Comment on "Segment Synthesis in Logical Data Base Design"

In [1] the problem of synthesizing relations from functional dependencies (FDs) is attacked by first finding a nonredundant covering of the FDs and then merging together FDs of the covering into relations based on common domain attributes. The authors claim in Proposition 1 that

> If $A \rightarrow B$ and $A \rightarrow C$ are FDs in a nonredundant covering, then the relation $R(A, B, C)$ has no transitive dependencies (i.e., the relation is in Codd's third normal form [2]).

Their proof runs as follows. Suppose $B \rightarrow C$ is an FD in the closure of the covering, i.e., there is a transitive dependency in R. Then, since $A \rightarrow B$ and $B \rightarrow C$ are in the covering, $A \rightarrow C$ must be redundant, which is a contradiction.

That argument, however, is specious. One must show also that $B \rightarrow C$ can be derived from the covering without using $A \rightarrow C$. For example, suppose the covering is $\{A \rightarrow B, A \rightarrow C, B \rightarrow A\}$. Now, $B \rightarrow C$ is in the closure but $A \rightarrow C$ is not redundant.

To correct the proof, one must use the full power of the transitive dependency, that is, $A \rightarrow B$, $B \not\rightarrow A$, and $B \rightarrow C$. It is the second FD, $B \not\rightarrow A$, that drives home the proof. For if $B \rightarrow C$ is derived from the covering using $A \rightarrow C$, then it can be shown that $B \rightarrow A$, which contradicts $B \not\rightarrow A$. I know of no short proof for this latter fact. A rigorous proof appears in [3] as does a corrected version of the proof of Proposition 1.

The authors also claim that their algorithm produces the fewest possible relations that both cover the given FDs and are in the third normal form. This claim is falsified by the nonredundant set of FDs, $\{EMP\# \rightarrow SS\#, SS\# \rightarrow EMP\#\}$, where $EMP\#$ is "employee number" and $SS\#$ is "social security number." If these two FDs are inverses of each other, then the single relation $R(\underline{EMP\#}, \underline{SS\#})$ covers the FDs, whereas the algorithm in [1] synthesizes two relations, one for each FD. An algorithm that synthesizes a provably minimal number of relations is presented in [3].

Philip A. Bernstein
Dept. of Computer Science
Univ. of Toronto
Toronto, Canada M55 1A7

February 23, 1976

References

1. C. P. Wang and H. H. Wedekind, "Segment Synthesis in Logical Data Base Design," *IBM J. Res. Develop.* **19**, 71 (1975).
2. E. F. Codd, "Further Normalization of the Data Base Relational Model," in *Data Base Systems*, Courant Computer Sciences Science Symposium 6, Prentice-Hall Publishing Co., Inc., Englewood Cliffs, NJ, 1972, pp. 33-64.
3. P. A. Bernstein, "Normalization and Functional Dependencies in the Relational Data Base Model," Technical Report No. 60, Computer Systems Research Group, Univ. of Toronto, Canada, October 1975.

The authors reply:

Although not explicitly stated in our paper, it has been our underlying assumption that the declaration of the functional relations has to be complete and reflect the tree relationships in the real-world information. If A and B are functionally related and inverse to each other, then both $A \rightarrow B$ and $B \rightarrow A$ must be explicitly stated. Multiple functional relations between any two attributes must also be explicitly distinguished. Otherwise, there is no way to recover that information algorithmically.

Based on this premise, Proposition 1 in our paper is correct but falls short of treating the equivalent key problem in a general way. We commend Dr. Bernstein's work to extend the proof and look forward to his publication on this subject.

Dr. Bernstein's letter also stated that, given $\{EMP\# \rightarrow SS\#, SS\# \rightarrow EMP\#\}$, our algorithm would synthesize two relations, $R1(\underline{EMP\#}, SS\#)$ and $R2(\underline{SS\#}, EMP\#)$, whereas his algorithm would produce a single relation R. The fact is, however, that any synthesis process must preserve the information content before minimizing the number of relations. Unfortunately, the relation R does not preserve the information contained in $R1$ and $R2$ because there is no way to ascertain whether the relationship between $EMP\#$ and $SS\#$ in R is functional and mutually inverse or nonfunctional, as could occur in $TEACH(\underline{TEACHER}, \underline{STUDENT})$.

C. P. Wang
IBM System Communications Division
Yorktown Heights, NY 10598

H. H. Wedekind
Technical University
Darmstadt, West Germany

March 8, 1976