes were considered to be the fundamental units of information (which were referred to as *objects*), but these sets were the lhs of FDs rather than all the attributes in an FD as in Sect. 9. Two object normal forms were proposed based on the requirement of being able to insert unique objects into a relation without violating the constraints. It was then shown that one of the normal forms is equivalent to BCNF and the other is equivalent to single key BCNF. In a later paper [9], this approach was extended to the multiple relation setting with inclusion and exclusion dependencies also being allowed. None of the semantic properties considered in this paper are comparable to the one used by Biskup and the other major difference is that MVDs are considered in this paper.

An approach similar to the one used in Sect. 9 was used by Chan in his investigations into the relationship between update anomalies and BCNF [10]. He considered the cases where facts are defined by the constraints and, alternatively, independently of the constraints and defined insertion anomalies, deletion anomalies and replacement anomalies. As mentioned earlier, we feel that a more complete investigation of the relationship between normal forms and fact-based insertion and deletion anomalies requires the incorporation of null values into the definitions of dependency satisfaction and normal forms and have not pursued the issue in this paper. There are several other differences between Chan's work and ours. Chan considered only FD constraints but investigated the relationship between normal forms and update anomalies in multiple relations using the weak-instance approach mentioned earlier, whereas we have considered both FDs and MVDs but only in the context of single relations. Moreover, Chan's definition of a replacement anomaly, while similar in principle to ours, differs in important from the definition of a modification anomaly given in Sect. 9. In his definition, the attributes whose values can be changed are a fixed set of attributes defined by the user and replacements which violate key uniqueness are permitted. In contrast, in our definitions any value can be changed as long as the corresponding attributes is part of a constraint but key violations are not permitted.

11 Conclusions

In this paper, we have addressed the problem of providing a formal justification for the use of 4NF in relational database design. We have formally defined three different semantic properties that it is desirable that a relation scheme should possess. These properties are: an absence of redundancy, an absence of key-based update anomalies and an absence of fact-based update anomalies. For each of the properties, normal forms were defined (called semantic normal forms) which encapsulate the relevant property. The relationship between 4NF and the semantic normal forms was then investigated. For some of the semantic normal forms, we proved that, depending on the types of constraints permitted, either BCNF or 4NF are equivalent conditions to the semantic normal forms, but for other semantic normal forms the equivalent syntactic normal forms are weaker than BCNF or 4NF. In particular, for the semantic normal form which is free of a key-based modification anomaly in which no candidate key value is changed, we proved that in the case of the only constraints being FDs, the equivalent syntactic normal form is a new normal form which lies between 3NF and BCNF. Similarly, in the

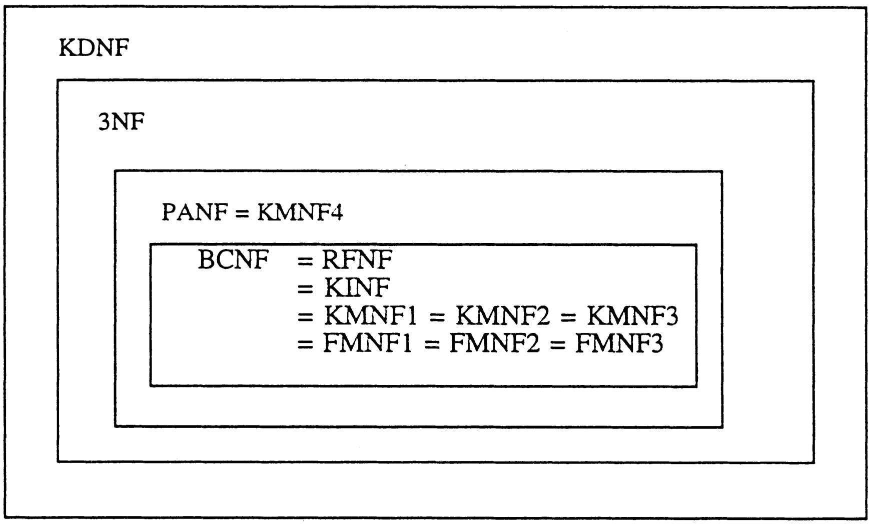


Fig. 9. The relationship between normal forms for the FD case

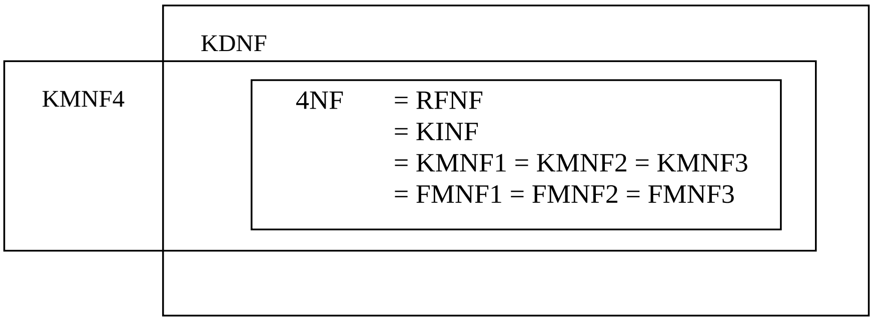


Fig. 10. The relationship between normal forms for the FD and MVD case

case of both FD and MVD constraints, the equivalent syntactic normal form was shown to be a weaker condition than 4NF. The relationship between the normal forms introduced in this paper and the classical normal forms is summarised in Figs. 9 and 10.

There are several other issues related to the work in this paper that warrant further investigation. Firstly, we have assumed in this paper that nulls are not present and a more thorough approach would be to extend the results of this paper to the case where nulls are present. A second issue is the justification of the normal forms used when join dependencies [28] are present. The definition of redundancy given in Sect. 3 is applicable to any type of relational constraint and we are currently investigating the normal forms which ensure an absence of redundancy in the presence of join dependencies. Interestingly, preliminary research [34] has shown that a condition that is weaker than both PJ/NF [17] and 5NF [21] is equivalent to the absence of redundancy. The final issue is to extend the approach used in this paper to develop a formal foundation for database design in the newer data models such as nested relational and object-oriented models. Little work has been devoted to database design for object-oriented models and while normal forms have been defined for nested relations [21, 24, 25, 29] none have been derived from fundamental semantic objectives. Preliminary research has indicated that the redundancy elimination approach used in Sect. 3 extends naturally to nested relations and may provide the basis for understanding and deriving nested normal forms.

References

1. Aho, A.V., Sagiv, Y., Ullman, J.D.: Equivalences among Relational Expressions. SIAM Journal of Computing **8**, 218–246 (1979)
2. Atzeni, P., DeAntonellis, V.: Relational Database Theory. Redwood City, CA: Benjamin/Cummings, 1993
3. Beeri, C.: On the Membership Problem for Functional and Multivalued Dependencies in Relational Databases. ACM Transactions on Database Systems **5**, 241–259 (1980)
4. Beeri, C., Fagin, R., Howard, J.H.: A Complete Axiomatization for Functional and Multivalued Dependencies in Database Relations. Proc. ACM SIGMOD Conference. pp. 47–61, 1977
5. Beeri, C., et al.: Equivalence of Relational Database Schemes. SIAM Journal of Computing **10**, 352–370 (1981)
6. Bernstein, P.A.: Synthesizing Third Normal Form Relations from Functional Dependencies. ACM Transactions on Database Systems **1**, 277–298 (1976)
7. Bernstein, P.A., Goodman, N.: What Does Boyce-Codd Normal Form Do? Proc. VLDB Conference, pp. 245–259, 1980
8. Biskup, J.: Boyce-Codd Normal Form and Object Normal Form. Inform Process Lett **32**, 29–33 (1989)
9. Biskup, J., Dublish, P.: Objects in Relational Database Schemes with Functional, Inclusion and Exclusion Dependencies. Theoretical Informatics and Applications **27**, 183–219 (1993)
10. Chan, E.P.F.: A Design Theory for Solving the Anomalies Problem. SIAM Journal of Computing **18**, 429–448 (1989)
11. Codd, E.F.: Further Normalization of the Database Relational Model. In: Courant Computer Science Symposia 6: Data Base Systems (R. Rustin, ed.), pp. 33–64. Englewood Cliffs, N.J.: Prentice-Hall, 1972
12. Codd, E.F.: Recent Investigations in Relational Database Systems. IFIP Conference, pp. 1017–1021, 1974
13. Date, C.J.: An Introduction to Database Systems. Reading, MA: Addison-Wesley, 1990
14. Desai, B.C., Goyal, P., Sadri, F.: Fact Structure and its Application to Updates in Relational Databases. Information Systems **12**, 215–221 (1987)
15. Fagin, R.: Multivalued Dependencies and a New Normal Form for Relational Databases. ACM Transactions on Database Systems **2**, 262–278 (1977)
16. Fagin, R.: A Normal Form for Relational Databases that is based on Domains and Keys. ACM Transactions on Database Systems **6**, 387–415 (1981)

17. Fagin, R.: Normal Forms and Relational Database Operators. Proc. ACM SIGMOD Conference. pp. 153–160, 1979
18. Honeyman, P.: Testing Satisfaction of Functional Dependencies. Journal of the Association for Computing Machinery **29**, 668–677 (1982)
19. Jajodia, S.: Recognizing Multivalued Dependencies in Relation Schemas. The Computer Journal **29**, 458–459 (1986)
20. Jajodia, S., Ng, P.A.: Update Sets Approach to Databases. Proc. IEEE 7th International Computer Software and Applications Conference, pp. 194–200, 1983
21. Kent, W.: A Simple Guide to the Five Normal Forms in Relational Database Theory. Communications of the ACM **26**, 120–125 (1983)
22. Maier, D.: The Theory of Relational Databases. Rockville, Md: Computer Science Press 1983
23. Maier, D., Mendelzon, A.O., Sagiv, Y.: Testing Implications of Data Dependencies. ACM Transactions on Database Systems **4**, 455–469 (1979)
24. Mok, W.K., Ng, Y.K., Embley, D.W.: An Improved Nested Normal Form for Use in Object-Oriented Software Systems. Proc. 2nd International Computer Science Conference - Data and Knowledge Engineering: Theory and Applications, pp. 446–452, 1992
25. Ozsoyoglu, Z.M., Yuan, L.Y.: A New Normal Form for Nested Relations. ACM Transactions on Database Systems **12**, 111–136 (1987)
26. Ozsoyoglu, Z.M., Yuan, L.Y.: Reduced MVDs and Minimal Covers. ACM Transactions on Database Systems **12**, 377–394 (1987)
27. Paredaens, J., et al.: The Structure of the Relational Database Model. Berlin: Springer-Verlag, 1989
28. Rissanen, J.: Theory of Joins for Relational Databases - A Tutorial Survey. In: Mathematical Foundations of Computer Science, Lecture Notes in Computer Science 64, pp. 537–551. Berlin: Springer-Verlag, 1979
29. Roth, M.A., Korth, H.F.: The Design of ∅1NF Relational databases into Nested Normal Form. ACM SIGMOD International Conference on the Management of Data, pp. 143–159, 1987
30. Sagiv, Y.: Can We Use the Universal Relation Instance Assumption Without Using Nulls? Proc. ACM SIGMOD Conference, pp. 108–120, 1981
31. Thalheim, B.: Open Problems in Database Theory. In: Proceedings 1st Symposium on Mathematical Fundamentals of Database Systems, Lecture Notes in Computer Science no. 305, pp. 241–247. Springer Verlag, 1988
32. Ullman, J.D.: Principles of Database and Knowledge-Base Systems. Computer Science Press, 1988
33. Vardi, M.Y.: Fundamentals of Dependency Theory. In: Trends in Theoretical Computer Science (E. Borger, ed.), pp. 171–224. Rockville: Computer Science Press, 1988
34. Vincent, M.W., A New Redundancy Free Normal Form for Relational Databases. In: Proc. International Workshop on Database Semantics (B. Thalheim ed.), pp. 109–118, Prague, 1995
35. Vincent, M.W., Semantic Justification of Normal Forms in Relational Database Design. Ph.D. Thesis, Department of Computer Science, Monash University, 1994
36. Vincent, M.W., Srinivasan, B.: Redundancy and the Justification for Fourth Normal Form in Relational Databases. International Journal of Foundations of Computer Science **4**, 355–365 (1993)
37. Vincent, M.W., Srinivasan, B.: Update Anomalies and the Justification for 4NF in Relational Databases. Information Sciences **81**, 87–102 (1994)
38. Vossen, G.: Data Models, Database Languages and Database Management Systems. Addison-Wesley, 1990

39. Vossen, G.: A New Characterization of Fd Implication with an Application to Update Anomalies. Inform Process Lett **29**, 131–135 (1988)
40. Yuan, L.Y., Ozsoyoglu, Z.M.: Design of Desirable Database Schemes. J Comput Syst Sci **45**, 435–470 (1992)
41. Zaniolo, C.: A New Normal Form for the Design of Relational Database Schemata. ACM Transactions on Database Systems **7**, 489–499 (1982)