BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: Census/SRD/RR-88/06

USING GRAPH THEORY TO ANALYZE SKIP PATTERNS IN QUESTIONNAIRES

by

Jim Fagan
Brian V. Greenberg
Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

Recommended by: Lawrence Ernst

Report completed: February, 1988

Report issued: December 22, 1988

I. INTRODUCTION

For most survey questionnaire forms (especially those administered by an interviewer) questions are not asked in a strictly linear fashion. For example, if the response to some question is "YES", the interview will follow it with a question different from that were the response "NO". If the questions on a survey form are numbered sequentially, Q1, Q2,...,QN, then one fully completed response form might look Q1-Q5-Q7-Q12-Q50 and another might look like Q1-Q8-Q12-Q39-Q50 where N=50. The remaining questions in each case are not missing in the usual sense of missing data, but were not asked and are "not applicable" based on the responses provided at an earlier stage and on the underlying structure of the questionnaire. These diverse patterns of response and interrogation are frequently referred to as skip patterns.

A natural mathematical structure for the analysis of skip patterns and underlying structure of questionnaires is the directed graph. In this paper, we will exploit the graph theoretic structure inherent in questionnaires in order to adjust for missing data: that is, we attempt to recognize questions that should have been answered but were left blank and based on the pattern of responses, differentiate them from non-applicable questions.

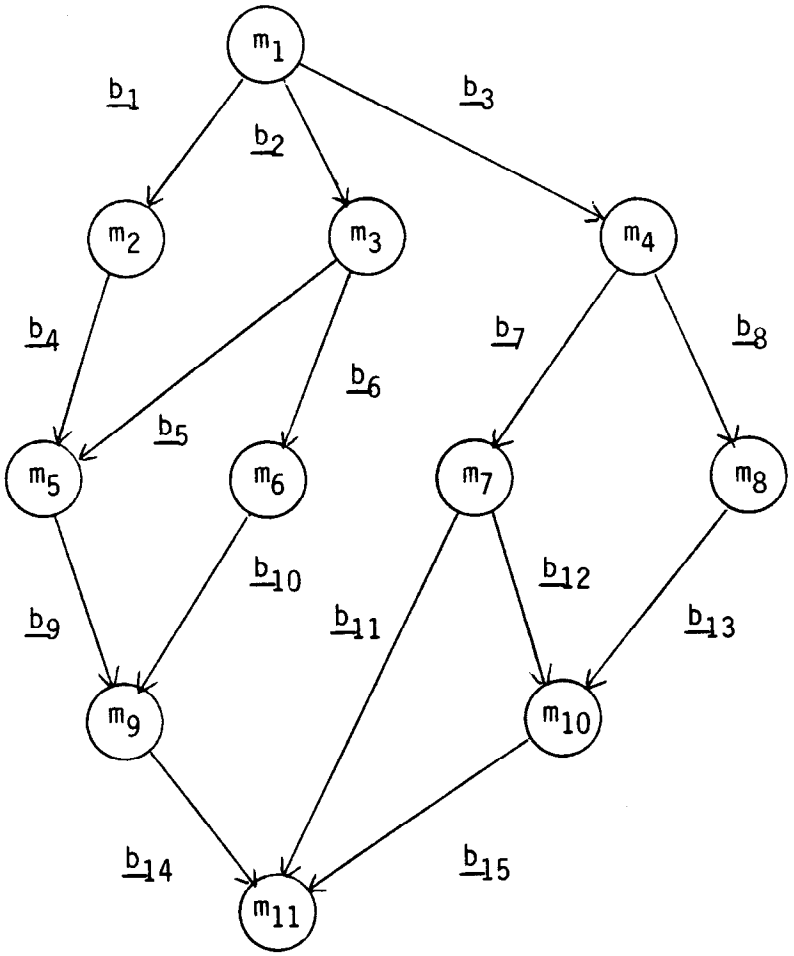
In Section II, we introduce basic definitions and terminology of graph theory, establish results that will be needed in the later sections when applying graph theoretic analysis to questionnaire forms, and provide examples. In Section III, we show how a questionnaire can be viewed as a graph and examine its properties from this perspective. In the remaining sections, we show how this structure can be employed to impute for missing data, analyze skip patterns, and edit questionnaires. In Section IV, we describe two computer programs that implement the techniques developed in the earlier section, and Section V is a summary. In Appendix I and Appendix II we provide samples of output from the computer program based on examples discussed in this report.

II. DIRECTED GRAPHS

A. Basic Definitions and Examples

A (directed) graph is a pair G = (N,A) where N is a finite set, call the set of nodes, and A is a set of ordered pairs of nodes whose elements are called arcs. A typical arc, a, is written as a = (n_i, n_j) where n_i and n_j are in N. We say that a is an arc from node n_i to node n_j and that n_i and n_j are the endpoints of a.

Example 1:



In this graph, the node set is N = $\{m_i: i=1,\ldots,11\}$ and the arc set is $\{b_j: j=1,\ldots,15\}$. Note that $b_9 = (m_5, m_9)$ and also note that (m_3, m_6) is an arc but (m_6, m_3) is not an arc.

A chain (of length r) is a sequence of arcs and nodes;

n_{k_i}, $\underline{a}_{k_1}$, n_{k_2}, $\underline{a}_{k_2}$, n_{k_3},...,$\underline{a}_{k_r}$, $n_{k_{r+1}}$; such that the endpoints of $\underline{a}_{k_i}$ are n_{k_i} and $n_{k_{i+1}}$ for i=1,...,r. In Example 1, the sequence: m_1, $\underline{b}_2$, m_3, $\underline{b}_6$, m_6, $\underline{b}_{10}$, m_9 is a chain. We will at times refer to this chain as a set of nodes: { m_1, m_3, m_6, m_9 } or as a set of arcs: {$\underline{b}_2$, $\underline{b}_6$, $\underline{b}_{10}$} depending upon the context. The node n_{k_1} (resp., arc $\underline{a}_{k_1}$) is called the initial node (resp., initial arc) of the chain and the node $n_{k_{r+1}}$ (resp., arc $\underline{a}_{k_r}$) is called the terminal node (resp., arc). In the chain above, m_1 is the initial node and m_9 is the terminal node, and $\underline{b}_2$ is the initial arc and $\underline{b}_{10}$ is the terminal arc.

A path is a chain in which all nodes are distinct, and the example of a chain above is also a path. A graph in which every chain is a path is called acyclic, (that is, the graph contains no cycles). A path properly contained in no other path is called a maximal path. That is, m_3, m_6 m_9 is a path which is not maximal yet m_1, m_3, m_6, m_9, m_{11} is a maximal path. We end with one crucial assumption: for all graphs to be considered here, there exists a unique node n_s and a unique node n_t such that every maximal path has n_s as initial point and n_t as terminal point. In terms of Example 1, m_1 is our unique initial point, and m_{11} is our unique terminal point. These points will be called, respectively, a source and a sink, and throughout this report, we will deal exclusively with acyclic directed graphs with a source and sink, see [2]. For the purposes of this report, we will call the source our initial point and the sink our terminal point.

If n_i and n_j are nodes (resp., $\underline{a}_i$ and $\underline{a}_j$ are arcs) of a graph, and if there exists a path with initial node n_i (resp., arc $\underline{a}_i$) and terminal node n_j (resp., arc $\underline{a}_j$), we say we have a path from n_i to n_j (resp., $\underline{a}_i$ to $\underline{a}_j$). If we have a path from n_i to n_j (resp. $\underline{a}_i$ to $\underline{a}_j$) we say that n_i preceeds n_j (resp. $\underline{a}_i$ preceeds $\underline{a}_j$) and that n_j succeeds n_i. If there is an arc from n_i to n_j, we say that n_i is an immediate predecessor to n_j and that n_j is an immediate successor to n_i.

Definitions: Let G = (N, A) be a graph and let $n \in N$, then

(a) the (node) cover of n, denoted by $C_N(n)$, is the set of immediate successors to n,

(b) the lower (node) ideal of n, denoted by $L_N(n)$, is the set of successors to n along with the node n itself,

(c) the upper (node) ideal of n, denoted by $U_N(n)$, is the set of predecessors of n along with the node n itself,

(d) the (node) ideal of n, denoted by $B_N(n)$, is the union of $L_N(n)$ and $U_N(n)$.

Definition: Let G = (N,A) be a graph and let $\underline{a} \in A$, then

(a) the (arc) cover of $\underline{a}$, denoted by $C_A(\underline{a})$ is the set of immediate successors to $\underline{a}$,

(b) the lower (arc) ideal of $\underline{a}$, denoted by $L_A(\underline{a})$, is the set of successors to $\underline{a}$ along with the arc $\underline{a}$ itself,

(c) the upper (arc) ideal of $\underline{a}$, denoted by $U_A(\underline{a})$, is the set of predecessors of $\underline{a}$ along with the arc $\underline{a}$ itself,

(d) the (arc) ideal of $\underline{a}$, denoted by $B_A(\underline{a})$, is the union of $L_A(\underline{a})$ and $U_A(\underline{a})$.

Notation: When G = (N,A) is a graph, we will denote the cardinality of N by **N** and the cardinality of A by **A**.

Example 2: Referring to the graph of Example 1, we have

$$C_N(m_7) = \{m_{10}, m_{11}\}$$

$$L_N(m_7) = \{m_7, m_{10}, m_{11}\}$$

$$U_N(m_7) = \{m_1, m_4, m_7\}$$

$$B_N(m_7) = \{m_1, m_4, m_7, m_{10}, m_{11}\}$$

$$C_A(\underline{b}_9) = \{\underline{b}_{14}\}$$

$$L_A(\underline{b}_9) = \{\underline{b}_9, \underline{b}_{14}\}$$

$$U_A(\underline{b}_9) = \{\underline{b}_1, \underline{b}_2, \underline{b}_4, \underline{b}_5, \underline{b}_9\}$$

$$B_A(\underline{b}_9) = \{\underline{b}_1, \underline{b}_2, \underline{b}_4, \underline{b}_5, \underline{b}_9, \underline{b}_{14}\}.$$

B. Representing Graphs by Matrices

If G = (N,A) is a graph, we can represent G by a matrix, c_N, often called the (node) incidence matrix. This matrix is square, the rows and columns are both indexed by N, and

$$c_N(i,j) = \begin{cases} 1 & \text{if there is an arc from node } n_i \text{ to node } n_j \\ 1 & \text{if } i=j \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for the graph is Example 1, c_N is:

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}
m_1	1	1	1	1							
m_2		1			1						
m_3			1		1	1					
m_4				1			1	1			
m_5					1				1		
m_6						1			1		
m_7							1			1	1
m_8								1		1	
m_9									1		1
m_{10}										1	1
m_{11}											1

For an arbitrary node, n, the cover of n, $C_N(n)$ corresponds to the positive positions in row n of $(c_N - I)$ where I is the NxN identity matrix. We can also define an (arc) incidence matrix of G, c_A, as the square matrix whose rows and columns are indexed by the arcs of G. We define

$$c_A(i,j) = \begin{cases} 1 & \text{if the terminal node of arc } \underline{a}_i \\ & \text{is the initial node of arc } \underline{a}_j \\ 1 & i=j \\ 0 & \text{otherwise.} \end{cases}$$

For the graph in Example 1, we obtain the following arc incidence matrix below. As with the node incidence matrix the cover of arc $\underline{a}$ corresponds to the positive elements in row $\underline{a}$ of $(c_A - I)$ where I is the AxA identity matrix.

Arc Incidence Matrix for Example 1

	$\underline{b}_1$	$\underline{b}_2$	$\underline{b}_3$	$\underline{b}_4$	$\underline{b}_5$	$\underline{b}_6$	$\underline{b}_7$	$\underline{b}_8$	$\underline{b}_9$	$\underline{b}_{10}$	$\underline{b}_{11}$	$\underline{b}_{12}$	$\underline{b}_{13}$	$\underline{b}_{14}$	$\underline{b}_{15}$
$\underline{b}_1$	1			1											
$\underline{b}_2$		1			1	1									
$\underline{b}_3$			1				1	1							
$\underline{b}_4$				1					1						
$\underline{b}_5$					1				1						
$\underline{b}_6$						1				1					
$\underline{b}_7$							1				1	1			
$\underline{b}_8$								1					1		
$\underline{b}_9$									1					1	
$\underline{b}_{10}$										1				1	
$\underline{b}_{11}$											1				
$\underline{b}_{12}$												1			1
$\underline{b}_{13}$													1		1
$\underline{b}_{14}$														1	
$\underline{b}_{15}$															1

If G = (N,A) is a graph, we can also represent G by a matrix, b_N, often called the (node) ideal matrix. The matrix is square, both rows and columns are indexed by N, and

$$b_N(i,j) = \begin{cases} 1 & \text{if node } n_i \text{ preceeds node } n_j \\ 1 & i=j \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for the graph in Example 1, b_N is:

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}
m_1	1	1	1	1	1	1	1	1	1	1	1
m_2		1			1				1		1
m_3			1		1	1			1		1
m_4				1			1	1		1	1
m_5					1				1		1
m_6						1			1		1
m_7							1			1	1
m_8								1		1	1
m_9									1		1
m_{10}										1	1
m_{11}											1

For an arbitrary node, n, the lower ideal, $L_N(n)$ corresponds to the positive positions in row n and the upper ideal, $U_N(n)$ corresponds to the positive entries in column n.

We can also define an (arc) ideal matrix, b_A, which is also an upper triangular square matrix whose rows and columns are indexed by the arcs of G. We define

$$b_A(i,j) = \begin{cases} 1 & \text{if arc } \underline{a}_i \text{ proceeds arc } \underline{a}_j \\ 1 & \text{if } i=j \\ 0 & \text{otherwise.} \end{cases}$$

For the graph in Example 1, we obtain the arc ideal matrix below. As with the node matrix for an arc $\underline{a}$, the lower arc ideal corresponds to the positive entries in row $\underline{a}$, and the upper arc ideal corresponds to the positive entries in column $\underline{a}$.

Arc Ideal Matrix for Example 1

	$\underline{b}_1$	$\underline{b}_2$	$\underline{b}_3$	$\underline{b}_4$	$\underline{b}_5$	$\underline{b}_6$	$\underline{b}_7$	$\underline{b}_8$	$\underline{b}_9$	$\underline{b}_{10}$	$\underline{b}_{11}$	$\underline{b}_{12}$	$\underline{b}_{13}$	$\underline{b}_{14}$	$\underline{b}_{15}$
$\underline{b}_1$	1			1					1					1	
$\underline{b}_2$		1			1	1			1	1				1	
$\underline{b}_3$			1				1	1			1	1	1		1
$\underline{b}_4$				1					1					1	
$\underline{b}_5$					1				1					1	
$\underline{b}_6$						1				1				1	
$\underline{b}_7$							1				1	1			1
$\underline{b}_8$								1					1		1
$\underline{b}_9$									1					1	
$\underline{b}_{10}$										1				1	
$\underline{b}_{11}$											1				
$\underline{b}_{12}$												1			1
$\underline{b}_{13}$													1		1
$\underline{b}_{14}$														1	
$\underline{b}_{15}$															1

Remark: Starting with the cover matrix, c_N, one can easily derive the ideal matrix, b_N. If we let C be the **N**x**N** matrix:

$$C(i,j) = \begin{cases} 1 & \text{if } c_N^{(\mathbf{N})}(i,j) > 1 \\ 0 & \text{otherwise,} \end{cases}$$

it is not hard to show that $b_N = C$, where **N** is the number of nodes and $c_N^{(\mathbf{N})}$ is the $\mathbf{N}^{th}$ power of c_N.

One can, however, introduce a number of programming simplifications. If we let $D_{(1)} = c_N$, and

$$D_{(k+1)}(i,j) = \begin{cases} 1 & \text{if } (c_N D_{(k)})(i,j) > 1 \\ 0 & \text{otherwise,} \end{cases}$$

for k=1,...,**N**-1, one observes that $b_N = D_{(\mathbf{N})}$. Thus, we reduce the problem of finding b_N to one of multiplying two zero-one matrices. Note further that if we had let k above range from 1 to M where M **N**, then $D_{(\mathbf{N})} = D_{(\mathbf{N}+1)}$. Accordingly, we could have defined $E_{(k)}$ to be a family of NxN zero-one matrices where $E_{(1)} = c_N$ and

$$E_{(k+1)}(i,j) = \begin{cases} 1 & \text{if } (E_{(k)}E_{(k)})(i,j) > 1 \\ 0 & \text{otherwise.} \end{cases}$$

If $k > \log_2 N$, then $b_N = E_{(k)}$. Thus we need only form $E_{(k+1)}$ for k between 1 and K where K is the smallest integer greater than or equal to $\log_2$ **N**. For example, if there were 1000 questions on a questionnaire, we would only have to perform 10 multiplications. More typically, if a questionnaire has only 64 questions, we would have to perform only 6 multiplications. Of course, since we are only dealing with zero-one matrices considerable reductions can be realized through bit minipulation and other simplifications. The point is that by merely entering the cover matrix, one very easily can obtain the node ideal matrix, b_N. Similar observations apply to the arc cover and ideal matrices.

Remark: It can be conceptually convenient to think of the (node) bi-ideal matrix, J_N, of a graph G. The matrix, J_N, is a square matrix, both rows and columns are indexed by N, and

$$J_N(i,j) = \begin{cases} 1 & \text{if } b_N(i,j) = 1 \text{ or } b_N(j,i) = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Note that J_N is a symmetric matrix and that node i is in the ideal of node j if and only if, $J_N(i,j) \neq 0$.

If $\{\underline{v}_i\}_{i\in I}$ is a finite set of zero-one vectors (that is, each component is a zero or one) and $\underline{v}_j = (v_{j1}, v_{j2}, \ldots, v_{jN})$, we can define the product $\prod_{i\in I} \underline{v}_i$ to be the zero-one vector $\underline{u} = (u_1, \ldots, u_N)$ where $u_j = \prod_{i\in I} v_{ij}$ for all $j=1,\ldots,N$. If r_i denotes the i^{th} row of the node bi-ideal matrix, J_N, then the non-zero elements in the vector r_i correspond to nodes in the ideal of node n_i. If $\{n_i\}_{i\in I}$ is a set of nodes, then the non-zero coordinates of $\prod_{i\in I} r_{n_i}$ corresponds to nodes in all of the ideals $B_N(n_i)$ for $i\in I$. That is, if $\{n_i\}_{i\in I}$ is an arbitrary set of nodes, then $\bigcap_{i\in I} B_N(n_i)$ corresponds to coordinates equal to one in each of the rows r_{n_i}. Identical considerations apply in finding the intersection of a set of arc ideals as well as upper and lower ideals.

C. Further Definitions Needed for Skip Patterns

Definition: Let G = (N,A) be a graph, M N, and S A, we say that S is a consistent set of arcs if $\underline{a} \in B_A(\underline{b})$ for all $\underline{a},\underline{b} \in S$. We say that M is a consistent set of nodes if $m \in B_N(n)$ for all $m, n \in M$. A set of arcs or nodes that is not consistent is called inconsistent.

Example 3: Referring to the graph of Example 1, the set of arcs

$$S = \{\underline{b}_3, \underline{b}_{11}, \underline{b}_{15}\}$$

is not consistent, but the set

$$S = \{\underline{b}_2, \underline{b}_5, \underline{b}_9, \underline{b}_{14}\}$$

is consistent.

If S is a consistent subset of arcs, one can order the elements of $S = \{\underline{s}_i \mid = i=1,\ldots,n\}$ such that $\underline{s}_i$ proceeds $\underline{s}_{i+1}$ for $i=1,\ldots,n-1$. For any $k=1,\ldots,n-1$, if the terminal point of $\underline{s}_k$ is different from the initial point of $\underline{s}_{k+1}$ there exists a chain of arcs $\underline{t}_1,\ldots,\underline{t}_{n_k}$ such that the initial point of $\underline{t}_1$ equals the terminal point of $\underline{s}_k$, and the terminal point of $\underline{t}_{n_k}$ equals the initial point of $\underline{s}_{k+1}$. Thus, every consistent set of arcs can be embedded in a chain C. If either the source or sink of G is not in C, we can form a chain from the source to $\underline{s}_1$ and a chain from $\underline{s}_n$ to the sink thus embedding the original consistent set of arcs in a maximal chain. Thus we have shown the following.

Proposition 1: Let G = (N,A) be a graph and $S \subset A$. The set S is consistent if and only if there exists at least one maximal chain containing S.

Definition: If G = (N,A) is a graph and $S \subset A$, define

$$B(S) = \bigcap_{\underline{s} \in S} B_A(\underline{s}).$$

Example 4: Referring to Example 1, let

$$S = \{\underline{b}_2, \underline{b}_{14}\}.$$

Then $B(S) = \{\underline{b}_2, \underline{b}_5, \underline{b}_6, \underline{b}_9, \underline{b}_{10}, \underline{b}_{14}\}$

and $B(B(S)) = \{\underline{b}_2, \underline{b}_{14}\}$.

Remark: The set B(S) need not be consistent even if S is consistent as can be seen from Example 4.

Proposition 2: Let G = (N,A) be a graph and $S \subset A$.

(i) B(S) consists of all arcs $\underline{a} \in A$ such that $\{\underline{a}, \underline{s}\}$ is consistent for all $\underline{s} \in S$.

(ii) If S is consistent, the B(S) consists of all arcs $\underline{a} \in A$ such that $\{\underline{a}\} \cup S$ is consistent.

(iii) If S is consistent, then for $\underline{a} \in B(S)$ there exists a maximal chain containing $\{\underline{a}\} \cup S$.

(iv) The set S is consistent if and only if $S \subset B(S)$. Accordingly, every maximal chain containing S is contained in B(S).

Proof:

(i) If $\underline{a} \in B(S)$, then $\underline{a} \in B_A(\underline{s})$ for all $\underline{s} \in S$, so $\{\underline{a}, \underline{s}\}$ is consistent. If $\{\underline{a}, \underline{s}\}$ is consistent for all $\underline{s} \in S$, then $\underline{a} \in B_A(\underline{s})$ for all $\underline{s} \in S$ so $\underline{a} \in B(S)$.

(ii) Follows from (i).

(iii) Follows from (ii) and Proposition 1.

(iv) If S is consistent, and $\underline{t} \in S$, then $\{\underline{t}, \underline{s}\}$ is consistent for all $\underline{s} \in S$, so $\underline{t} \in B(S)$ by (i). If $S \subset B(S)$ and $\underline{s}, \underline{t} \in S$, then $\{\underline{s}, \underline{t}\}$ is consistent since $\underline{t} \in B(S)$, hence $\underline{t} \in B_A(\underline{s})$. The last assertion follows from (i).

Proposition 3: Let G = (N,A) be a graph and S A be a consistent set of arcs. The set B(S) is consistent if and only if B(S) is a maximal chain. In this case, B(S) is the unique maximal chain containing S.

Proof: Assume B(S) is consistent and let C be a maximal chain containing B(S). Since S is consistent, $S \subset B(S)$ and so $S \subset C$. Hence, $C \subset B(S)$ by Proposition 2 (iv), thus $C = B(S)$. Going the other direction, we need only note that every chain is consistent.

Lemma 1: Let $G = (N,A)$ be a graph and $E \subset F \subset A$.

(i) $B(F) \subset B(E)$.

(ii) If F is consistent, then E is consistent and $F \subset B(E)$.

Proof:

(i) $$B(F) = \bigcap_{f \in F} B_A(f) \subset \bigcap_{e \in E} B_A(e) = B(E) .$$

The containment holds since the index of the second intersection is a subset of the index of the first

(ii) It is clear that E is consistent. Thus $E \subset B(E)$ (Proposition 2, (iv)) and $F \subset B(F)$ by (i) above.

Proposition 4: Let $G = (N,A)$ be a graph and $S \subset A$. Then

(i) $S \subset B(B(S))$,

(ii) $B(S) = B(B(B(S)))$,

(iii) if S is consistent, then $B(B(S)) \subset B(S)$.

Proof:

(i) $\underline{q} \in B(B(S))$ if $\{\underline{t}, \underline{q}\}$ is consistent for all $\underline{t} \in B(S)$. But, if $\underline{s} \in S$, then $\{\underline{s}, \underline{t}\}$ is consistent for all $\underline{t} \in B(S)$. Thus $\underline{s} \in B(B(S))$ and so $S \subset B(B(S))$.

(ii) By (i) and the Lemma, $B(B(B(S))) \subset B(S)$. Since $T \subset B(B(T))$ for all subsets $T \subset A$, letting $T = B(S)$, we have by (i) $B(S) \subset B(B(B(S)))$. The result follows.

(iii) Since S is consistent, $S \subset B(S)$, so $B(B(S)) \subset B(S)$ by the Lemma.

Proposition 5: Let $G = (N,A)$ be a graph and $S \subset A$. The set S is consistent if and only if $B(B(S))$ is consistent.

Proof: If $B(B(S))$ is consistent, then according to Lemma 1 and Proposition 4, S is consistent since $S \subset B(B(S))$. If S is consistent, then $B(B(S)) \subset B(S)$ by Proposition 4. Let $\underline{b}_1, \underline{b}_2 \in B(B(S))$. Since $\underline{b}_1 \in B(S)$ and $\underline{b}_2 \in B(B(S))$, $\{\underline{b}_1, \underline{b}_2\}$ is consistent. That is, $B(B(S))$ is consistent.

Proposition 6: Let $G = (N,A)$ be a graph, $S \subset A$ a consistent set of arcs, and C^* the family of all maximal chains containing S. Then

(i) $\bigcup_{C \in C^*} C = B(S)$,

(ii) $\bigcap_{C \in C^*} C = B(B(S))$.

Proof:

(i) Follows from Proposition 2, parts (iii) and (iv).

(ii) Let $C \in C^*$. By repeated applications of Lemma 1, we have $B(B(S)) \subset B(B(C))$. Since C is a maximal chain, by Proposition 2, we have $C = B(C) = B(B(C))$; that is,

$$B(B(S)) \subset \bigcap_{C \in C^*} C .$$

Let $\underline{c} \in \bigcap_{C \in C^*} C$, and let $\underline{b} \in B(S)$ be arbitrary. There exists a chain containing $\underline{b}$ and S, so $\{\underline{b}\} \cup S$ is consistent. Thus, there exists a chain $C_1 \in C^*$ such that $\{\underline{b}\} \cup S \subset C_1$. But $\underline{c} \in C_1$ and hence $\{\underline{b}, \underline{c}\}$ is consistent. Since $\underline{b} \in B(S)$ was arbitrary, $\underline{c} \in B(B(S))$, hence

$$\bigcap_{C \in C^*} C \subset B(B(S)) .$$

Corollary: Let G = (N,A) be a graph and let $S \subset A$ be a consistent set of arcs. Then B(B(S)) consists of those arcs that must be in every maximal chain containing S.

Definition: Let G = (N,A) be a graph and $S \subset A$. Define P(S) to be the set of nodes occuring either as an initial or terminal point of some arc in S.

Proposition 7: Let G = (N,A) be a graph and $S \subset A$ a consistent set of arcs. Then P(S) is a consistent set of nodes and every node in P(B(S)) is consistent with every node in P(S).

Proof: Clear.

Example 5: Referring to Example 1, let

$$S = \{\underline{b}_2, \underline{b}_{14}\}.$$

Then $P(S) = \{m_1, m_3, m_9, m_{11}\}$

and $P(B(S)) = \{m_1, m_3, m_5, m_6, m_9, m_{11}\}$.

Definition: Let G = (N,A) be a graph and $S \subset A$ be a consistent set of arcs. We say that $w \in N$ is a waist point of G if every maximal chain in G contains w. We say that $w_S \in N$ is a waist point of S if every maximal chain containing S also contains w_S. We denote the waist points of G by W(G) and the waist points of S by W(S).

Remark: Looking ahead to the graph drawn in Figure 2 (page 26), the only waist nodes (other than the source and sink) are node m_{11} and m_{13}.

Lemma 2: Let G = (N,A) be a graph and let C be a maximal chain in G. Denoting the points of C by P(C), then

$$\bigcap_{\underline{s} \in C} P(I_A(\underline{s})) = P(C) .$$

Proof: Let $p \in P(C)$ and let $\underline{t} = (p,q) \in C$. Then $\{\underline{s},\underline{t}\}$ is consistent for all $\underline{s} \in C$, so $\underline{t} \in I_A(\underline{s})$ for all $\underline{s} \in C$, so $\underline{t} \in \bigcap_{\underline{s}\in C} P(I_A(\underline{s}))$, hence

$$\bigcap_{\underline{s}\in C} P(I_A(\underline{s})) \supset P(C).$$

To go the other way, assume $p \in N$, $p \notin P(C)$. After relabeling, represent the chain C by the sequence of nodes and arcs: $n_1, \underline{a}_1, n_2, \underline{a}_2, \ldots, \underline{a}_{k-1}, n_k$, where $\underline{a}_i = (n_i, n_{i+1})$ for $i=1,..,k-1$. Note that both $C \cap U_N(p)$ and $C \cap L_N(p)$ are non-empty. Let

$$\alpha = \text{Max}\{i \mid n_i \in C \cap U_N(p)\}$$

$$\beta = \text{Min}\{i \mid n_i \in C \cap L_N(p)\}.$$

Since G is an acyclic directed graph, $\alpha < \beta$ and we consider the arc $\underline{a}_\alpha = (n_\alpha, n_{\alpha+1})$. We will show that $p \notin P(I_A(\underline{a}_\alpha))$ by writing

$$P(I_A(\underline{a}_\alpha)) = U_N(n_\alpha) \cup L_N(n_{\alpha+1}).$$

If p were in $U_N(n_\alpha)$, then $n_\alpha \in L_N(p)$, which is impossible since

$$\alpha < \beta = \text{Min}\{i \mid n_i \in C \cap L_N(p)\}.$$

If p were in $L_N(n_{\alpha+1})$, then $n_{\alpha+1} \in U_N(p)$, which is impossible since

$$\alpha+1 > \alpha = \text{Max}\{i \mid n_i \in C \cap U_N(p)\}.$$

Thus, $p \notin P(I_A(\underline{a}_\alpha))$, so $p \notin \bigcap_{\underline{s}\in C} P(I_A(\underline{s}))$. Hence $\bigcap_{s\in C} P(I_A(\underline{s}) \subset P(C)$.

Theorem 1: Let $G = (N,A)$ be a graph and let C^* be the family of maximal chains in G. Then

$$W(G) = \bigcap_{C\in C^*} P(C) = \bigcap_{\underline{s}\in A} P(I_A(\underline{s})).$$

Proof: The first equality is the definition of W(G), so we need only show the second.

Let $p \in \bigcap_{C \in C^*} P(C)$, and let $\underline{s} \in A$ be arbitrary. Since $I_A(\underline{s})$ contains at least one maximal chain, C_1, $p \in P(C_1) \subset P(I_A(\underline{s}))$, and since $\underline{s}$ was arbitrary,

$$\bigcap_{C \in C^*} P(C) \subset \bigcap_{\underline{s} \in A} P(I_A(\underline{s})).$$

Let $p \in \bigcap_{\underline{s} \in A} P(I_A(\underline{s}))$ and let C_1 be an arbitrary maximal chain. Then

$$\bigcap_{\underline{s} \in A} P(I_A(\underline{s})) \subset \bigcap_{\underline{s} \in C_1} p(I_A(\underline{s})) = P(C_1)$$

by Lemma 2. Since C_1 was arbitrary.

$$\bigcap_{\underline{s} \in A} P(I_A(\underline{s})) \subset \bigcap_{C \in C^*} P(C)$$

and the result is proven.

Lemma 3: Let $G = (N,A)$ be a graph, $S \subset A$ a consistent subset of arcs, and C a maximal chain containing S. Then

$$\bigcap_{\underline{s} \in C} [P(I_A(\underline{s})) \cap P(B(S))] = \bigcap_{\underline{s} \in C} P(I_A(\underline{s})).$$

Proof: Note first that

$$\bigcap_{\underline{s} \in C} [P(I_A(\underline{s})) \cap P(B(S))] = P(B(S)) \cap \bigcap_{\underline{s} \in C} P(I_A(\underline{s})) \quad .$$

But

$$P(B(S)) \cap [\bigcap_{\underline{s} \in C} P(I_A(\underline{s}))] = \bigcap_{\underline{s} \in C} P(I_A(\underline{s}))$$

since

$$\bigcap_{\underline{s} \in C} P(I_A(\underline{s})) = P(C) \subset P(B(S))$$

by Proposition 2(iv).

Theorem 2: If $G = (N,A)$ is an acyclic directed graph with a source and sink and $S \subset A$ is a consistent set of arcs, then $P(S) \subset P(B(S))$ and

$G_S = (P(B(S)), B(S))$ is an acyclic directed graph with source and sink which is a subgraph of $G = (N,A)$. Furthermore, G and G_S have the same source and sink. The node incidence matrix of G_S is the node matrix of G with rows and columns deleted corresponding to nodes not in $P(B(S))$. The same applies to the arc incidence matrix. The waist points of G_S correspond to the waist points of S.

Theorem 3: Let $G = (N,A)$ be a graph, $S \subset A$ a consistent subset of arcs, and C^* the family of maximal chains containing S. Then

$$W(S) = \bigcap_{C \in C^*} P(C) = \bigcap_{\underline{s} \in B(S)} P(I_A(\underline{s})).$$

Proof: By observing that $C \subset B(S)$ for all $C \in C^*$, this theorem then can be proved in a manner similar to the proof of Theorem 2. An alternate approach is to appeal to Theorem 2 and view this result as a corollary.

Remark: It is Theorem 3 that provides a computationally efficient procedure to find all the waist points of a consistent set of arcs, S. Instead of considering all maximal chains containing S, one needs only consider rows and columns of the ideal matrix b_A of G corresponding to elements of S.

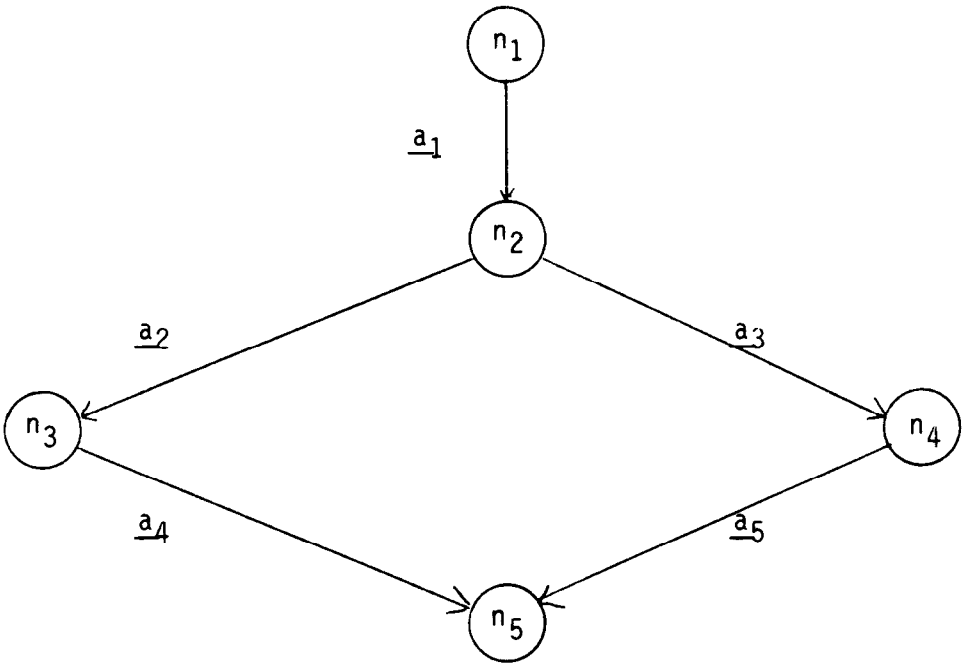
Remark: The results in Theorems 1 and 3 and some of the preceeding discussion might suggest that if S is a consistent set of arcs then

$$\bigcap_{\underline{s} \in B(S)} P(I_A(\underline{s}))$$

is equal to

$$P\left[\bigcap_{\underline{s} \in B(S)} I_A(\underline{s})\right].$$

This is not the case however as has been seen from the following example. Let G be represented by the graph:



Let the consistent set $S = \{\underline{a}_1\}$, so $B(S) = A$, the full arc set. Then

$$\bigcap_{\underline{s}\in B(S)} P(I_A(\underline{s})) = \{n_1, n_2, n_5\}$$

and

$$P\left(\bigcap_{\underline{s}\in B(S)} I_A(\underline{s})\right) = \{n_1, n_2\} .$$

Remark: Let $G = (N,A)$ be a graph and let $S \subset A$ be a consistent set of arcs. We can think of $B(S)$ as the saturation of S and

$$G_S = (P(B(S)), B(S))$$

as the localization of G at S. There is a one-one, onto correspondence between the maximal chains of G containing S (hence contained in $B(S)$) and the maximal chains in G_S. In particular, if S is a maximal consistent set of arcs, then G_S consists of a single (maximal) chain.

Remark: If $G = (N,A)$ is an acyclic directed graph with source and sink, we can induce a partial ordering on the arcs of G. We say that for $\underline{a}$, $\underline{b} \in A$,

$$\underline{a} \prec \underline{b} \quad \text{if} \quad \underline{a} \in L_A(\underline{b}).$$

Viewed this way, a set of arcs $S \subset A$ is consistent if and only if all elements of S are comparable under the partial ordering. Furthermore, each element of B(S) is comparable with every element of S and every subset of A all of whose elements are comparable with S are contained in B(S).

Remark: It is quite easy to determine if $S \subset B(S)$ by examing zero-one vectors; and hence, whether a set of arcs is consistent. In addition, given a set of arcs, S, one can easily find P(S), P(B(S)), B(B(S)) as well as all other constructs needed for the skip pattern analysis by using the rows and columns of the ideal matrix manipulating zero-one vectors.

III. GRAPHS AND QUESTIONNAIRES

A. Deriving a Graph from a Questionnaire

In most survey questionnaire forms, questions are not answered in a strictly linear fashion. That is, based on the response to a question, if the respondent answered one way he/she would be asked one subsequent question yet if he/she responded differently the respondent would be asked a different following question.

Example 6: (Taken from Wave I Questionnaire from the Survey of Income and Program Participation):

Question 3a.	Responses	
Were there any weeks in the	- Yes	(Skip to 3c)
4-month period when ... wanted a job?	- No	(Skip to 9a)

The rules guiding the sequence of questions asked based on responses furnished is often referred to as a skip pattern. In the analysis of a questionnaire both for purposes of design and subsequent data analysis, the underlying skip pattern plays a major role. In this report, we will show how the analysis of a questionnaire skip pattern can be used in the area of edit and imputation.

Our first objective is to show how a questionnaire can be represented as an acyclic (directed) graph with source and sink. For each question on a questionnaire we associate a node in an associated graph, that is, if $\{Q_i\}$ i=1,...,n is the set of questions, we have a set of nodes, $\{N_i\}$ i=1,...,n where question Q_i corresponds to node N_i. If some response to question Q_t allows a set of questions $\{Q_s\}_{s\varepsilon S}$ to be asked next, we set up the arc from node N_t to each node $\{N_s\}_{s\varepsilon S}$. We identify each possible response to question Q_t with an arc from node N_t. In order to make this correspondence complete, we must introduce one "dummy question" on the questionnaire. This dummy question will require no response, and it is viewed as following all otherwise final questions on the questionnaire. That is, if any question has, in fact, no follow-up on the questionnaire itself and would terminate an interview, we act as if after asking a terminal question we "skip" to the dummy question. By adding this dummy question, we have a map from the set of questionnaires to the set of acyclic directed graphs having a source and sink, where for each questionnaire, questions correspond to nodes and response choices correspond to arcs. The first and last questions correspond to the source and sink (initial and terminal node) of the graph. We will have exactly one more node on the graph than questions on the questionnaire (due to the dummy question corresponding to the terminal node).

For each question there may be several responses that require the same "next" question to be asked. When this occurs, we treat each of these responses as being equivalent and recode them to so indicate. Thus, in the map from questionnaires to graphs as described above each equivalence class of responses corresponds to a single arc and questions correspond to nodes.

Example 7: (Taken from Wave I Questionnaire from the Survey of Income and Program Participation)

Question R11	Responses	
What is ...'s martial status?	-- Married	(Skip to 17)
	-- Widowed	(Skip to 19a)
	-- Never Married	(Skip to R12)
	-- Divorced	(Skip to 18)
	-- Separated	(Skip to 18)

Note that responses "divorced" or "separated" both direct next asking question Question 18, and are treated as equivalent for skip pattern analysis. They both correspond to the arc from Question R11 to Question 18.

Example 8: The following is a slightly modified extract from the 1979 Research Panel, Income Survey Development Program (ISDP) questionnaire.

Questions	Responses	
2a. During the period outlined on this calendar, did... do any work at a job or a business?	-- Yes	(Skip to 3a)
	-- No	(Skip to 2c)
2c. Did...do any temporary, part-time, or seasonal work even for a few days during this period?	-- Yes	(Skip to 3a)
	-- No	(Skip to 2e)
2e. What were the main reasons... did not work at a job during this 3-month period?	-- Taking care of home and family	(Skip to 2f)
	-- Going to school	
	-- Could not find work	
	-- Didn't want to work	
	-- Retired	
	-- Too old to work	
	-- Ill, injured, or disabled	
	-- Other - Specify	
2f. During the 3-month period did...spend any time looking for work?	-- Yes	(Skip to 2g)
	-- No	(Skip to 2h)
2g. How many weeks did...spend looking for work?	-- Number of weeks	(Skip to 3d)
	-- All	
	-- DK	

2h. Did... want a regular job, either full or part time, at any time during this period?	-- Yes	(Skip to 2i)
	-- No -- DK	(Skip to 3d)
2i. What were the reasons... did not look for work during this period?	-- Believes no work available in line of work or area -- Couldn't find any work -- Lacks necessary schooling, training experience -- Employers think too old or too young -- Can't arrange child care -- Family responsibilities -- Going to school -- Other	(Skip to 3d)
3a. Were there any full weeks during this 3-month period in which...did not have a job or business (exclude temporary layoff)?	-- Yes	(Skip to 3b)
	-- No -- DK	(Skip to 3d)
3b. During the weeks when...did not have a job or business, did...spend any time looking for work?	-- Yes	(Skip to 3c)
	-- No	(Skip to 3d)
3c. How many weeks did..spend looking for work?	-- Number of weeks -- All -- DK	(Skip to 3d)
3d. During the 3-month period did...receive any unemployment or other compensation because of layoff, slack work, or strike?	-- Yes	(Skip to 3e)
	-- No	(Skip to 3f)

3e. What was the source of this compensation?	-- Unemployment from the State or local unemployment office -- Supplemental Unemployments Benefits -- Union strike benefits -- Other - Specify	(Skip to 3f)
3f. Did...receive any income during these 3 months to make up for pay lost because of illness or injury?	-- Yes -- No -- DK	(Skip to 3g) (Skip to END)
3g. What was the source of this income?	-- Worker's compensation -- State temporary sickness or disability -- Own accident, disability, sickness insurance policy -- Other source or don't know	(Skip to END)
End (Dummy Question)	- Final node for this segment (no response required)	

When we draw the graph associated with this questionnaire, we get Figure 1. By renumbering the nodes and arcs we have the graph in Figure 2. Note, for example, that node m_9 in Figure 2 corresponds to question 3.b in the questionnaire, and arc a_{13} in Figure 2 corresponds to a "Yes" response to question 3.b. Note also that there are fourteen genuine questions on the questionnaire but fifteen nodes on the graph. The fifteenth node corresponds to the dummy question. Following these figures we include the node and the arc incidence matrices for this graph and the arc and ideal matrices.

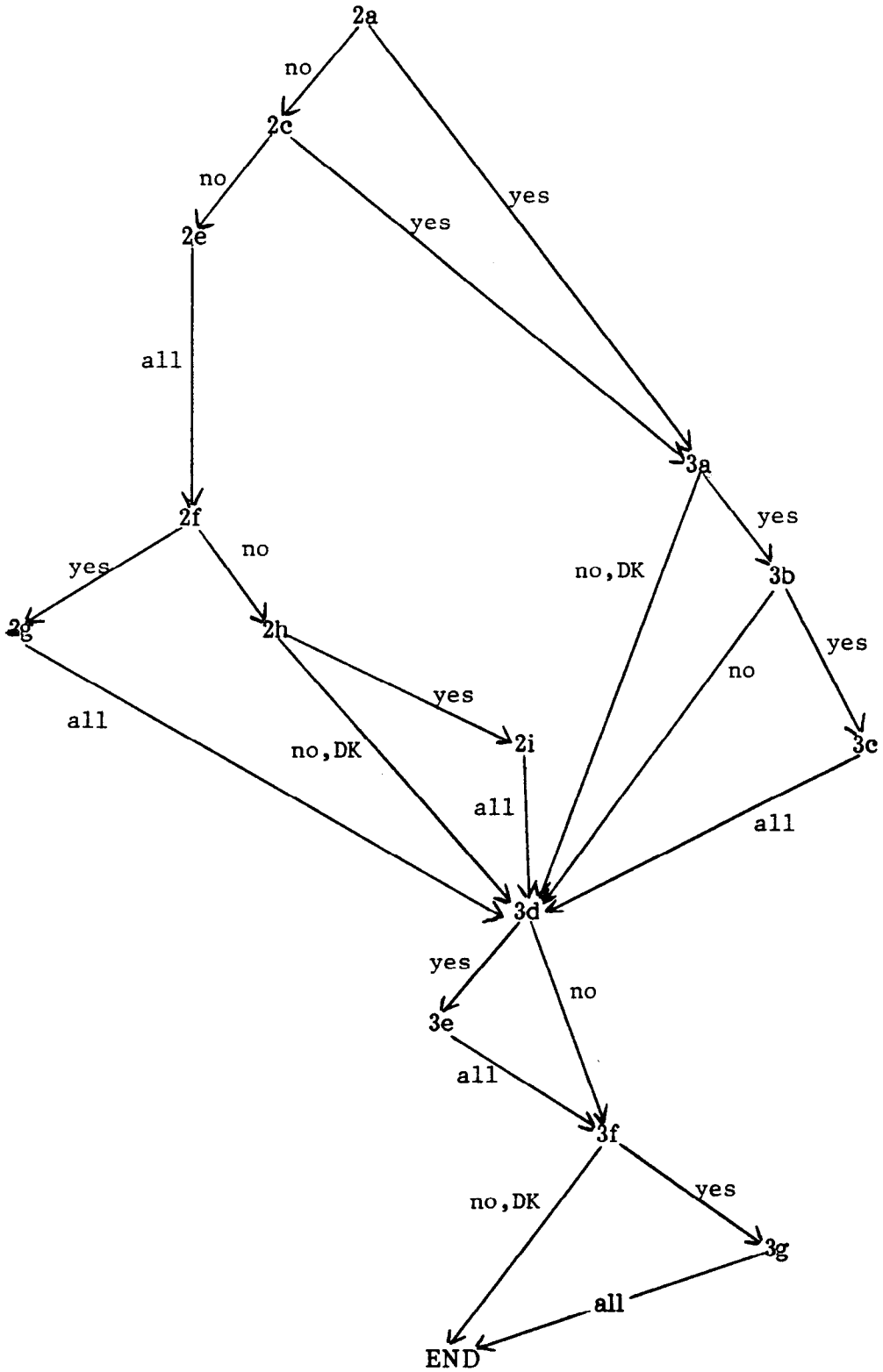

2a
no
2c
no
yes
2e
yes
all
3a
2f
yes
yes
no
no,DK
3b
2g
2h
yes
no
yes
all
2i
no,DK
3c
all
all
3d
yes
no
3e
all
3f
no,DK
yes
3g
all
END

Figure 1

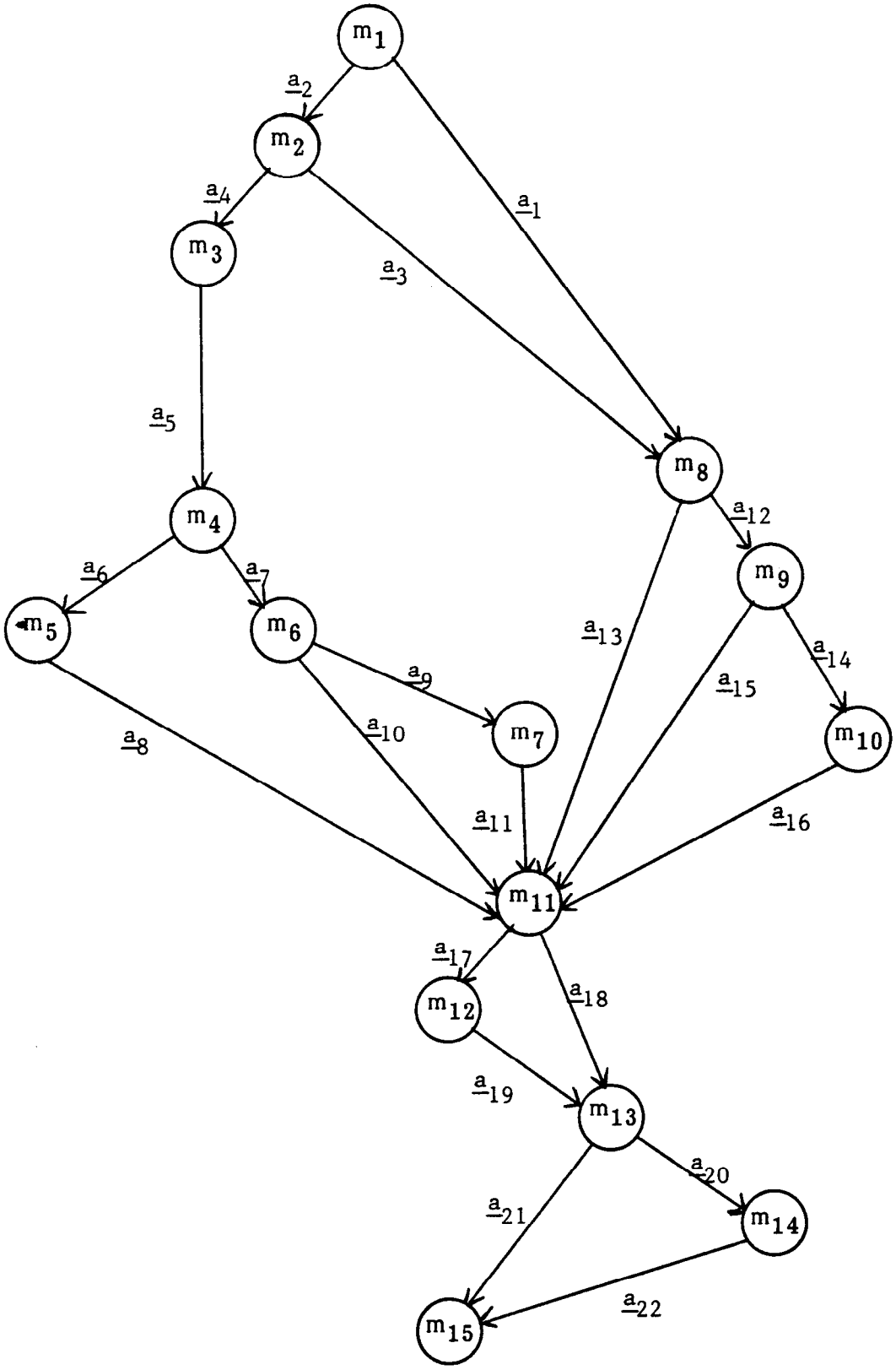

m1
a2
m2
a1
a4
m3
a3
a5
m8
m4
a12
a6
a7
m9
m5
m6
a13
a14
a9
a15
a10
m7
m10
a8
a11
a16
m11
a17
m12
a18
a19
m13
a20
a21
m14
a22
m15

Figure 2

Arc Incidence Matrix For Example 8

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}	a_{20}	a_{21}	a_{22}
a_1	1											1	1									
a_2		1	1	1																		
a_3			1									1	1									
a_4				1	1																	
a_5					1	1	1															
a_6						1		1														
a_7							1		1	1												
a_8								1									1	1				
a_9									1		1											
a_{10}										1							1	1				
a_{11}											1						1	1				
a_{12}												1		1	1							
a_{13}													1				1	1				
a_{14}														1		1						
a_{15}															1		1	1				
a_{16}																1	1	1				
a_{17}																	1		1			
a_{18}																		1		1	1	
a_{19}																			1	1	1	
a_{20}																				1		1
a_{21}																					1	1
a_{22}																						1

Arc Ideal Matrix For Example 8

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}	a_{16}	a_{17}	a_{18}	a_{19}	a_{20}	a_{21}	a_{22}
a_1	1											1	1	1	1	1	1	1	1	1	1	1
a_2		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
a_3			1									1	1	1	1	1	1	1	1	1	1	1
a_4				1	1	1	1	1	1	1	1						1	1	1	1	1	1
a_5					1	1	1	1	1	1	1						1	1	1	1	1	1
a_6						1		1									1	1	1	1	1	1
a_7							1		1	1	1						1	1	1	1	1	1
a_8								1									1	1	1	1	1	1
a_9									1		1						1	1	1	1	1	1
a_{10}										1							1	1	1	1	1	1
a_{11}											1						1	1	1	1	1	1
a_{12}												1		1	1	1	1	1	1	1	1	1
a_{13}													1				1	1	1	1	1	1
a_{14}														1		1	1	1	1	1	1	1
a_{15}															1		1	1	1	1	1	1
a_{16}																1	1	1	1	1	1	1
a_{17}																	1		1	1	1	1
a_{18}																		1		1	1	1
a_{19}																			1	1	1	1
a_{20}																				1		1
a_{21}																					1	
a_{22}																						1

Node Incidence Matrix for Example 8

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}	m_{12}	m_{13}	m_{14}	m_{15}
m_1	1	1						1							
m_2		1	1					1							
m_3			1	1											
m_4				1	1	1									
m_5					1						1				
m_6						1	1				1				
m_7							1				1				
m_8								1	1		1				
m_9									1	1	1				
m_{10}										1	1				
m_{11}											1	1	1		
m_{12}												1	1		
m_{13}													1	1	1
m_{14}														1	1
m_{15}															1

Node Ideal Matrix For Example 8

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}	m_{11}	m_{12}	m_{13}	m_{14}	m_{15}
m_1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
m_2		1	1	1	1	1	1	1	1	1	1	1	1	1	1
m_3			1	1	1	1	1				1	1	1	1	1
m_4				1	1	1	1				1	1	1	1	1
m_5					1						1	1	1	1	1
m_6						1	1				1	1	1	1	1
m_7							1				1	1	1	1	1
m_8								1	1	1	1	1	1	1	1
m_9									1	1	1	1	1	1	1
m_{10}										1	1	1	1	1	1
m_{11}											1	1	1	1	1
m_{12}												1	1	1	1
m_{13}													1	1	1
m_{14}														1	1
m_{15}															1

B. Analysis of Response Forms: Missing Items and Consistent Responses

Given an unanswered data item on a questionnaire, either it is missing and needs to be imputed, or the question is not applicable by virtue of other responses and the underlying skip pattern for the questionnaire.

For example, to use an extreme case, suppose on the ISDP segment in Example 8, a respondent answered question 2a with the response "NO", 3a with the response "YES", and that's all. What could be said about this questionnaire? One notes that questions 2e, 2f, 2g, 2h, and 2i are not applicable. Questions 2c, 3b, 3d, and 3f are missing and must be imputed. Questions 3c, 3e and 3g have an undetermined status which depends on the responses, respectively of 3b, 3d, and 3f which are missing and must be imputed. Furthermore, the response to question 2c must be imputed as a "YES".

As a second example, suppose a questionnaire was filled in as follows: Question 2a with response "YES", Question 3a with response "NO", Question 3d with response "NO", and Question 3f with response "NO". It is then clear that this questionnaire is complete, and all missing responses are not applicable.

Through a analysis of the underlying skip pattern of a questionnaire, one can determine, for many questions, which response variables are missing and must be imputed and which are not applicable. For some of the variables that are missing and must be imputed, a unique valid imputation can be recognized based solely on the structure of the skip pattern. Of course, such an analysis will not resolve all missing values. The deterministic information uncovered by the skip pattern analysis must be followed by survey-specific imputation procedures. That is, given that Question 3b is missing and must be imputed in the first example above, we cannot determine a value to be imputed solely by an analysis of the skip pattern. In some sense, this skip pattern analysis can be thought of as a preprocessor for response forms with blanks, the output of which is a questionnaire having fewer indeterminate blanks and some deterministic imputations.

Example 9:

Let us consider another example where the responses to the ISPD questionnaire extract are:

Question 2a ---- response "NO"
Question 2c ---- response "YES"
Question 2e ---- response "(any)"
Question 2f ---- response "NO"
Question 2h ---- response "YES"
Question 3f ---- response "YES"

Note the response of "YES" to question 2c is inconsistent with the fact that Questions 2e, 2f and 2h were also responded to. There is an inconsistency in the recorded response form and it is easily detected by looking at the questionnaire graph Figure 1. Clearly, if we delete the response of "YES" to Question 2c and impute the response "NO", the revised questionnaire will be consistent.

We could also have observed this by shifting our focus to the graph in Figure 2. We consider the response arcs

$$\{\underline{a}_2, \underline{a}_3, \underline{a}_5, \underline{a}_7, \underline{a}_9, \underline{a}_{20}\}$$

and observe the following inconsistent pairs:

$$(\underline{a}_3, \underline{a}_5)$$
$$(\underline{a}_3, \underline{a}_7)$$
$$(\underline{a}_3, \underline{a}_9).$$

By "deleting" arc $\underline{a}_3$ and "imputing" arc $\underline{a}_4$ we obtain a consistent arc set, namely:

$$\{\underline{a}_2, \underline{a}_4, \underline{a}_5, \underline{a}_7, \underline{a}_9, \underline{a}_{20}\}.$$

Remark: The term "inconsistent" when used to discuss the questionnaire is used in its usual sense, but when applied to the arc sets it is used as in the definition of Section II. The terminology in Section II for graphs was chosen to mirror common usage in discussing questionnaires and response forms.

The essence of the methods in this report revolves about the following observation: given that question K is responded to, in order for question L to also be responded to, L must either follow K or precede K based on a questionnaire skip pattern. Thus, in the graph derived from the questionnaire, node L must either follow node K or precede node K where the words "preceed" and "follow" are applied as graph terminology introduced in Section II. That is, n_K must be in the ideal generated by n_L, (and hence n_L must be in the ideal generated by n_K) in order that question L and question K are simultaneously present on some questionnaire response form. Accordingly, if $\{n_i\}_{i\in I}$ is a set of consistent responses, then a node, n_K, that is not in the intersection of the ideals $B_N(n_i)$ must be non-applicable.

Considering only nodes does not suffice, and we must also consider arcs that preceed and follow a reported arc. For example, in the ISDP questionnaire above (Example 9), if question 2c was responded to with a response of "YES", i.e., arc $\underline{a}_3$ was answered, then arc $\underline{a}_5$ cannot be answered, nor could arc $\underline{a}_4$. Both of these arcs fail to be in the arc ideal of $\underline{a}_3$, $B_A(\underline{a}_3)$. Accordingly, arcs $\underline{a}_4$ and $\underline{a}_5$ do not constitute viable responses when $\underline{a}_3$ is treated as a valid response (and conversely). Thus, we must examine the structure of arc relations as well as node relations, and this will be treated in the following sections.

IV. GRAPH THEORETIC ANALYSIS OF QUESTIONNAIRE STRUCTURE

In this section, we will describe methodologies and programs to implement graph theoretic procedures for analysis of skip patterns in questionnaires. The first program to be described will start with a questionnaire and will form the graph theoretic constructions needed to examine skip patterns. The second program is used to edit response forms using the skip pattern structure of the questionnaire.

A. GRAPH 1 -- Questionnaire Structure

In the first program, call it GRAPH 1, we read in the questionnaire and set up the basic graph theoretic constructs. Before using this program, a user will number each question on the questionnaire, and the only restriction is that if Question K follows Question L, then K is greater than L.

A user first enters into the program the number of questions in the questionnaire, **N**. The program forms the set of nodes, one for each question and the program then prompts the user, one node at a time, to specify the nodes that are the immediate successors. The program then sets up a node incidence matrix c_N and arc incidence matrix c_A based on the information provided.

After creating the ideal matrix, b_N, from the incidence matrix c_N, and the ideal matrix b_A, from the incidence matrix c_A, the program supplies diagnostic information about the questionnaire. Recall that since each question is numbered, when we speak of question 7 and node 7 we mean the same thing. The program provides the following information.

(a) It prints out the input information as a check.
(b) For each question it lists the questions immediately following.
(c) For each arc, it tells which nodes the arc goes between.
(d) For each node, (respectively, arc) it lists all questions (respectively, response) that follow (not necessarily immediately).
(e) For each node, (respectively, arc) it lists all questions (respectively, responses) that preceed it.
(f) It lists all possible response patterns. That is, it lists all possible complete questionnaires if the user requests.

We entered the ISDP questionnaire extract into this program and in Appendix I we have the computer generated diagnostics for this graph.

B. GRAPH 2 -- Analysis of Response Forms

The arc ideal matrix and the node ideal matrix created in GRAPH 1 are passed to a second program, call it GRAPH 2, which will analyze survey response forms. Survey response forms are entered into this program, one response record at a time.

The program will first determine if all responses on a record are consistent (with respect to the skip pattern structure), and if not, the program will select a set of responses to delete so that the remaining responses are mutually consistent. The criterion built into the system is to delete as few responses as possible so that the remaining are mutually consistent. In fact, one can assign preferance factors (weights) to each field so that the system will locate a weighted minimal set of fields to delete. These weights are to be provided by the user before the system is executed. After responses causing inconsistencies (if any) have been deleted they are treated as not reported. The remaining responses on the record are mutually consistent.

After a record has been processed through the program GRAPH 2, each field will be assigned one of four status flags:

(a) valid response,
(b) not applicable,
(c) missing and must be imputed,
(d) status cannot be determined.

Below we present the procedures embedded in GRAPH 2. In this discussion we will use the following notation:

N = set of all nodes,
R^* = set of reported nodes,
R = set of mutually consistent reported nodes,
M = set of nodes which must be imputed,

U = set of nodes whose status cannot be determined,

A = set of all arcs,

Q^* = set of reported arcs (i.e., arcs corresponding to a reported response),

Q = set of valid reported arcs (i.e., arcs corresponding to a set of mutually consistent reported responses).

When we read in a questionnaire, the program records the questions responded to and the responses provided. That is, we have a list of reported nodes, R^*, and a set of reported arcs, Q^*. If $Q^* \subset B(Q^*)$ the questionnaire is consistent (Proposition 2) so we set $R=R^*$ and $Q=Q^*$ and proceed to examine the non-reported nodes.

If $Q^* \not\subset B(Q^*)$, we compile a listing of all pairs of mutually inconsistent responses. As noted in Section II, two responses, $\underline{a}$ and $\underline{b}$, are inconsistent if $\underline{a} \notin B_A(\underline{b})$. We delete a (weighted) minimal set of response arcs so that the remaining will be mutually consistent. The consistent arcs form the set Q, and R consists of initial points of arcs in Q. The remaining nodes, N-R, are considered not reported.

At this stage, we have a consistent set of reported arcs, Q, and a set of consistent nodes, R. The set of arcs containing all possible valid responses is B(Q) (Proposition 2), and the set consisting of all nodes that can possibly be applicable is P(B(Q)) (Proposition 7). Hence, the nodes which are <u>not applicable</u> are N-P(B(Q)), and we denote this set by L.

To determine the nodes that are missing and <u>must</u> be imputed, we find the set of waist nodes of Q (Theorem 3). Note that each maximal chain corresponds to a completed response form. The set of waist nodes are those nodes that must line on every maximal chain containing Q and, hence, correspond to questions that <u>must</u> be answered given that questions corresponding to Q were answered. Thus the set of nodes that are <u>missing and must be imputed</u>, M, are those nodes in

$$\bigcap_{\underline{q} \in B(Q)} P(B_A(\underline{q}))$$

other than the valid reported nodes. Hence $M = \bigcap_{g \in B(Q)} P(B_A(g)) - R$.

The status of the remaining nodes cannot be determined. Whether they are missing and must be imputed or are not applicable depends on responses to the nodes currently missing and yet to be imputed. Denoting these undeterminable nodes by U, we have

$$U = N - R - M - L.$$

Example 10: Returning to Example 9,

$$Q^* = \{\underline{a}_2, \underline{a}_3, \underline{a}_5, \underline{a}_7, \underline{a}_9, \underline{a}_{20}\}, \quad \text{so}$$

$$R^* = \{m_1, m_2, m_3, m_4, m_6, m_{13}\}.$$

The arc set Q^* is inconsistent with inconsistent pairs

$$(\underline{a}_3, \underline{a}_5)$$
$$(\underline{a}_3, \underline{a}_7)$$
$$(\underline{a}_3, \underline{a}_9).$$

By deleting response a_3, all inconsistencies are removed, and we have:

$$Q = \{\underline{a}_2, \underline{a}_5, \underline{a}_7, \underline{a}_9, \underline{a}_{20}\}$$

$$R = \{m_1, m_3, m_4, m_6, m_{13}\}$$

$$B(Q) = \{\underline{a}_2, \underline{a}_4, \underline{a}_5, \underline{a}_7, \underline{a}_9, \underline{a}_{11}, \underline{a}_{17}, \underline{a}_{18}, \underline{a}_{19}, \underline{a}_{20}, \underline{a}_{22}\}$$

$$P(B(Q)) = \{m_1, m_2, m_3, m_4, m_6, m_7, m_{11}, m_{12}, m_{13}, m_{14}, m_{15}\}$$

$$N\text{-}P(B(Q)) = \{m_5, m_8, m_9, m_{10}\}.$$

Thus the non-applicable questions are:

$$L = \{m_5, m_8, m_9, m_{10}\}.$$

By directly computing

$$\bigcap_{g \in B(Q)} P(B_A(g))-R$$

one finds that the missing nodes are:

$$M = \{m_2, m_7, m_{11}, m_{14}, m_{15}\} .$$

The set of nodes with status undetermined is

$$U = N-R-M-L$$

which, in this example is:

$$U = \{m_{12}\} .$$

One observes further that the responses to the following questions can be inferred:

Question	Response	
2c (node m_2)	NO	(arc a_4)
2i (node m_7)	(ANY)	(arc a_{11})
3g (node m_{14})	(ANY)	(arc a_{21}).

Since we cannot determine the response to question 3d (node m_{11}) the response status of question 3e (node m_{12}) is undetermined.

Remark: Examples 9 and 10 as well as the two examples on page 27 were run through the GRAPH 2 program, and the output is in Appendix II.

Remark: It is interesting to observe how one can obtain some of these results by drawing upon the localization graph G_Q and the information in Theorem 2 and the Remarks following. For $G_Q = (P(B(Q)), B(Q))$, we have the following graph in Figure 3. It is clear that all nodes in G_Q are waist nodes except m_{12}.

Thus, the waist nodes are; $\{m_2, m_7, m_{11}, m_{14}, m_{15}\}$, so each of these nodes (with the exception of node m_{15}) corresponds to a question that is missing and must be imputed. The node m_{12} is not a waist node and thus has the status undetermined, and it cannot be determined without knowing the response to node m_{11}, which is missing and must be imputed. Note that there is only a single arc leading from node m_2, and that corresponds to a "NO" response to question 2c, hence we have an implied imputation.

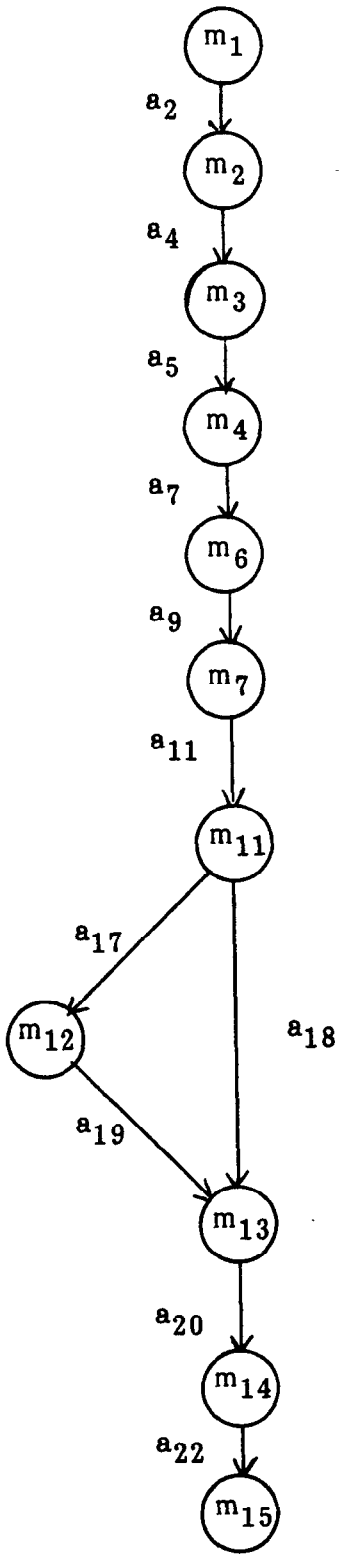

m1
a2
m2
a4
m3
a5
m4
a7
m6
a9
m7
a11
m11
a17
a18
m12
a19
m13
a20
m14
a22
m15

FIGURE 3

C. Method for Determining Responses to Delete on an Inconsistent Questionnaire

Given an inconsistent response form, our objective is to identify a subset of responses and change only them so that all remaining responses are mutually consistent. We denote those responses targeted for change as a deletion set. As a rule, the goal is to change as few responses as possible (and hence retain as many reported values as possible). By assigning weights (preference factors) one can delete a weighted minimal set of responses.

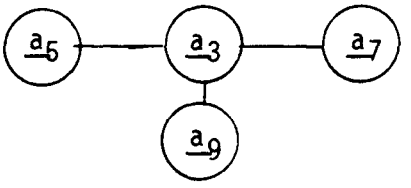
In the setting of skip pattern analysis one determines a deletion set, D, in the following manner. Consider all inconsistent pairs of responses and let D be a subset of responses such that at least one response from each failed pair is in D. For Example 9, the pairs of inconsistent responses are:

$$(\underline{a}_3, \underline{a}_5)$$
$$(\underline{a}_3, \underline{a}_7)$$
$$(\underline{a}_3, \underline{a}_9).$$

The singleton set $\{\underline{a}_3\}$ = D has the required property (i.e., $\underline{a}_3$ is an element in each failing pair) so that $\{\underline{a}_3\}$ = D is a deletion set. Thus, by "deleting" the response $\underline{a}_3$ (i.e., the response to question 2c) all remaining responses are consistent with respect to the underlying skip pattern. Of course, for more complex patterns of inconsistencies, the choice of deletion set is not quite so obvious.

In the program GRAPH 2, in order to determine the responses to delete on an inconsistent questionnaire, we construct the failed edit graph. The failed edit graph is an undirected graph used to represent inconsistencies on a questionnaire. The nodes on the failed edit graph correspond to responses on the questionnaire involved in an inconsistency, (i.e., to inconsistent response arcs on the questionnaire graph). Arcs on the failed edit graph correspond to inconsistent pairs of arcs on the original questionnaire. That is, if $\underline{a}$ and $\underline{b}$ are reported arcs on the questionnaire, and $\underline{a} \notin B_A(\underline{b})$, then there are nodes on the failed edit graph for $\underline{a}$ and for $\underline{b}$, and there is an arc between the nodes $\underline{a}$ and $\underline{b}$.

Using the Example 9, the failed edit graph is:



By removing node a_3 in the failed edit graph above (corresponding to response a_3 on the original graph) all other responses are mutually consistent.

The failed edit graph is disconnected by removing a set of nodes and by removing all arcs incident with a deleted node. When no arcs are left, the graph is said to be totally disconnected, and the nodes removed correspond to responses to be deleted. All other remaining responses will be mutually consistent.

There are a variety of ways to determine a minimal disconnecting set for a graph; one can employ a set covering procedure, devise reasonable heuristic algorithms, or rely on strictly graph-theoretic methods. For the program GRAPH 2 , we have taken the last option, and the procedure embedded into this program can be thought of as a clique generating approach. We briefly describe below the approach taken in GRAPH 2.

Let G = (V,H) be an arbitrary graph (not necessarily connected) and let $\bar{G} = (V,\bar{H})$ be the complementary graph. That is, the nodes of $\bar{G}$ are the same as those of G, but (v_1,v_2) is an arc in $\bar{H}$ if and only if (v_1,v_2) is not an arc in H. If G = (V,H) is an arbitrary graph, a clique is a set of nodes, W, of V such that if $w_1,w_2, \in W$ then (w_1,w_2) is an arc in H. A maximal clique is a clique properly contained in no other clique. If we attach weights to each vertex of a graph and define the weight of a clique to be the sum of the weights of the elements in the clique, we can define a maximal weighted clique. Clearly all maximal weighted cliques are maximal cliques (assuming all weights are positive).

A minimal (weighted) disconnecting set of a graph G corresponds to the complement of a maximal (weighted) clique in the complement graph, $\bar{G}$. Thus, given an edit failing record, r, to find the minimal deletion set, form the failed edit graph, G, and find a maximal (weighted) clique, C, in $\bar{G}$. The minimal (weighted) deletion set then corresponds to V-C.

Cliques have been extensively studied, and exact algorithms exist to find all maximal cliques for an arbitrary graph. We have programmed into GRAPH 2 an algorithm to find all maximal cliques of a graph as described in [3]. Having all maximal cliques of a graph, one can easily determine all maximal weighted cliques and hence all weighted minimal deletion sets. By having all minimal weighted deletion sets for an edit failing record, a user of the edit methodology has the option of selecting from among all alternative sets of fields to delete on edit failing records.

Remark: It is likely that a "reasonable heuristic" may be preferable in this program to the clique generation procedure referred to above. The process of disconnecting the failed edit graph is carried out in an external subroutine of GRAPH 2. It would be quite easy to swap the currently residing graph disconnecting routine in favor of any other. Such a replacement would not alter the flow or the underlying logic of GRAPH 2 nor alter the program performance. An example of a heuristic technique to disconnect the failed edit graph is given in [1]. This procedure was developed to disconnect the failed edit graph when editing economic data under ratio edits.

V. SUMMARY

The objective of this report has been to describe a methodology and programs to implement it for the analysis of skip patterns in questionnaires. The methods and programs can be employed to analyze the underlying skip pattern structure for questionnaires during the design process and to analyze the skip patterns on individual questionnaire response forms.

The underlying skip pattern structure of questionnaires is often very complex, and current techniques to deal with them are often quite complicated, ad hoc, and accordingly, error prone. Typically, for each survey instrument having a complex skip pattern, special purpose specifications for the analysis of response forms are written and computer code is developed from these specifications. We have described in this report two structured, parameter driven programs that can be employed to (1) provide users with a better understanding of the basic questionnaire and (2) to use in editing questionnaire forms.

In the early sections of this report we present the appropriate mathematical model, namely, the directed graph, and establish relationships within a graph that will be needed in later sections. We next show how these graph theoretic constructs have a simple representation in terms of zero-one matrices. In the next two sections we showed how a questionnaire and its skip pattern structure can be modeled as a directed graph and we showed how the graph theoretic relationships developed earlier apply. Finally, two computer programs that implement the procedures discussed earlier are described, and samples of computer generated analysis from these programs are included in Appendixes I and II.

The program, GRAPH II, described above for editing skip patterns in questionnaire forms is not meant to be a comprehensive edit and imputation package. Its primary goal is to recognize non-applicable questions on a response form and differentiate them from responses that are missing and must be imputed. In the process of doing this, one frequently can discover deterministic imputations for selected missing items. In addition, if any inconsistencies (with respect to the skip pattern) appear on a response form they will be detected by this program and a minimal set of responses will be deleted so that the remaining are consistent. After all inconsistencies on a record have been resolved and missing fields have been flagged as non-applicable or needing imputation, survey specific imputation routines can be brought to bear in the creation of a complete and consistent record.

Using the program GRAPH 1 by itself to analyze skip pattern structure, or using it in conjunction with GRAPH 2 to analyze response forms can enhance operations for processing surveys whose questionnaires have a complex underlying skip pattern.

REFERENCES

[1] Greenberg, B. (1981). Developing an Edit System for Industry Statistics. Computer Science and Statistics: Proceedings of the 13th Symposium of the Interface, 11-16. Springer-Verlag, New York.

[2] Harary, F. (1969). Graph Theory. Addison-Wesley, New York.

[3] Mulligan, G.D. and Cornell D.G. (1972). Corrections to Bierstone's Algorithm for Generating Cliques. Journal of the Association for Computing Machinery. 19, 2. 244-7.

APPENDIX I

On the following pages, we show the output generated by GRAPH 1 when the questionnaire in Example 8 was entered and run. The diagnostic information is potentially valuable for examining the questionnaire structure for design and subsequent analysis.

the number of nodes is 15

question	node number:	number of different responses:	questions immediately following:
2a. did you work	1	2	2 8
2c. part-time work	2	2	3 8
2e. why no work	3	1	4
2f. look for work	4	2	5 6
2g. weeks looking	5	1	11
2h. want a job	6	2	7 11
2i. why no look	7	1	11
3a. weeks no job	8	2	9 11
3b. weeks looking	9	2	10 11
3c. how many weeks	10	1	11
3d. unemployment	11	2	12 13
3e. source	12	1	13
3f. income for injur	13	2	14 15
3g. source	14	1	15
(dummy question)	15	none	none

the number of arcs is 22

```
arc   1 goes from node  1 to node  8
arc   2 goes from node  1 to node  2
arc   3 goes from node  2 to node  8
arc   4 goes from node  2 to node  3
arc   5 goes from node  3 to node  4
arc   6 goes from node  4 to node  5
arc   7 goes from node  4 to node  6
arc   8 goes from node  5 to node 11
arc   9 goes from node  6 to node  7
arc  10 goes from node  6 to node 11
arc  11 goes from node  7 to node 11
arc  12 goes from node  8 to node  9
arc  13 goes from node  8 to node 11
arc  14 goes from node  9 to node 10
arc  15 goes from node  9 to node 11
arc  16 goes from node 10 to node 11
arc  17 goes from node 11 to node 12
arc  18 goes from node 11 to node 13
arc  19 goes from node 12 to node 13
arc  20 goes from node 13 to node 14
arc  21 goes from node 13 to node 15
arc  22 goes from node 14 to node 15
```

question number	question name	response	arc number	skip path
1	2a. did you work			
		res 1: yes	1	skip to 8
		res 2: no	2	skip to 2
2	2c. part-time work			
		res 1: yes	3	skip to 8
		res 2: no	4	skip to 3
3	2e. why no work			
		res 1: any	5	skip to 4
4	2f. look for work			
		res 1: yes	6	skip to 5
		res 2: no	7	skip to 6
5	2g. weeks looking			
		res 1: any	8	skip to 11
6	2h. want a job			
		res 1: yes	9	skip to 7
		res 2: no-dk	10	skip to 11
7	2i. why no look			
		res 1: any	11	skip to 11
8	3a. weeks no job			
		res 1: yes	12	skip to 9
		res 2: no-dk	13	skip to 11
9	3b. weeks looking			
		res 1: yes	14	skip to 10
		res 2: no	15	skip to 11
10	3c. how many weeks			
		res 1: any	16	skip to 11
11	3d. unemployment			
		res 1: yes	17	skip to 12
		res 2: no	18	skip to 13
12	3e. source			
		res 1: any	19	skip to 13
13	3f. income for injur			
		res 1: yes	20	skip to 14
		res 2: no-dk	21	skip to 15
14	3g. source			
		res 1: any	22	skip to 15
15	(dummy question)	none	none	none

question	potential subsequent questions
1:	1 2 3 4 5 6 7 8 9 10 11 12 13 14
2:	2 3 4 5 6 7 8 9 10 11 12 13 14
3:	3 4 5 6 7 11 12 13 14
4:	4 5 6 7 11 12 13 14
5:	5 11 12 13 14
6:	6 7 11 12 13 14
7:	7 11 12 13 14
8:	8 9 10 11 12 13 14
9:	9 10 11 12 13 14
10:	10 11 12 13 14
11:	11 12 13 14
12:	12 13 14
13:	13 14
14:	14

question	potential preceeding questions
1:	1
2:	1 2
3:	1 2 3
4:	1 2 3 4
5:	1 2 3 4 5
6:	1 2 3 4 6
7:	1 2 3 4 6 7
8:	1 2 8
9:	1 2 8 9
10:	1 2 8 9 10
11:	1 2 3 4 5 6 7 8 9 10 11
12:	1 2 3 4 5 6 7 8 9 10 11 12
13:	1 2 3 4 5 6 7 8 9 10 11 12 13
14:	1 2 3 4 5 6 7 8 9 10 11 12 13 14

arc	potential subsequent arcs
1:	1 12 13 14 15 16 17 18 19 20 21 22
2:	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
3:	3 12 13 14 15 16 17 18 19 20 21 22
4:	4 5 6 7 8 9 10 11 17 18 19 20 21 22
5:	5 6 7 8 9 10 11 17 18 19 20 21 22
6:	6 8 17 18 19 20 21 22
7:	7 9 10 11 17 18 19 20 21 22
8:	8 17 18 19 20 21 22
9:	9 11 17 18 19 20 21 22
10:	10 17 18 19 20 21 22
11:	11 17 18 19 20 21 22
12:	12 14 15 16 17 18 19 20 21 22
13:	13 17 18 19 20 21 22
14:	14 16 17 18 19 20 21 22
15:	15 17 18 19 20 21 22
16:	16 17 18 19 20 21 22
17:	17 19 20 21 22
18:	18 20 21 22
19:	19 20 21 22
20:	20 22
21:	21
22:	22

arc	potential preceeding arcs
1:	1
2:	2
3:	2 3
4:	2 4
5:	2 4 5
6:	2 4 5 6
7:	2 4 5 7
8:	2 4 5 6 8
9:	2 4 5 7 9
10:	2 4 5 7 10
11:	2 4 5 7 9 11
12:	1 2 3 12
13:	1 2 3 13
14:	1 2 3 12 14
15:	1 2 3 12 15
16:	1 2 3 12 14 16
17:	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
18:	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18
19:	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 19
20:	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21:	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 21
22:	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 22

APPENDIX II

On the following three pages, we display three possible response records run through GRAPH 2. The first two response forms corresponds to the examples on page 27, and the last response form corresponds to Example 9 on page 28 and continued as Example 10 on pages 33 through 34.

respondent # 1

--responses on form--

question name	node number	response arc	response provided	# response code
2a. did you work	1	2	no	2
2c. part-time work	2	3	yes	1
2e. why no work	3	5	any	1
2f. look for work	4	7	no	2
2h. want a job	6	9	yes	1
3f. income for injur	13	20	yes	1

the following pairs of response arcs are inconsistent:

3	5
3	7
3	9

the following pairs of points have inconsistent arcs eminating from them:

2	3
2	4
2	6

the deleted response arcs are: 3

responses have been deleted for questions: 2

the edited questionaire is below. a * indicates possible imputation values

node number	response status		valid response codes
1	valid response		2
2	missing, to be imputed	*	2
3	valid response		1
4	valid response		2
6	valid response		1
7	missing, to be imputed	*	1
11	missing, to be imputed	*	1 2
12	status undeterminable		
13	valid response		1
14	missing, to be imputed	*	1

the inapplicable questions are: 5 8 9 10

respondent # 2

--responses on form--

question name	node number	response arc	response provided	response code
2a. did you work	1	2	no	2
3a. weeks no job	8	12	yes	1

the edited questionaire is below. a * indicates possible imputation values

node number	response status		valid response codes
1	valid response		2
2	missing, to be imputed	*	1
8	valid response		1
9	missing, to be imputed	*	1 2
10	status undeterminable		
11	missing, to be imputed	*	1 2
12	status undeterminable		
13	missing, to be imputed	*	1 2
14	status undeterminable		

the inapplicable questions are: 3 4 5 6 7

respondent # 3

--responses on form--

question name	node number	response arc	response provided	response code
2a. did you work	1	1	yes	1
3a. weeks no job	8	13	no-dk	2
3d. unemployment	11	18	no	2
3f. income for injur	13	21	no-dk	2

the edited questionaire is below. a * indicates possible imputation values

node number	response status	valid response codes
1	valid response	1
8	valid response	2
11	valid response	2
13	valid response	2

the inapplicable questions are: 2 3 4 5 6 7 9 10 12 14