THE REPRESENTATION OF DATA BASE RELATIONS THROUGH DIGRAPHS

by

ZAHIRUL KABIR CHOWDHURY

B. S. (ME), Bangladesh University of Engineering & Technology, 1979
M. S. (IE), Kansas State University, 1983

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Computer Science

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1984

Approved by:

Major Professor

TABLE OF CONTENTS

THIS BOOK CONTAINS NUMEROUS PAGES WITH THE ORIGINAL PRINTING BEING SKEWED DIFFERENTLY FROM THE TOP OF THE PAGE TO THE BOTTOM.

THIS IS AS RECEIVED FROM THE CUSTOMER.

THIS BOOK CONTAINS NUMEROUS PAGES WITH DIAGRAMS THAT ARE CROOKED COMPARED TO THE REST OF THE INFORMATION ON THE PAGE.

THIS IS AS RECEIVED FROM CUSTOMER.

ACKNOWLEDGEMENT

The author wishes to express sincere appreciation to Dr. Paul S. Fisher, major professor, for his guidance and encouragement throughout the report work.

Gratitude is extended to Dr. Elizabeth A. Unger and Dr. Richard A. McBride for serving on the graduate committee. Special thanks to Dr. Cleopas Angaye for his assistance in some of the analysis.

Last, but certainly not least, the author is very much indebted to his wife, without whose encouragement and support this graduate program may not have materialized. It is to her that this work is dedicated.

LIST OF FIGURES

C H A P T E R - 1

1. Introduction

Through a rapid process of evolution the theory of data base design has reached a state of elegance in the present decade. The three data models, Hierarchical, Network, and Relational, form the basis of most of the seemingly inumerable data base management systems (DBMS) now available commercially. Among the "three great data models", as Ullman (1982) phrases them, the relational view towards data aggregates has been the focus of most of the recent researchers.

The relational model is an unstructured model based upon set theoretic notations, whereas, the hierarchical and network models are structured and tied to graph theoretic notations.

The pioneering work on relational data base theory and normalization of relations was performed by Codd (1970). Codd introduced the First Normal Form (1NF), Second Normal Form (2NF), and Third Normal Form (3NF) of relations. Each of these normal forms is strictly stronger than the lower normal forms. Later Codd and Boyce (1972) introduced the Boyce-Codd Normal Form (BCNF) which is stronger than 3NF. These normal forms are explained in section 2.1. Fagin (1977) proposed decomposition into Fourth Normal Forms (4NF) which includes multivalued dependencies in relational schemes. The previous normal forms assumed only functional dependence (explained in section 2) between attributes of a relation. In a later paper Fagin (1981) introduced Domain Key Normal Form (DKNF) and Project Join Normal Form

(PJNF) which are combined under Fifth Normal Form (5NF). An explanation of 4NF and 5NF is beyond the interest of this report. Interested readers may consult Fagin (1977) and Fagin (1981).

The relational data model, being based upon set theoretic notations, could be established on strong mathematical foundation and is subjected to extensive abstract mathematical treatments to improve the model to eliminate all types of anomalies (insertion, deletion,and update) from the resulting data base schema and to capture more meaning of the data in a data base. However, instead of a complete abstract mathematical treatment a graphical representation will always help the designer in viewing the structure of the relation and communicate more meaning of data to the user. A pictorial representation of a relation will help in understanding the interactions among the attributes of a relation. Properties of functions, for example, can be identified by examining their graphs. When the data base grows too large due to inclusion of new attributes it becomes very difficult to comprehend the interactions among the attributes and obtain logical meaning. A graphical representation of the relations in the schemata can reveal information otherwise incomprehensible due either to size or complexity. This report is an effort to relate the concepts in a relational data model to a graphical representation. In this report we examine the relationships between directed graphs (digraphs) and relational data model.

Grant (1982) made an investigation into the relationships between the connectedness in digraphs and normal forms in relational data bases. The author of this report examined Grant's work and designed a set of lemmas which would help in forming a foundation for the representation of relational data bases through digraphs. The relational data base and its related terms are described in section 2. Digraphs and their related terms, along with a construction methodology of digraphs for the representation of a relation is described in section 3. In section 4 several lemmas have been presented, relating digraphs to the different normal forms of relational data bases. Each of the lemmas has been supported by a proof.

CHAPTER - 2

2. Relational Data Base

A data base deals with information from entities of the real world. A personnel data base of an enterprise, for example, may contain such entities as EMPLOYEE, DEPARTMENT, and MANAGER. Each entity again is described by a set of properties. The EMPLOYEE entity, for example, might be described by such properties as Employee Number (EMP#), Employee Name (EMP_NAME), Employee Age (EMP_AGE), Employee Salary (EMP_SALARY), etc. These properties of entities are called attributes of the entities. Each attribute of an entity may have any value at a particular instance from its domain of values.

If X = (A1 A2 A3 .... An) is a set of attributes with domains S1, S2, ..., Sn, not necessarily distinct, then a relation R on these n domains is a set of n-tuples such that the ith component of each tuple is from domain Si. The Relation R is some times denoted as R = S1 x S2 x S3 x ..... x Sn, which is a subset of the cross product of the domains on which it is defined. A relational schemata in the relational model consists of a collection of relational schemes. Each relational scheme again is an intension or abstract of a relation and is denoted by $R(\Gamma, F)$, where Γ is a set of attributes and F is the set of functional dependencies that holds in that relation. Functional dependencies are semantic constraints that represent relationships among collections of data in the real world and constrain the tuple values possible in a relation. We say a set of attributes X functionally determines a set of attributes Y, denoted by X --> Y, if for each assignment of values to the attributes of X there is only one value associated with each attribute in the set Y for that particular assignment of values to X. In other words, if any two tuples of the relation containing X and Y agree in the values of the attributes in X they must also agree in the values of the attributes in Y.

From a given set of functional dependencies it possible to derive a set of other dependencies among the same set of attributes by following a set of axioms proposed by Armstrong (1974) which are known as Armstrong's Axioms. Let $R(\Gamma,F)$ be a scheme with attribute set l and dependencies F. Let F+ denote the closure of F and let the set of one or more attributes Z, Y, and W be subsets of Γ. Then Armstrongs Axioms are:

Axiom-1 (Reflexivity): If $Z \subseteq Y \subseteq \Gamma$ then $Y \dashrightarrow Z \in F+$

Axiom-2 (Augmentation): If $Y \dashrightarrow Z \in F+$ and $W \not\subseteq Y \cup Z$ then $YW \dashrightarrow ZW \in F+$

Axiom-3 (Transitivity): If $Y \dashrightarrow Z$ and $Z \dashrightarrow W \in F+$ then $Y \dashrightarrow W \in F+$

Axiom-4 (Pseudotranstivity): If $Y \dashrightarrow Z$ and $ZW \dashrightarrow V \in F+$ then $YW \dashrightarrow V \in F+$

Axiom-5 (Union): If $Y \dashrightarrow Z$ and $Y \dashrightarrow W \in F+$, where $Z \not\subseteq W$ and also $W \not\subseteq Z$, then $Y \dashrightarrow ZW \in F+$

Axiom-6 (Decomposition): If $Y \dashrightarrow Z \in F+$ and $W \subset Z$ then $Y \dashrightarrow W$

Using Armstrongs axioms a number of other dependencies can be derived from the set F of dependencies. The minimal subset of dependencies that can be derived from F by using Armostrongs axioms is called the closure of F and is denoted by F+.

In order to be able to determine each tuple of a relation uniquely each relational scheme has a set of one or more keys. A subset X of the set of attributes of a relation R is a key for R iff X determines all the attributes of R and no subset of X also determines all of R's attributes.

The keys in a relation are called candidate keys. One of the candidate keys is arbitrarily chosen as a primary key. A super key in a relational scheme is any set of attributes that contain a key. The attributes which are part of a key are called prime attributes; all the other attributes are called non-prime attributes.

Relational schemes are normalized to incorporate desirable properties in the data base. The aims of normalization are to eliminate the update, insertion, and deletion anomalies and to remove redundancies in the data base. The two techniques of normalization are synthesis of the data base as proposed by Bernstein (1976) and by succesive decomposition proposed by Codd. Codd introduced the 1NF, 2NF, and 3NF of relatinal schemes. Later on Codd (1972) introduced a stronger normal form BCNF. In the following section these normal forms are defined and explained briefly.

2.1. Normal Forms

2.1.1. First and Second Normal Forms

A relation is in first normal form if all of its attributes have simple domains. Let R be a relation in first normal form (1NF) with the set of attributes Γ = { A1, A2, A3, A4, ..., An}. Let X be a subset of Γ and let it be a candidate key in R. Then R is in 2NF if there is no non-prime attribute A in R such that A is partially dependent on X. This implies that no subset of X can determine a non-prime attribute and any set of attributes Y determining A, where Y does not contain A, is not a proper subset of any key of R.

2.1.2. Third Normal Form

In the relation R above let A and B be sets of one or more attributes such that A $\not\subseteq$ B. Then, according to Codd (1972), A is said to be transitively dependent on the key X if X-->B, B-/->X, and B-->A. The *relation R* is in 3NF if R is in 2NF and there is no non-prime attribute A in R which is transitively dependent on any candidate key X in R.

2.1.3. Boyce-Codd Normal Form

Boyce-Codd normal form or BCNF is stronger than 3NF. If A is any attribute in R, then R is in BCNF iff any set of attributes X determining A, such that A $\not\subseteq$ X, implies that X contains a key. The BCNF is stronger then 3NF in the sense that BCNF restricts transitivity of both prime and nonprime attributes on the key whereas 3NF restricts transitivity of only nonprime attributes.

Clearly BCNF implies 3NF since the implication holding for all attributes forces it to hold for non-prime ones as well. Also, 3NF implies 2NF because if X contains a key it cannot be a proper subset of a key.

In order to find the relationship between normal forms of relations and digraphs we need to tie such concepts as keys and normal forms in relational model to similar concepts in digraphs. In the following section a brief discussion on digraphs is presented.

CHAPTER - 3

3. Digraphs

A digraph D is a collection of vertices and directed arcs. More formally, a digraph D is represented as D = (V, T, M) where V is the nonempty set of vertices that participate in the digraph D and T is the set of directed arcs. M is called the directed incidence mapping that maps every arc of T onto some ordered pair of vertices (Vi,Vj). An arc t is in digraph D iff there is a pair of vertices in V, not necessarly distinct, such that t originates from one of the vertices of the pair and terminates on the other. A vertex is denoted by a point and an arc by a line segment between a pair of vertices (Vi,Vj) with an arrow directed from Vi to Vj. The symbolism $t \simeq (v,w)$ will be used to denote an arc t originating from v and directed towards w. An example of a digraph with four vertices and seven directed arcs is shown in Fig. 1.

A relation can be represented by a digraph. The attributes of a relation R can be presented as a set of vertices and the functional dependencies among the attributes can be represented by a set of directed arcs. Before we endeavour to present a methodology of constructing a digraph from a given relation we present some terminology describing the local structure of a directed graph. Only the terminology that are relevant in this research are described in the following subsection.

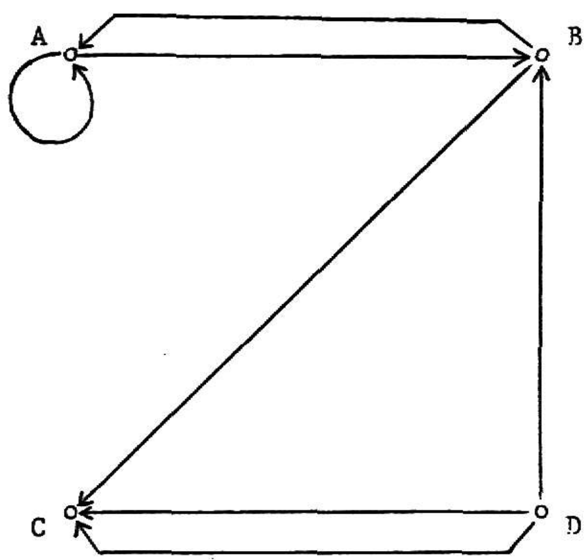


Figure 1: Directed Graph (Digraph)

3.1. Terminology

If in the set of arcs T there exists two arcs t1 and t2 such that t1 $\simeq$ (v,w) and t2 $\simeq$ (v,w) then t1 and t2 are said to be strictly parallel. A digraph D contains a loop if there is an arc t $\simeq$ (v,w) in T such that v = w, ie, the originating vertex and the terminal vertex is the same vertex in D. A digraph is simple if it does not contain strictly parallel arcs and any loop.

The sequence of directed arcs followed to reach from one vertex to another is called a path. If it takes a sequence of n arcs t1, t2,....tn, not necessarily distinct, to reach from vertex v to w we say the path is of length n. If v = w then the path is said to be a closed path. A simple path is one in which all vertices are distinct. A closed simple path is called a cycle. A digraph is said to be cyclic if it contains at least one cycle otherwise it is acyclic. A vertex w is reachable from a vertex v if there is a directed path from v to w; in other words if by following a sequence of directed arcs from v in the direction of the arrows we can reach w then w is reachable from v.

If one allows traversing of any arc in the wrong direction in traveling from one vertex to another, one can define semipaths, simple semipaths, closed semipaths, and semicycles.

The number of arcs originating from a particular vertex is termed the out-degree of that vertex and the number of arcs terminating on a vertex is termed the in-degree of that vertex. A vertex with zero in-degree and zero out-degree is termed a free or isolated vertex.

The set of all digraphs can be divided into four classes according to the degree of connectedness among the vertices. Connectedness is a measure of interaction among the vertices of a digraph. There are four types of connectedness, some times called degrees: 1) strong (degree 3), 2) unilateral (degree 2), 3) weak (degree 1), and 4) disconnected (degree 0). The following are the definitions of these categories:

SC. A digraph D is called strongly connected iff for every pair of vertices (v,w) each one is reachable from the other.

UC. A digraph D is unilaterally connected iff, for any pair of vertices (v,w) at least one is reachable from the other.

WC. A digraph D is weakly connected iff it is not disconnected and there is at least one pair of vertices (v,w) such that they are not reachable from each other.

DC. A digraph D is disconnected iff there exists at least one pair of vertices (v,w) such that there is no path or semipath connecting the two vertices.

It is clear from the definitions above that SC implies UC.

The minimal set of vertices in a digraph that reach to all other vertices in the digraph is called a vertex basis. A digraph may contain more than one vertex basis. Orthogonal to the concept of vertex basis is the concept of contrabasis. The contrabasis of a digraph is the minimal set of vertices such that all other vertices in the digraph can reach at least one vertex in that set. In other words contrabasis is the set of vertices each of which has a zero out-degree. The concept of vertex basis corresponds to the notion of a key in relational databases in the sense that every attribute in a relation is in the closure of its key and evry vertex in a digraph is reachable from the vertex basis. A vertex basis containing a single vertex in it is called a singleton vertex basis, otherwise the basis is nonsingleton. The concepts of relational databases and the concepts of digraphs appear to be isomorphic: attributes map to vertices, functional dependencies map to directed arcs, and keys map to vertex bases. It is a natural outcome of this correspondance to expect that the type of normality of relations would map to the category of connectedness.

In order to investigate the possible relationships between normal forms of relations and connectedness in digraphs it is necessary to formalize the construction methodology of digraphs to represent corresponding relations unambiguously and consistently. In the next subsection we present such a construction methodology.

3.2. Diagraph Construction

Given a relation R the corresponding digraph D of the relation can be obtained by execution of a finite number of steps of the methodology presented in this section. For any relation R = $\{\Gamma, F\}$ the corresponding digraph D of R is obtained by following the steps below:

1. For every functional dependency Y --> W where $Y \nsubseteq \Gamma$ and $W \subseteq \Gamma$ decompose the dependency to obtain Y --> W1, ..., Y --> Wn, where W1, W2, ... , Wn W and are single attributes.
2. For every attribute $A \in \Gamma$ construct a vertex in the digraph and label it by the name of the attribute for convenience.
3. For every dependency Y --> Z in F where each Z is a single attribute do one of the following:
   a) If Y is a single attribute draw a directed arc originating from the vertex representing Y and terminating on the vertex representing Z.
   b) If Y contains more than one attribute then do the following
      b1) construct a new vertex for Y and label it distinctly.
      b2) draw a directed arc originating from Y and terminating on Z.
      b3) for each attribute A in Y draw a directed arc from Y to the vertex representing the attribute A.

b4) for each set $W \subseteq Y$ containing more than one attribute draw a directed arc from Y onto the vertex representing W iff there is at least one dependency in F of the form W --> Z where $Z \in \Gamma$.

The construction methodology described above will produce an unique digraph D for a relation R. This methodology produces a few more considerations that we need to mention. If a vertex is constructed from the left hand side of a dependency and contains more then one attribute then it is termed a "concatenated vertex". Each vertex in the digraph that corresponds to a proper subset of attributes in a concatenated vertex is termed a "component vertex" of that particular concatenated vertex.

The connectedness of the resulting digraph will fall in one of the categories defined in the previous section. By obtaining the connectedness of the digraph and examining the nature of the vertices in the vertex basis and its characteristics it is possible to assert the type of normality that would be obtained in the corresponding relation. A set of lemmas have been presented in the following chapter which are results of a number of observations of the relationship between digraphs and relations.

C H A P T E R - 4

4. Relationship Between Connectedness and Normality

By examining the characteristics of digraphs of the four categories of connectedness several interesting results were obtained. These results are explained under the heading of each category.

4.1. Strong Digraphs

If for a relation R the corresponding digraph D is strongly connected the relation R is always found to be in BCNF. The following lemma is proposed from this observation.

LEMMA - 1:

let R (Γ,F) be a relation where Γ is the set of attributes and F is the set of functional dependencies. *Then if G is the digraph of R and is* strongly connected then R is in BCNF.

PROOF :

Suppose the digraph G of R is strongley connected and R is not in BCNF. Let X be a key in R. Since R is assumed not to be in BCNF then there exists sets A and B of one or more attributes in R such that A --> B $\in$ F+ and A is not a key in R. Therefore X --> A but A -/-> X. *Then in the* corresponding digraph D of R, A is reachable from X but X is not reachable from A and therefore D is not strongly connected. This is a contradiction. Therefore no two sets of *attributes A and B with such property* exist in R, and R is in BCNF.

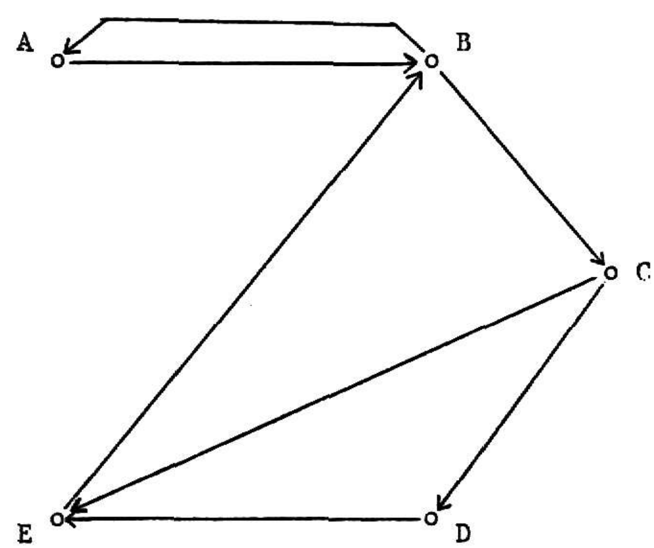


Figure 2: Strongly Connected Digraph

The lemma above asserts that strongly connected digraphs represent relations in BCNF but it does not assert that all BCNF always map to strongly connected digraph. Infact digraphs of other forms may also represent BCNF relations which are presented in subsequent lemmas.

As an example of the case of strongly connected digraphs, let us consider a relation $R = (\Gamma, F)$ where $\Gamma = (A\ B\ C\ D\ E)$ and $F = \{A\text{-->}B,\ B\text{-->}C,\ C\text{-->}ED,\ E\text{-->}BA,\ D\text{-->}E\}$. The corresponding digraph D is shown in Figure-2. By examining the digraph it is clear that starting at any vertex one can reach to all other vertices in the digraph. Therefore each vertex is a basis. By mapping the basis to the candidate keys in the relation each attribute in R is a candidate key since all other attributes is in its closure. Hence it is clear that R is in BCNF.

4.2. Unilateral Digraphs

In case of unilateral digraphs, for every pair of vertices at least one must reach the other, it implies that the basis must be singleton. If a digraph of a relation is unilaterally connected containing a concatenated vertex then that concatenated vertex must be a vertex basis from the fact that the unilateral digraphs contain singleton vertex basis and there is an arc terminating on a concatenated vertex only if the attributes in it are a proper subset of attributes of another concatenated vertex in the same relation. Any subset of vertices of an unilaterally connected digraph is also unilaterally connected with a singleton vertex basis. This suggests the observation that unilaterally connected digraphs containing concatenated vertices must contain extraneous attributes in the concatenated vertices. The following lemma states this result.

LEMMA - 2:

Let G be the digraph of $R(\Gamma,F)$ where Γ is the set of attributes and F is the set of functional dependencies. Let G be not strongly connected but unilaterally connected such that G contains concatenated vertices. Then $R(\Gamma,F)$ contains extraneous attributes.

PROOF :

Since G contains concatenated vertices then according to the construction of G, R contains some functional dependencies with more than one attribute on the left side of the dependencies which map to the concatenated vertices of G. Let X be a concatenated vertex in G. Then by construction of G there exists a component vertex for each attribute X1, X2,....,Xn $\in$ X. Since G is unilaterally connected there is a component vertex Xi such that Xi reaches to all other vertices X1, X2,....,Xi-1, Xi+1,...,Xn. Therefore X1, X2,.....,Xn $\in$ (Xi)+. But X = (X1, X2,...,Xn) and so X $\in$ (Xi)+.

Let A be the right side of the dependency for which X is the left side. Then A $\in$ (X)+. Since X $\in$ (Xi)+ so also is (X)+. Therefore A $\in$ (Xi)+ and the attributes X-Xi $\in$ X are extraneous in the dependency X --> A.

The Lemma reveals the presence of extraneous attributes in the relations represented by unilaterally connected digraphs with concatenated vertices in them. It does not assert, however, that relations with extraneous attributes always map to an unilateral digraph. In actuality a digraph with concatenated vertices indicate the presence of extraneous attributes in the corresponding relation if one of the following two conditions hold:

1) For a concatenated vertex X there is at least one component vertex Y such that if Z is the set of attributes obtained by the union of the attributes of the vertices reachable from Y then $X \subseteq Z$.

2) Any of the noncomponent vertices reachable from a concatenated vertex X through a path length of 1 is also reachable from any of the component vertices of X.

A unilateral digraph containing concatenated vertices can be reduced to an equivalent unilateral digraph without any concatenated vertices in it. According to the proof of Lemma-2, for a concatenated vertex, say Y=(ABC), there is one component vertex, say A, which reaches to all other component vertices of Y. If now for every noncomponent vertex which is reachable from Y through paths of length 1 we draw an arc from A to those vertices and delete Y and all arcs originating from Y then the resulting digraph will still be unilateral. Figure-3 depicts this situation. The concatenated vertex Y in digraph D is removed and the resulting digraph D' is still unilaterally connected.

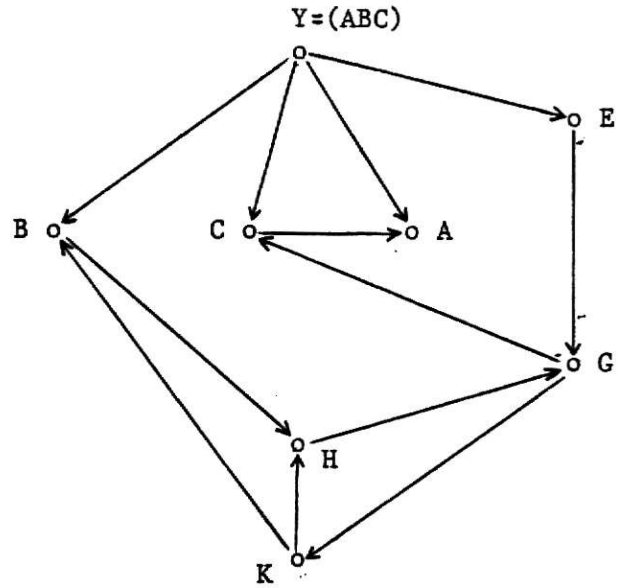
Since any unilateral digraph with concatenated vertices can be reduced to an equivalent unilateral digraph without any concatenated vertices in it, Lemma-3 will consider only the simpler case of unilateral digraph with no concatenated vertex.

LEMMA - 3:

Let D be the digraph of relation $R(\Gamma, F)$ where Γ is the set of attributes and F is the set of functional dependencies. Let D be unilaterally connected and not strongly connected, and has no concatenated vertex in it. Then $R(\Gamma, F)$ is in 2NF.

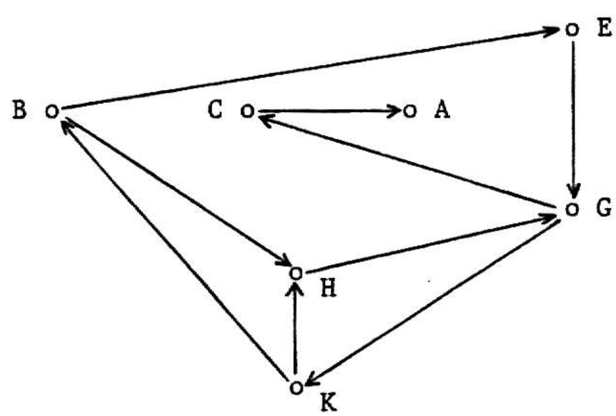
PROOF :

Suppose R is not in 2NF, then there exists a key X and an attribute A, not in X, in Γ such that Xi --> A and Xi $\subset$ X. In the corresponding digraph D of R either the basis Xv corresponding to X is not singleton, i.e., Xv is a set of vertices or the basis vertex Xv corresponding to X is a con-



F (ABC —> E, E —> G, G —> C, C —> A, G —> K,
H —> G, K —> H, B —> H, K —> B)

Digraph - D with Concatenated Vertex Y



F (B —> E, E —> G, G —> C, C —> A, G —> K,
H —> G, K —> H, B —> H, K —> B)

Digraph - D' Reduced From D (Y eliminated)

Figure 3: Unilateral Digraphs and Elimination of Extraneous Attributes

catenated vertex with component vertex Xi. Since the vertex basis of an unilateral digraph is a singleton vertex, the basis vertex Xv cannot be nonsingleton. For the later case, Xv in D cannot be a concatenated vertex since it contradicts that D does not contain any concatenated vertex in it. Therefore, there exists no such key X and an attribute A in R with the property that Xi --> A and Xi $\subset$ X eliminating any possibility of a partial dependence of a nonprime attribute on a key. Therefore, R is in 2NF.

NOTE:

In a special case of unilateral digraphs with n vertices and n-1 vertex bases the digraph will represent a relation in BCNF. The nonbasis vertex will be in the contrabasis and must represent the only nonprime attribute of the corresponding relation. The left hand sides of all dependencies are keys in the relation since each of them map to a vertex basis in the digraph.

4.3. Weak Digraphs

The weakly connected and disconnected cases have more than one interpretation. Therefore another avenue of approach is considered to relate these types of digraphs to the types of normality in corresponding relations. The way the vertices in a particular vertex basis and the vertices in the corresponding contrabasis partitions the whole digraph and the type of vertex basis (concatenated or nonconcatenated and singleton or nonsingleton) determines the type of normality to expect in the corresponding relation. If the vertex basis and the corresponding contrabasis form a partition in the digraph then the type of normality depends on whether the basis is singleton or not. The following lemma is proposed from this observation.

LEMMA - 4:

Let the relation R(Γ ,F), where Γ is the set of attributes and F is the set functional dependencies, have the weakly connected digraph D such that the basis and contrabasis are disjoint and forms a partition of the vertices in D. Then R is in BCNF if the basis is singleton, it is 1NF otherwise.

PROOF :

Suppose R is not in BCNF and X is a key of R. Since R is assumed not to be in BCNF then there exist two sets A and B of one or more attributes such that A --> B and A is not a key. Therefore X --> A, A -/-> X, and A --> B. In the corresponding digraph D of R, X is in the basis, B is in the contrabasis but A is neither in the basis nor in the contrabasis. Therefore the basis and contrabasis do not form a partition of the vertices in D. This contradicts the requirement that the basis and contrabasis form a partition in D. Therefore there are no attributes A and B and no key X with the property that X-->A, A-/->X, and A-->B in R.

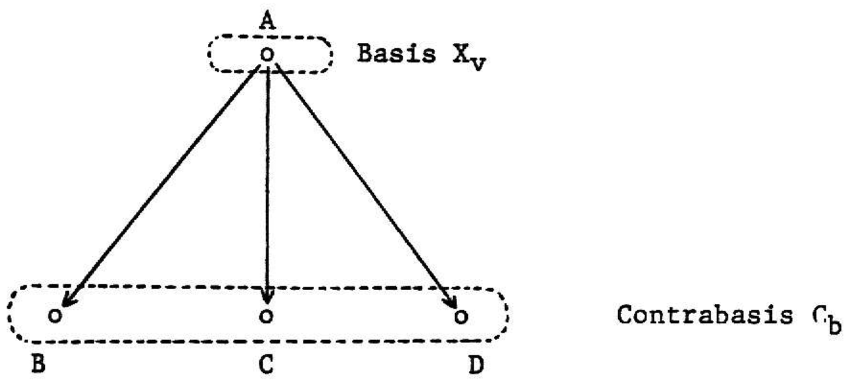
Now we need to show that the basis of D must be singleton for R to be in BCNF. Suppose the basis of D is not a singleton vertex. Then there exists a path from each vertex in the basis to some vertex in the contrabasis. In the corresponding relation R there will be functional dependencies with attributes in each vertex of the basis as left hand side of a dependency. Now since each basis of a digraph maps to a key in R then there exists a key X in R such that Xi $\subset$ X and Xi --> A, where A is some attribute in R. Hence A is partially dependent on key X and R is not in 2NF and therefore not in BCNF.

With the basis and contrabasis creating a partition eliminates any transitivity on prime or nonprime attribute as shown in the first half of the proof and in addition to the partitioning a singleton vertex basis eliminates the possibility of partial dependency on the key. Therefore all dependencies in R are the result of a key and R is in BCNF.

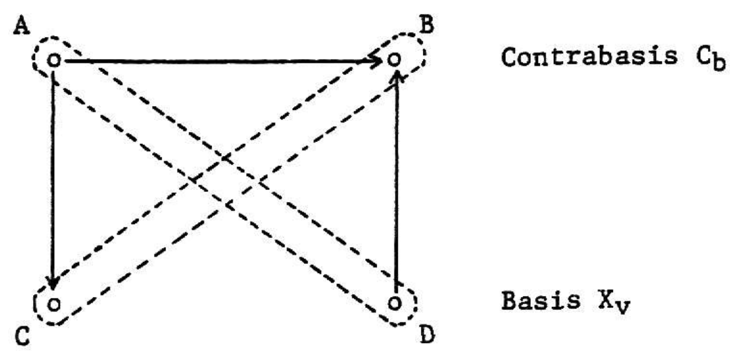
In Figure-4(a) the vertex basis of the digraph is A and the contrabasis Cb = (B C D). In Figure-4(b) the vertex basis is Xv = (A D) and the contrabasis is Cb = (B C). In both the cases the union of the vertices in the basis and contrabasis is the whole digraph but the difference is that the former one is singleton whereas the basis of the later is nonsingleton. The relation R(Γ1,F1) corresponding to Figure-4(a) with Γ1 = (A B C D) and F1 = {A-->BCD} is in BCNF whereas the relation R(Γ2,F2) corresponding to Figure-4(b) with Γ2 = (A B C D) and F2 = {A-->BC, D-->B} is in 1NF.

Weakly connected digraphs with disjoint basis and contrabasis where their union does not equal the digraph, determine the type of relations they represent depending upon whether the basis is singleton or not as well as whether the singleton basis is a concatenated vertex or not. Three different cases may arise. The following three lemmas consider these three cases. All these three lemmas could be designed as three different cases of a single lemma. The division into three lemmas is to make the proofs of these lemmas easier.

If the union of vertices in the vertex basis and contrabasis of a digraph does not constitute the whole digraph then there are vertices in the digraph which are neither in the vertex basis nor in the contrabasis. This set of vertices indicate possibility of transitive dependency in the corresponding relation. Thus 3NF may not be obtained in the relations of such digraphs. But if the basis is a singleton nonconcatenated vertex then there exists no possibility of a partial dependency and the corresponding



4(a): Singleton Vertex Basis



4(b): Nonsingleton Vertex Basis

Figure 4: Two Cases of Weak Digraphs

relation will be in 2NF. In fact it can be generalised that whenever the vertex basis of a digraph is a singleton nonconcatenated vertex the corresponding relation will at least be in 2NF. The following lemma concerning such weakly connected digraphs is motivated by this observation.

LEMMA - 5:

Let the relation $R(\Gamma, F)$, with Γ as the set of attributes and F as the set of functional dependencies, have the weak digraph D. Let D have a disjoint basis and contrabasis such that the union of their vertices does not equal the total vertices in the digraph and let the basis be a singleton nonconcatenated vertex, then R is in 2NF.

PROOF :

Suppose R is not in 2NF but is in 1NF. Let X be a key in R. Since R is assumed not to be in 2NF then there exists partial dependencies in R. Therefore there exists an attribute $A \in \Gamma$ such that Xi-->A where $Xi \subset X$. Therefore X contains more than one attribute in it. In the corresponding digraph D the vertex basis Xv corresponding to X is either a concatenated vertex or the basis is not singleton. Both of these contradicts the fact that the vertex basis is a singleton, nonconcatenated vertex. Therefore there is no key X in R that contains more than one attribute and hence there is no partial dependency on the key in R and R is in 2NF.

We now show that R is not in 3NF. Suppose R is in 3NF and X be a key in R. Then there exist no two nonprime attributes $A,B \in \Gamma$ such that X-->A, A-/->X, and A-->B holds in R. In the corresponding digraph D of R there is no vertex corresponding to such an attribute A that is neither in the basis nor in the contrabasis. In other words, the basis and contrabasis form a partition in D. But this is a contradiction since then the union of the vertices in the vertex basis and the vertices in the contrabasis will constitute the whole digraph. Therefore there must at least be one vertex in D which is not a member of the vertex basis nor is it a member of the contra-basis, and thus introduces a transitivity involoving the key in R and two nonprime attributes in R. Therefore R is not in 3NF.

As an illustration of the result we consider the relation $R(\Gamma, F)$ with Γ = (A B C D E) and F = {A-->DE, B-->DA, D-->C}. Figure-5 shows the corresponding digraph D. The vertex basis is B and the contrabais is Cb = (C E). The vertices A and D are non prime and constitutes a transitivity of the vertices C and E on the key. The relation is thus not in 3NF but clearly is in 2NF.

If the vertex basis is a singleton, concatenated vertex then for weakly connected digraphs, with disjoint basis and contrabasis, the corresponding relation will be in 2NF only if there is no arc from any of the component vertices to any noncomponent vertex of the corresponding concatenated vertex basis. This restricts all the component vertices to be in the set that forms the contrabasis. This assures no partial dependency in the relation that the digraph represents. The following lemma is motivated from this analysis.

LEMMA - 6:

Let the relation $R(\Gamma, F)$, with the set of attributes 1 and set of functional dependencies F, have the weak digraph D with disjoint basis and contrabasis such that the union of their vertices does not equal the total vertices in the digraph. Let the basis be a singleton, concatenated vertex with the component vertices in the contrabasis, then R is only in 2NF.

PROOF :

Suppose R is not in 2NF and let X be a key in R. There exists an attribute $A \in \Gamma$ such that Xi-->A $\in$ F+ where Xi $\subset$ X. Then in the

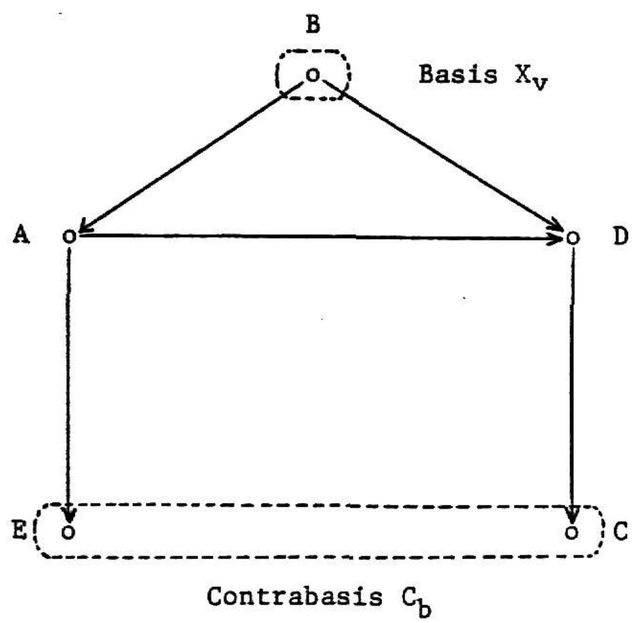
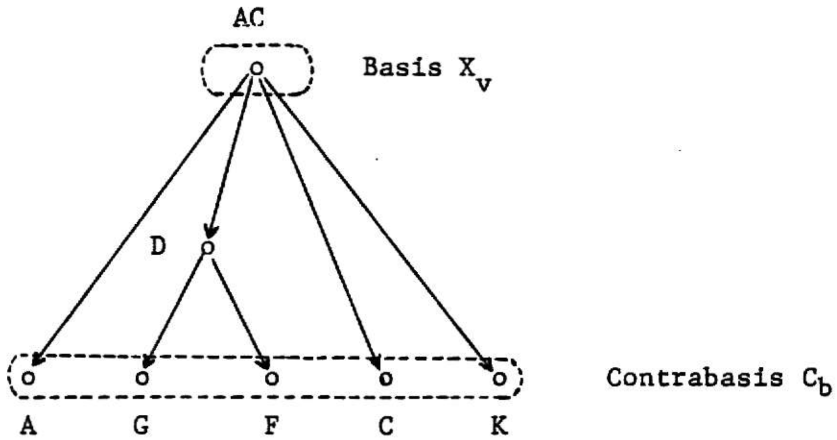


Figure 5: Weak <u>D</u>igraph with Basis $\cup$ Contrabasis $\neq$ Digraph

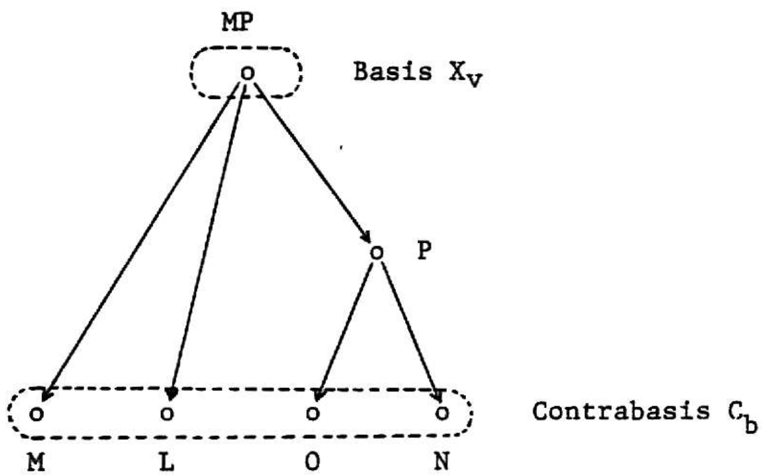
corresponding digraph D there is a directed path from vertex Xi to the vertex representing the attribute A. According to the construction of D there is a directed arc from X to its component vertex Xi. Hence, there is a path from X to A on which Xi appears. By the definition of contrabasis, Xi cannot be a member of the contrabasis. This contradicts the fact that all the component vertices of the concatenated vertex basis is in the contrabasis. Therefore there is no attribute A in R with the property that for some $Xi \subset X$, $Xi \dashrightarrow A \in F+$. Therefore $R(\Gamma, F)$ is in 2NF.

If the vertex basis of the digraph is a concatenated vertex and all of its component vertices are in the contrabasis the relation it represents is in 1NF. The corresponding key in the relation will be the left side of a dependency containing more than one attribute. For each component vertex, which is not in the contrabasis, there would be a dependency in the relation with the attributes of the component vertex on the left hand side. This set of attributes is a subset of the key. Therefore the relation contains a partial dependency.

The two situations are illustrated in Figure-6. In digraph D1 the vertex basis is Xv = (AC), and it is a concatenated vertex. The contrabasis is Cb = (A G F C K). Both the component vertices, A and C, are in the contrabasis eliminating any possibility of partial dependency of nonprime attributes on either A or C. In digraph D2 the vertex basis is Xv = (MP) and is a concatenated vertex. The contrabasis is Cb = (M L O N). The component vertex P is not a member of the contrabasis. In the relation $R(\Gamma, F)$ represented by the digraph the key is (MP) and F = {MP --> L, P --> ON}. Hence the nonprimes O and N are partially dependent on the key MP. The relation, therefore, is in 1NF.



Digraph D_1



Digraph D_2

Figure 6: Two Weak Digraphs with Concatenated Vertex Basis

The vertex basis of a digraph is a set of vertices each with a nonzero out-degree except in case of disconnected digraph. Disconnected digraphs will be discussed later. Each vertex with a nonzero out-degree, by the construction of the digraph, is the left side of some functional dependency that holds in the relation it represents. If the vertex basis Xv of a digraph is not singleton then the set of attributes in the key of the relation, to which the basis maps, is a union of the attributes on the left side of two or more functional dependencies. This obviously means that the relation exhibits partial dependency. In fact since the vertex basis of both strongly connected and unilaterally connected digraphs is always singleton this situation may occur only in cases of weakly connected and disconnected digraphs. The following lemma concerning weak digraphs is based upon this fact.

LEMMA - 7:

Let $R(\Gamma, F)$ be a relation with the digraph D. Let D have a disjoint basis and contrabasis with the basis being nonsingleton, then R is only in 1NF.

PROOF :

Since the domains of all attributes are considered simple then either R is in 1NF or in some higher normal form. We only need to show that R is not in 2NF.

Suppose R is in 2NF and let X be a key in R. Then there is no attribute $A \in \Gamma$ such that Xi --> $A \in$ F+ where $Xi \subset X$. For that case in the corresponding digraph D of R there is no path of length 1 or more from the vertex representing Xi, a member or a component of the corresponding basis Xv, to any vertex A not in the basis. Then either Xv is a singleton nonconcatenated vertex in which case no Xi vertex exists in Xv or Xv is a singleton concatenated vertex with all component vertices in the contrabasis. Either of these conditions contradicts the fact that the basis Xv is not singleton. Therfore R is not in 2NF.

Disconnected digraphs also show the same behavior whenever the vertex basis is not singleton. The digraph in Figure-4(b) is weakly connected since vertices A and D are not reachable from each other. The vertex basis is Xv = (A D) and the contrabasis is Cb = (B C). The relation represented by this digraph is $R(\Gamma,F)$ with Γ = (A B C D) and F = {A --> BC, D --> B}. The only key in the relation is AD which clearly shows the existence of partial dependency of both B and C on the key.

4.4. Disconnected Digraph

All of the lemmas presented concerns digraphs which are connected. Connected digraphs, however, cannot handle attributes in the relation which are neither dependent on any attribute nor do functionally determine the value of any other attribute. These attributes create disconnected digraphs. Considering the interactions among the attributes a disconnected digraph can have any of the following combinations of structural elements:

a) all n vertices each with zero in-degree and zero out-degree

b) m vertices with zero in-degree and zero out-degree and one or more groups of connected vertices such that none of n-m vertices participates in more than one group of connected vertices.

c) two or more groups of connected vertices, D1, D2, ...,Dn, such that each of n vertices participates in exactly one connected group.

In Figure-7 one example for each type of disconnected digraph is presented. In the digraph Da, corresponding to case (a), no pair of

Digraph D_a

Digraph D_b

Digraph D_c

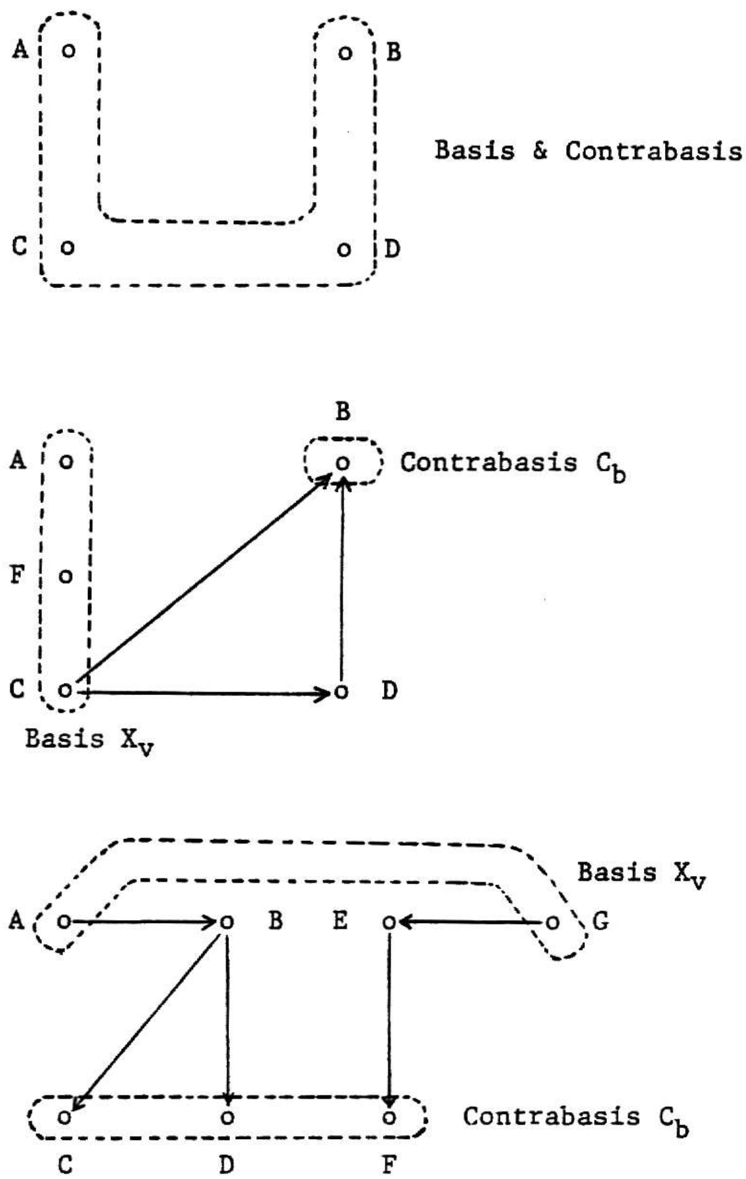


Figure 7: Three Classes of Configurations for Disconnected Digraphs

vertices is connected. The basis and contrabasis are same, Xv = Cb = (A B C D). They form a duobasis. Since the vertex basis maps to a key in the relation, the key contains all of the attributes in the relation and the relation is in BCNF by default.

LEMMA - 8:

Let D be the digraph of relation R(Γ,F), with the set of attributes Γ and the set of functional dependencies F. Let D be a disconnected digraph such that R is its own duobasis. Then R is in BCNF.

PROOF :

Since R is its own doubasis then the basis and contrabasis are the same. Then all the attributes in the corresponding relation participate in the vertex basis. Therefore the key in the relation contains all the attributes of the relation. Since there is no nonprime attribute the relation R is in BCNF.

In Figure-7 digraphs Db and Dc, examples of case (b) and case (c) respectively, indicate that in case (b) and in case (c) the vertex basis cannot be singleton. In digraph Db of Figure-7 the vertex basis Xv = (A F C) and contrabasis Cb = (B). The key in the corresponding relation is AFC. Attributes B and C are then certainly partially dependent on the key. A similar situation exists in digraph Dc. The vertex basis Xv = (A G) and contrabasis Cb = (C D F). The only difference is that in digraph Dc there is no attribute with in-degree and out-degree both zero. By lemma-7 the relations corresponding to digraph Db and Dc are in 1NF.

In fact, whenever a disconnected digraph contains groups of connected vertices the digraph and thus the corresponding relation can be decomposed into component digraphs. At most one of the component digraphs may remain disconnected containing all the isolated or the free vertices and nothing else. Steps c and d below for a decomposition can be followed in any order:

a) Set n to the number of vertices in the digraph.
b) Set m to the number of isolated vertices in the digraph.

c) If m = 0 then go to step-b else do the following

   1) Create digraph D1 such that number of vertices Nd = 0
   2) For each vertex A with in-degree = out-degree = 0 introduce A in the digraph D1 and set Nd = Nd + 1. When Nd = m go to step-b.

d) For the n-m connected vertices create digraphs D2, D3,....,Dn such that each D1+i for i = 1 to n-1 map to a group of connected vertices.

Combining all the isolated vertices in one digraph will produce a BCNF relation. The connectivity of the digraphs formed from the sets of connected vertices will fall in any of the three previously discussed categories and the corresponding relation can be analysed using the lemmas presented.

4.5. Normalization of Relations Through Digraph Representation

In Bernstien's algorithms for normalization of relations, free or isolated attributes are lost from the data base during the process of synthesis as extraneous attributes. Such relations presented as digraphs will create disconnected digraphs. Analysis of these relations through disconnected digraph representation will prevent elemination of these attributes from the data base and eliminate inadvertent loss of informations.

Normalization of relations is possible through analysis of the digraphs of relations. The conditions that must hold in a digraph for existence of extraneous attributes in the relation it represents have been discussed in section-4.2. The elimination of these undesirable properties from the digraphs is also very simple.

It has been observed that redundant functional dependencies may exist in a relation only if in its digraph at least one pair of vertices are K-connected, where K > 1. In other words, if one of the vertices of the pair is reachable from the other through more than one path. In such cases redundancy exists iff one of the paths is of length one.

More formally stated, a functional dependency X —> Y between two sets of attributes X and Y is redundant iff in its digraph D

a) there is an arc originating from X and terminating on Y and

b) Y is reachable from X through a path of length > 1.

This form of redundancy can be eliminated by simply deleting the arc from X to Y and incorporating the equivalent change in the corresponding relation.

Deo (1974) discussed the techniques of identifying the connectivity of a digraph by representing the digraph through its adjacency matrix or some times called relation matrix in the calculus of relations. The author also presents an algorithm for finding the components of a disconnected digraph and then examining the connectivity of each component.

Decomposition of a relation into BCNF is possible through decomposition of the digraph of the relation into strongly connected components. Let D= (V,T) be a digraph with vertex set V and arc set T. Ds=(Vs,Ts) is a generated subgraph of D iff: 1) Vs is a subset of V; and 2) Ts is the set of all arcs in T connecting the vertices in Vs. A generated subgraph is a maximal strongly connected component or some times called a fragment, if it is strongly connected and no superset of the vertices is also strongly connected. Since strongly connected digraphs represents relations in BCNF, finding the strongly connected components of the digraph of a relation is a easy way of decomposing the relation into BCNF relations. Tarjan (1972) presented an algorithm for decomposing a digraph into strongly connected components.

Similarly, decomposition into unilateral components can result in decomposing the relation into a set of 2NF relations.

CHAPTER - 5

5. Summary and Conclusion

This research is an effort to explore the possibility of explaining the relational model of data through graph theoretic notations. It is possible to tie the concepts of data base relations to the concepts in directed graph theory. Structural mapping between the two can be defined by 1) mapping attributes into digraph vertices; 2) functional dependencies into arcs connecting pairs of vertices; and 3) candidate keys into vertex bases. A methodology for constructing a digraph from the set of attributes and functional dependencies of a relation is presented. This methodology will produce an unique digraph for a given relation.

Digraphs are classified into four categories according to the connectivity of vertices. These categories are: 1) strong digraphs, 2) unilateral digraphs, 3) weak digraphs, and 4) disconnected digraphs. It is possible to predict the normality that exists in a relation by examining the connectivity of its digraph and properties of its basis and contrabasis. The union of the attributes in the vertex bases represents the set of prime attributes in the relation and each basis represents a candidate key of the relation, whereas contrabasis represents the set of attributes each of which does not appear in the left hand side of any dependency independently. Several lemmas were presented which formalized the relationship between the connectivity of digraphs and the normality of the corresponding relations. Strong digraphs always represent relations in BCNF. Unilateral digraphs represent relations in 2NF. The cases of weak digraphs and disconnected digraphs are complex. Whenever the basis and contrabasis partitions the set of vertices and the basis is singleton the corresponding relation is in BCNF. Whenever the basis is not singleton the relation is in 1NF. Disconnected digraphs indicate the necessity of decomposition of the relation through decomposition of its digraph into connected components.

The necessary conditions that must hold in a digraph to indicate the presence of extraneous attributes in the dependencies of the relation and the conditions for existence of redundant functional dependencies were discussed. The process of elimination of these undesirable properties is very easy.

There are many other areas for further research. For example, it would be worthwhile trying to develope efficient algorithms to decompose a digraph into components that will result in a decomposition of the relation into BCNF and still preserve the dependencies and reconstructibilty. It is possible to develop computer algorithms to detect and eliminate extraneous attributes and redundant functional dependencies by examining the adjacency matrix of the digraph.

Present research does not include multivalued dependencies among attributes. Further research can be performed to investigate the possible representation of multivalued dependencies in graph theoretic notations.

BIBLIOGRAPHY

1. Armstrong, W. W., "Dependency Structures of Database Relationships", Proc. 1974 IFIP Congress, North Holland, Amesterdam, pp. 580-583.

2. Ausiello, G., D'atri, A., and Sacca', D., "Graph Algorithms for the Synthesis and Manipulation of Data Base Schemes", Procceedings of the International Workshop WG 80 on Graphtheoretic Concepts in computer science, Bad Honnef/Bonn, June 1980, Springer-Verlag, pp. 212-233.

3. Batini, C., "Schema Hypergraphs: A Formalism to Investigate Logical Data Base Design", Procceedings of the International Workshop WG 80 on Graphtheoretic Concepts in Computer Science, Bad Honnef/Bonn, June 1980, Springer-Verlag, pp. 177-194.

4. Bernstein, P. A., "Synthesizing Third Nomal Form Relations from Functional Dependencies", ACM Transactions on Database Systems, Vol.1, No. 4, Dec. 1976, pp. 277-298.

5. Busacker, R. G. and Saaty, T. L., "Finite Graphs and Networks", McGraw-Hill, New York, 1965.

6. Codd, E. F., "A Relational Model of Data for Large Shared Data Banks", Communications of the ACM, Vol. 13, No. 6, June 1970, pp. 377-387

7. Codd, E. F., "Further normalization of the Database Relational Model", Data Base Systems, Prentice-Hall, Englewood Cliffs, New Jersey, 1972, pp. 33-64.

8. Deo, N., "Graph Theory with Applications to Engineering and Computer Science", Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

9. Fagin, R., "Multivalued Dependencies and a New Normal Form for Relational Database", ACM Transactions on Database Systems, Vol. 2, No. 3, Sept. 1977, pp. 262-278.

10. Fagin, R., "A Normal Form for Relational Database That Is Based on Domains and Keys", ACM Transactions on Database Systems, Vol. 6, No. 3, Sept 1981, pp. 387-415.

11. Grant, D. J., "The Relationship Between Normality of Relations and Connectedness of their Digraphs", Unpublished Term-Paper, Computer Science Department, Kansas State University, Fall-1982.

12. Ullman, J. D., "Principles of Database Systems", Computer Science Press, Rockville, Maryland, 1982.

THE REPRESENTATION OF DATA BASE RELATIONS
THROUGH DIGRAPHS

by

ZAHIRUL KABIR CHOWDHURY

B. S. (ME), Bangladesh University of Engineering & Technology, 1979
M. S. (IE), Kansas State University, 1983

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Computer Science

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1984

The Relational Data Model is based on set theoretic notation. It has been subjected to extensive mathematical treatment by recent database theorists. But instead of only abstract mathematical treatment a graphical representation of the relation will always help the designer view the interactions among the attributes and can reveal informations otherwise incomprehensible due either to size or complexity of the database. The present research investigated possible relationships between graph theoretic notations and relational data model. It has been observed that attributes in a relation can be mapped to vertices in the corresponding digraph, the functional dependencies can be mapped to directed edges, and the keys in the relation can be mapped to vertex bases in the digraph.

Interesting relationships were observed between connectedness in digraphs and normal forms of the corresponding reltions. The relations with strong digraphs are always in BCNF whereas the relations with unilateral digraph are always in 2NF. The relations with weak or disconnected digraphs may result in various normal forms but can still be classified considering the type of their vertex bases and contrabasis. The analysis of relational data bases through adjacency matrix of the corresponding set digraphs is possible. The attributes in a relation which do not bear any dependency with any other attribute of the relation are lost during the process of synthesis by Bernstien's algorithms. These attributes map to isolated vertices in disconnected digraphs and may be possible to retain them in the normalised data base if analysed through adjacency matrix of the corresponding digraphs.

The detection of extraneous attributes and redundant functional dependencies are rather easy tasks and their elimination is simple.