1970

Theoretical and computational aspects of clumps

Robert C. Heiser
Lehigh University

THEORETICAL AND COMPUTATIONAL ASPECTS OF CLUMPS

by

Robert C. Heiser

A THESIS

Presented to the Graduate Committee

of Lehigh University

in Candidacy for the Degree of

Master of Science

in

Information Sciences

Lehigh University

1970

This thesis is accepted and approved in partial fulfillment of the requirements for the degree of Master of Science.

20 May 1970
(date)

Robert F. Baum
Professor in charge

Daniel J. Hilman
Chairman of the Department

ACKNOWLEDGEMENTS

The author wishes to express his appreciation and gratitude to Professor Robert F. Barnes for his invaluable help and amazing patience throughout the preparation and revision of this paper; to Thomas Morrisette for his interesting discussions and suggestions; and to Ann and Debby for their thoughtfulness, kindness, and general moral support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Chapter 1 of this paper contains the general theory of clumps, Chapter 2 deals with three applications of this theory in associative information retrieval, and Chapter 3 presents a procedure for computing clumps.

Chapter 1 deals with the abstract structure (T,M), where T is a non-empty set and M is a binary, symmetric, non-negative extended real-valued function which is positive on the diagonal of T. Section 1.1 is devoted to general background. Section 1.2 deals with alternative approaches in defining (T,M). In Section 1.3, $M_t(A)$ is defined to be $\sum\{M(t,t'): t' \in A\}$ when A is a countable subset of T and $t \in T$. Observing that the countable subsets of T form a σ-ring of measurable sets, M_t is a positive measure, called the t-induced M-measure, for each $t \in T$. In Section 1.4, two subsets A and B of T are defined to touch each other when $M(a,b) > 0$ for some $a \in A$ and some $b \in B$. A class of subsets of T is separated if no two distinct sets in the class touch each other. A subset of T is connected whenever no decomposition containing two or more sets is separated. For each subset A of T, an A-component is a maximal connected subset of A, and the class of all A-components is shown to be separated. Fundamental properties of these and other concepts are stated, interrelated, and used in Section 1.5, which introduces the notion of an R-clump. A non-empty

subset A of T is an R-clump whenever for each $t \epsilon A$ and each countable $B \subseteq T$, $M_t(B - A) \leq M_t(A)$. There is always an R-clump cover of T. Measurable unions of R-clumps are R-clumps. If the union over a separated class is an R-clump, each set in that class is an R-clump. An atomic R-clump is one which is not expressible as the union of two properly included R-clumps. Each atomic R-clump is connected. If M_t is bounded for each $t \epsilon T$, using the fact that each T-component is then countable, it is shown that R-clumps are generable from atomic R-clumps by taking unions. Strong clumps are connected R-clumps.

Chapter 2 presents an interpretation of T as a set of vocabulary terms used to characterize a document collection and of M as an associative term-term matrix. Strong clumps are term sets which hopefully have intuitive as well as formal coherence. A concrete example is given as an illustration. Strong clumps may be used in man-machine negotiation, in system-managed informal vocabulary control, and in finding articulation points of T-components.

Chapter 3 presents a procedure for computing strong clumps. The analytic component of this procedure uses facts about non-atomic R-clumps of a given cardinality to isolate atomic R-clumps of that cardinality. The generative component uses atomic R-clumps to compute non-atomic R-clumps of higher cardinalities by taking unions. Finally, equivalent procedures, estimating procedures, and short-cut procedures for computing clumps are briefly discussed.

CHAPTER ONE: THE THEORY OF CLUMPS

Section 1.1: General Preliminaries and Notation.

The following notational conventions are observed. The phrase "if and only if" is abbreviated as "iff". Let A and B be sets. $\|A\|$, the cardinality of A, is the number of elements of A. The set A is either finite, countably infinite, or uncountable depending upon which of these mutually exclusive and exhaustive properties is ascribed to $\|A\|$. A is countable iff A is either finite or countably infinite. $\mathcal{P}(A)$, the power set of A, is the class of all subsets of A. B—A, the relative complement of A with respect to B, is $\{x: x\in B \text{ and } x\notin A\}$. $A\times B$, the Cartesian product of A with B, is $\{(a,b): a\in A \text{ and } b\in B\}$ (where (a,b) is the ordered pair whose first element is a and whose second element is b). Also, $\emptyset$ is the empty set.

Let $\mathcal{B}\subseteq\mathcal{P}(T)$. $\mathcal{B}$ is <u>disjointed</u> iff for each $B_1, B_2\in\mathcal{B}$, if $B_1\neq B_2$ then $B_1\cap B_2=\emptyset$. $\mathcal{B}$ is a <u>decomposition</u> of set A iff $\mathcal{B}$ is disjointed, $\emptyset\notin\mathcal{B}$, and $A=\bigcup\mathcal{B}$.

$\mathbb{N}$ is the set of all positive integers; and ω is the first infinite ordinal number. If $\{a_n: n\in\mathbb{N}\}$ is a set, then $\langle a_n\rangle$ is the sequence with terms $a_1, a_2, a_3, \ldots$.

$\mathbb{R}$ is the set of all non-negative real numbers; and $\mathbb{R}^*$ is the set of all non-negative extended real numbers. Thus, $\mathbb{R}^*=\mathbb{R}\cup\{\infty\}$, where ∞, whose name is "infinity", is

the only non-real which is a non-negative extended real. Now $0 \cdot \infty = 0$; and for each $r \in \mathbb{R}^*$, $r \leqslant \infty$, $r + \infty = \infty$, $r \cdot \infty = \infty$ if $r > 0$, and $\infty - r = \infty$ if $r \neq \infty$. The value of $\infty - \infty$ is not defined. Topologically, $\mathbb{R}^*$ is the obvious one-point compactification of $\mathbb{R}$.

Let A be countable, and let f be a function which maps A into $\mathbb{R}^*$. Then $\sum\{f(a): a \in A\}$ is the summation over all values which f attains as a function of each element of A. It follows by an elementary fact about series of non-negative reals that $\sum\{f(a): a \in A\}$ is a unique element in $\mathbb{R}^*$ if A is countable (i.e., an absolutely convergent series has a unique limit regardless of the arrangement of the terms of the series). $\sum\{f(a): a \in \emptyset\}$ is defined to be zero. Finally, f is said to be <u>bounded</u> iff for some $r \in \mathbb{R}$, $f(a) < r$ for each $a \in A$.

<u>Section 1.2</u>: <u>General Framework of the Theory</u>.

In its most general setting, the theoretical portion of this thesis is concerned with the properties of a non-empty set T associated with a reflexive and symmetric binary relation Θ on T which is labeled in a symmetric way by the function m which maps Θ into $\mathbb{R}^* - \{0\}$. In other words, one may consider a structure (T, Θ, m) with the following properties:

(m-i) $T \neq \emptyset$.

(m-ii) $\Theta \subseteq T \times T$.

(m-iii) m is a function which maps Θ into $\mathbb{R}^* - \{0\}$.

(m-iv) Θ is reflexive on T (i.e., for each $t \in T$, $t \Theta t$).

(m-v) Θ is symmetric on T (i.e., for each $t_1, t_2 \in T$, if $t_1 \Theta t_2$ then $t_2 \Theta t_1$).

(m-vi) m labels Θ symmetrically (i.e., for each $t_1, t_2 \in T$, if $t_1 \Theta t_2$ then $m(t_1, t_2) = m(t_2, t_1)$).

The function m from Θ into $\mathbb{R}^* - \{0\}$ may be extended to a function M from $T \times T$ into $\mathbb{R}^*$ so that $M(t_1, t_2) = m(t_1, t_2)$ iff $t_1 \Theta t_2$ and so that $M(t_1, t_2) = 0$ iff $t_1 \not\Theta t_2$.

Alternatively, one may consider this general setting as a non-empty set T associated with a function M which maps $T \times T$ into $\mathbb{R}^*$, which labels the diagonal of T positively, and which labels $T \times T$ symmetrically. In other words, one may consider the structure (T,M) with the following properties:

(M-i) $T \neq \emptyset$

(M-ii) M is a function which maps $T \times T$ into $\mathbb{R}^*$.

(M-iii) M labels the diagonal of T positively (i.e., for each $t \in T$, $M(t,t) > 0$).

(M-iv) M labels $T \times T$ symmetrically (i.e., for each $t_1, t_2 \in T$, $M(t_1, t_2) = M(t_2, t_1)$).

One may define a binary relation Θ on T so that for each $t_1, t_2 \in T$, $t_1 \Theta t_2$ iff $M(t_1, t_2) > 0$. It is clear from (M-iii) and (M-iv) that Θ is reflexive and symmetric on T.

It is this second approach which is actually used in this paper's development of the theory of clumps. A reflexive and symmetric relation Θ is required so that one may interpret "$t_1 \Theta t_2$" as "t_1 touches t_2", "t_1 is next

to t_2", or "t_1 is associated with t_2". Each element of T touches itself (Θ is reflexive on T); and for any two elements of T, if the first touches the second then the second touches the first (Θ is symmetric on T). Now one may also say that two subsets of T "touch each other" if some element of one touches some element of the other. Hence, it is more general to define Θ as a binary relation on $\mathcal{P}(T)$ so that for each $A,B\in\mathcal{P}(T)$, $A\Theta B$ iff $M(a,b)>0$ for some $a\in A$ and some $b\in B$. For this reason, the second of the above two approaches is technically preferable.

Section 1.3: Measures Induced by Elements of T.

With respect to the structure (T,M) satisfying properties (M-i) through (M-iv), the function M is a "weighting" function which evaluates the "strength" of the relation Θ between pairs of (singletons of) elements of T. Every element of T touches itself with some positive M-strength; and for any two elements of T, the first touches the second with the same M-strength with which the second touches the first. Since M is a weighting function, one can then develop a natural measure of a set relative to any fixed element t of T. Such a measure is called the t-induced M-measure; and for each $t\in T$, there is such a measure. The concepts of measurability and measure must be introduced and related to the "weighting" function M.

DEF'N 1: Let $\mathcal{M}\subseteq\mathcal{P}(T)$. Then $\mathcal{M}$ is a σ-ring in T iff:

(i) $\emptyset\in\mathcal{M}$;

(ii) If $\langle A_n \rangle$ is a sequence of elements of $\mathcal{M}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{M}$; and

(iii) If $A, B \in \mathcal{M}$, then $A - B \in \mathcal{M}$.

$\mathcal{M}$ is a <u>σ-algebra</u> in T iff $\mathcal{M}$ is a σ-ring in T for which $T \in \mathcal{M}$.

$\mathcal{M}_c$ is defined as $\{A \in \mathcal{P}(T)$: A is countable$\}$.

<u>REMARK</u>: $\mathcal{M}_c$ is clearly a σ-ring in T; and $\mathcal{M}_c$ is called the σ-ring of countable subsets of T. Clearly, $\mathcal{M}_c$ is a σ-algebra in T iff T is countable.

<u>DEF'N 2</u>: Let $\mathcal{M}$ be a σ-ring in T, and let μ be a function which maps $\mathcal{M}$ into $\mathbb{R}^*$. Then μ is a <u>positive measure</u> on $\mathcal{M}$ iff:

(i) μ is not identically ∞; i.e., for some $A \in \mathcal{M}$, $\mu(A) < \infty$; and

(ii) μ is countably additive; i.e., for each disjointed sequence $\langle A_n \rangle$ of elements of $\mathcal{M}$, $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$.

In addition, the elements of $\mathcal{M}$ are called <u>measurable sets</u>.

<u>LEMMA 1</u>: Let $\mathcal{M}$ be a σ-ring in T, and let μ be a positive measure on $\mathcal{M}$. Then:

(i) μ preserves zero; i.e., $\mu(\emptyset) = 0$.

(ii) μ is finitely additive; i.e., for each $A_1, A_2, \ldots, A_m \in \mathcal{M}$, if $\{A_1, A_2, \ldots, A_m\}$ is disjointed, then $\mu(\bigcup_{n=1}^{m} A_n) = \sum_{n=1}^{m} \mu(A_n)$.

(iii) μ is countably subadditive; i.e., for each sequence $\langle A_n \rangle$ of elements of $\mathcal{M}$, $\mu(\bigcup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \mu(A_n)$.

(iv) μ is monotone; i.e., for each $A, B \in \mathcal{M}$, if $A \subseteq B$,

then $\mu(A) \leq \mu(B)$.

(v) μ is subtractive; i.e., for each $A,B \in \mathcal{M}$, if $A \subseteq B$ and $\mu(A) < \infty$, then $\mu(B-A) = \mu(B) - \mu(A)$.

(vi) μ is continuous from below; i.e., if $\langle A_n \rangle$ is an increasing sequence of elements of $\mathcal{M}$, then $\langle \mu(A_n) \rangle$ converges monotonely to $\mu(\bigcup_{n=1}^{\infty} A_n)$.

(vii) μ is continuous from above; i.e., if $\langle A_n \rangle$ is a decreasing sequence of elements of $\mathcal{M}$ for which $\mu(A_k) < \infty$ for some $k \in \mathbb{N}$, then $\langle \mu(A_n) \rangle$ converges monotonely to $\mu(\bigcap_{n=1}^{\infty} A_n)$.

Proof: This standard measure theoretic result summarizes material presented on pages 30-32, 37-40 of reference [2].//

LEMMA 2: Let $\mathcal{M}$ be a σ-ring in T, let $c_1, c_2, \ldots, c_m \in \mathbb{R}^*$, and let $\mu_1, \mu_2, \ldots, \mu_m$ be positive measures on $\mathcal{M}$. Then $\sum_{k=1}^{m} c_k \mu_k$ is a positive measure on $\mathcal{M}$.

Proof: This standard result is found on page 117 of reference [2].//

DEF'N. 3: For each $t \in T$ and for each $A \in \mathcal{M}_c$, $M_t(A)$ is defined as $\sum \{M(t,t') : t' \in A\}$.

THEOREM 1: For each $t \in T$, M_t is a positive measure on $\mathcal{M}_c$.

Proof: Let $t \in T$ be fixed. It is clear from the remarks in the last paragraph of Section 1.1 that M_t is a function which maps $\mathcal{M}_c$ into $\mathbb{R}^*$, since A is countable iff $A \in \mathcal{M}_c$. Also, M_t is not identically ∞, since $M_t(\emptyset) = \sum \{M(t,t') : t' \in \emptyset\} = 0 < \infty$.

Finally, to show that M_t is countably additive, let $\langle A_n \rangle$ be a disjointed sequence of elements of $\mathcal{M}_c$. For each $n \in \mathbb{N}$, let $t_1^n, t_2^n, t_3^n, \ldots$ be an enumeration of A_n if A_n is

countably infinite, and let $t_1, t_2, \ldots, t_{\|A_n\|}$ be an enumeration of A_n if A_n is finite and non-empty. For each $j,k,m,n \in \mathbb{N}$, if $m \neq n$ or $j \neq k$ and if t_j^m and t_k^n are defined, then $t_j^m \neq t_k^n$ since $\langle A_n \rangle$ is disjointed. Then:

$$M_t\left(\bigcup_{n=1}^{\infty} A_n\right) = M_t(\{t_k^n : n,k \in \mathbb{N} \text{ and } t_k^n \text{ is defined}\})$$
$$= \sum \{M(t,t_k^n) : n,k \in \mathbb{N} \text{ and } t_k^n \text{ is defined}\}$$
$$= \sum_{n=1}^{\infty} \left(\sum \{M(t,t_k^n) : k \in \mathbb{N} \text{ and } t_k^n \text{ is defined}\}\right)$$
$$= \sum_{n=1}^{\infty} \left(\sum \{M(t,t') : t' \in A_n\}\right)$$
$$= \sum_{n=1}^{\infty} M_t(A_n). //$$

DEF'N. 4: Let $A \subseteq T$, let $\{c_1, c_2, \ldots, c_m\} \subseteq \mathbb{R}^*$, and let $\{t_1, t_2, \ldots, t_m\} \subseteq A$. Then $\{(c_k,t_k) : k \in \mathbb{N}, k \leq m\}$ is called an A-combination. For suggestive purposes, the A-combination $\{(c_k,t_k) : k \in \mathbb{N}, k \leq m\}$ is usually written as $\sum_{k=1}^{m} c_k t_k$.

Let $V(A)$ be the set of all A-combinations.

Let $\tau \in V(T)$; say $\tau = \sum_{k=1}^{m} c_k t_k$ where $\{c_1, c_2, \ldots, c_m\} \subseteq \mathbb{R}^*$ and $\{t_1, t_2, \ldots, t_m\} \subseteq T$. For each $A \in \mathcal{M}_c$, $M_\tau(A)$ is defined as $\sum_{k=1}^{m} c_k M_{t_k}(A)$.

REMARK: For each $\tau \in V(T)$, it is clear from LEMMA 2 and THEOREM 1 that M_τ is a positive measure on $\mathcal{M}_c$.

DEF'N. 5: For each $t \in T$, the positive measure M_t on $\mathcal{M}_c$ is called the t-induced M-measure on $\mathcal{M}_c$.

For each $\tau \in V(T)$, the positive measure M_τ on $\mathcal{M}_c$ is called the τ-induced M-measure on $\mathcal{M}_c$.

Section 1.4: Separated Classes and Connected Sets.

In this section, the "touching" relation θ is defined

on $\mathcal{P}(T)$ as described in the last paragraph of Section 1.2. A class $\mathcal{B}$ of subsets of T is separated whenever no two distinct sets in $\mathcal{B}$ touch each other. Let $A \subseteq T$. Then A is connected whenever no decomposition of A containing two or more sets is separated, and an A-component is a maximal connected subset of A. If $t \in T$, the adherency of t is the set of all elements of T which touch t. In this section, these notions are precisely defined, elementary properties regarding these concepts are developed, and these results are applied to the induced M-measures defined in the previous section. THEOREM 3 and THEOREM 5 are the major results in this section.

DEF'N. 6: Let $A, B \in \mathcal{P}(T)$. Then A touches B, written as $A \Theta B$, iff $M(a,b) > 0$ for some $a \in A$ and some $b \in B$.

LEMMA 3: Let $A, B \in \mathcal{P}(T)$. Then:

(i) $A \not\Theta \emptyset$.

(ii) $A \neq \emptyset$ iff $A \Theta A$.

(iii) If $A \Theta B$, then $B \Theta A$.

(iv) If $A_1 \subseteq A$, $B_1 \subseteq B$, and $A_1 \Theta B_1$, then $A \Theta B$.

Proof:

(i) Assume that $A \Theta \emptyset$. Then choose $a \in A$ and $t \in \emptyset$ so that $M(a,t) > 0$. Then $t \in \emptyset$, which is impossible; so $A \not\Theta \emptyset$.

(ii) Suppose that $A \neq \emptyset$. Choose $a \in A$. But $M(a,a) > 0$ by (M-iii); so $A \Theta A$. Conversely, if $A = \emptyset$, then $A \not\Theta A$ by (i).

(iii) Suppose that $A \Theta B$. Choose $a \in A$ and $b \in B$ so that $M(a,b) > 0$. Then $M(b,a) > 0$ by (M-iv); so $B \Theta A$.

(iv) Suppose that $A_1 \subseteq A$, $B_1 \subseteq B$, and $A_1 \Theta B_1$. Choose

$a \in A_1$ and $b \in B_1$ so that $M(a,b) > 0$. But $a \in A$ and $b \in B$, so $A \Theta B$.//

DEF'N. 7: Let $\mathcal{B} \subseteq \mathcal{P}(T)$. Then $\mathcal{B}$ is separated iff for each $B_1, B_2 \in \mathcal{B}$, if $B_1 \neq B_2$ then $B_1 \emptyset B_2$.

LEMMA 4: Let $\mathcal{B}, \mathcal{B}' \subseteq \mathcal{P}(T)$. Then:

(i) If $\mathcal{B}' \subseteq \mathcal{B}$ and $\mathcal{B}$ is separated, then $\mathcal{B}'$ is separated.

(ii) If $\mathcal{B}$ is separated, then $\mathcal{B}$ is disjointed.

(iii) If $\mathcal{B}_1 \cup \mathcal{B}_2 \subseteq \mathcal{B}$, $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$, and $\mathcal{B}$ is separated, then $\cup\mathcal{B}_1 \emptyset \cup\mathcal{B}_2$.

Proof:

(i) Suppose that $\mathcal{B}' \subseteq \mathcal{B}$ and $\mathcal{B}$ is separated. Let $B_1, B_2 \in \mathcal{B}'$, and suppose that $B_1 \neq B_2$. But then $B_1, B_2 \in \mathcal{B}$, and $\mathcal{B}$ is separated; so $B_1 \emptyset B_2$. Hence $\mathcal{B}'$ is separated.

(ii) Suppose that $\mathcal{B}$ is not disjointed. Then choose $B_1, B_2 \in \mathcal{B}$ so that $B_1 \neq B_2$ and $B_1 \cap B_2 \neq \emptyset$. Choose $t \in B_1 \cap B_2$. Then $M(t,t) > 0$ by (M-iii), so $B_1 \Theta B_2$. Hence, $\mathcal{B}$ is not separated.

(iii) Suppose that $\mathcal{B}_1 \cup \mathcal{B}_2 \subseteq \mathcal{B}$, $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$, and $\mathcal{B}$ is separated. Assume that $\cup\mathcal{B}_1 \Theta \cup\mathcal{B}_2$. Choose $b_1 \in \cup\mathcal{B}_1$ and $b_2 \in \cup\mathcal{B}_2$ so that $M(b_1,b_2) > 0$. Choose $B_1 \in \mathcal{B}_1$ and $B_2 \in \mathcal{B}_2$ so that $b_1 \in B_1$ and $b_2 \in B_2$. Hence, $B_1 \Theta B_2$. But $B_1, B_2 \in \mathcal{B}$ since $\mathcal{B}_1 \cup \mathcal{B}_2 \subseteq \mathcal{B}$, and $B_1 \neq B_2$ since $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$; so $\mathcal{B}$ is not separated, which contradicts the hypothesis. Thus, $\cup\mathcal{B}_1 \emptyset \cup\mathcal{B}_2$.//

THEOREM 2: Let $\mathcal{B}$ be a separated class of elements of $\mathcal{M}_c$, and let $A \in \mathcal{M}_c$. Let $B \in \mathcal{B}$ be fixed. Then:

(i) For each $t \in T$, $M_t(A) = 0$ iff $\{t\} \emptyset A$.

(ii) For each $t\in B$, $\{t\}\not\emptyset\cup\mathcal{B}-B$.

(iii) For each $t\in B$, $\{t\}\not\emptyset A\cap(\cup\mathcal{B}-B)$.

(iv) For each $t\in B$, if $\cup\mathcal{B}\in\mathcal{M}_c$ then $M_t(\cup\mathcal{B})=M_t(B)$.

(v) For each $t\in B$, $M_t(A-\cup\mathcal{B})=M_t(A-B)$.

Proof:

(i) Let $t\in T$. Then:

$M_t(A)=0$ iff $\sum\{M(t,a): a\in A\}=0$,

iff $M(t,a)=0$ for each $a\in A$,

iff $\{t\}\not\emptyset A$.

(ii) Let $t\in B$. Now $B\not\emptyset\cup(\mathcal{B}-\{B\})$ by LEMMA 4-iii since $\mathcal{B}$ is separated and since $(\mathcal{B}-\{B\})\cap\{B\}=\emptyset$. But $\mathcal{B}$ is disjointed by LEMMA 4-ii since $\mathcal{B}$ is separated; so $\cup(\mathcal{B}-\{B\})=\cup\mathcal{B}-B$. Hence, $B\not\emptyset\cup\mathcal{B}-B$. But $\{t\}\subseteq B$; so by LEMMA 3-iv, $\{t\}\not\emptyset\cup\mathcal{B}-B$.

(iii) Let $t\in B$. By (ii) and LEMMA 3-iv, it follows immediately that $\{t\}\not\emptyset A\cap(\cup\mathcal{B}-B)$.

(iv) Let $t\in B$, and suppose that $\cup\mathcal{B}\in\mathcal{M}_c$. Then:

$M_t(\cup\mathcal{B})=M_t((\cup\mathcal{B}-B)\cup B)$

$=M_t(\cup\mathcal{B}-B)+M_t(B)$ (by LEMMA 1-ii and THEOREM 1)

$=M_t(B)$ ($M_t(\cup\mathcal{B}-B)=0$ by (i), since $\{t\}\not\emptyset\cup\mathcal{B}-B$ by (ii)).

(v) Let $t\in B$. Now $(A-B)-\cup\mathcal{B}=A-\cup\mathcal{B}$, since $B\subseteq\cup\mathcal{B}$.

Also, $(A-B)\cap\cup\mathcal{B}=(A\cap\cup\mathcal{B})-B=A\cap(\cup\mathcal{B}-B)$, since $B\subseteq\cup\mathcal{B}$.

$M_t(A-B)=M_t(((A-B)-\cup\mathcal{B})\cup((A-B)\cap\cup\mathcal{B})$

$=M_t((A-B)-\cup\mathcal{B})+M_t((A-B)\cap\cup\mathcal{B})$

$=M_t(A-\cup\mathcal{B})+M_t(A\cap(\cup\mathcal{B}-B))$

$=M_t(A-\cup\mathcal{B})$ ($M_t(A\cap(\cup\mathcal{B}-B))=0$ by (i), since $t\not\emptyset A\cap(\cup\mathcal{B}-B)$ by (iii)).//

DEF'N. 8: Let $A \in \mathcal{P}(T)$. Then A is connected iff for each decomposition $\mathcal{B}$ of A, if $\|\mathcal{B}\| \geq 2$ then $\mathcal{B}$ is not separated; and A is disconnected iff A is not connected.

Let $\mathcal{C}_A$ be the class of all connected subsets of A.

LEMMA 5: For each $t \in T$, $\{t\}$ is connected; and $\emptyset$ is connected.

Proof: Let $t \in T$. The only decomposition of $\{t\}$ is $\{\{t\}\}$; and since $\|\{\{t\}\}\| = 1$, $\{t\}$ is trivially connected. Now $\emptyset$ has no decomposition and is also trivially connected.//

LEMMA 6: Let $A_1, A_2 \in \mathcal{P}(T)$ be non-empty and connected, and suppose that $A_1 \cup A_2$ is disconnected. Then $A_1 \emptyset A_2$.

Proof: Assume that $A_1 \Theta A_2$. Choose $a_1 \in A_1$ and $a_2 \in A_2$ so that $M(a_1, a_2) > 0$. Now $A_1 \cup A_2$ is disconnected, so choose a decomposition $\mathcal{B}$ of $A_1 \cup A_2$ so that $\|\mathcal{B}\| \geq 2$ and $\mathcal{B}$ is separated. Choose $B_1, B_2 \in \mathcal{B}$ so that $a_1 \in B_1$ and $a_2 \in B_2$. But $M(a_1, a_2) > 0$, so $B_1 \Theta B_2$. Since $\mathcal{B}$ is separated, $B_1 = B_2$.

Let $\mathcal{B}_1 = \{B \cap A_1 : B \in \mathcal{B}\} - \{\emptyset\}$, $\mathcal{B}_2 = \{B \cap A_2 : B \in \mathcal{B}\} - \{\emptyset\}$. Clearly $\mathcal{B}_1$ is a decomposition of A_1. For each $B', B'' \in \mathcal{B}$, if $B' \neq B''$, then $B' \emptyset B''$ since $\mathcal{B}$ is separated; so $B' \cap A_1 \emptyset B'' \cap A_1$ by LEMMA 3-iv. Hence, by LEMMA 4-i, $\mathcal{B}_1$ is a separated decomposition of A_1. But A_1 is a connected non-empty set, so $\|\mathcal{B}_1\| = 1$. Hence, $\mathcal{B}_1 = \{B_1 \cap A_1\}$; so $B_1 \cap A_1 = A_1$. Similarly, $B_2 \cap A_2 = A_2$. From the previous paragraph, $B_1 = B_2$; so $A_1 \cup A_2 \subseteq B_1$. But $B_1 \in \mathcal{B}$ and $\mathcal{B}$ is a decomposition of $A_1 \cup A_2$; so $B_1 = A_1 \cup A_2$. Thus, $\mathcal{B} = \{A_1 \cup A_2\}$; so $\|\mathcal{B}\| = 1$, which contradicts the choice of $\mathcal{B}$.//

DEF'N. 9: Let $A, B \in \mathcal{P}(T)$, Then B is an A-component iff B is maximal in the partial ordering $(\mathcal{C}_A, \subseteq)$.

B is a <u>component</u> (or a <u>genus</u>) iff B is a T-component.

<u>THEOREM 3</u>: Let $A \in \mathcal{P}(T)$, and let $\mathcal{J}_A$ be the class of all A-components. Then:

(i) $\emptyset \in \mathcal{J}_A$ iff $A=\emptyset$ iff $\mathcal{J}_A = \emptyset$.

(ii) $\mathcal{J}_A$ is separated.

(iii) $A=\cup \mathcal{J}_A$.

(iv) $\mathcal{J}_A$ is a decomposition of A iff $A \neq \emptyset$.

(v) If A_1 and A_2 are distinct A-components, then $M(a_1,a_2)=0$ for each $a_1 \in A_1$ and each $a_2 \in A_2$.

<u>Proof</u>:

(i) Suppose that $A \neq \emptyset$. Choose $t \in A$. Now $\{t\}$ is connected by LEMMA 5, and $\emptyset \subseteq \{t\}$; so $\emptyset$ is not maximal in $(\mathcal{C}_A, \subseteq)$. Hence, $\emptyset \notin \mathcal{J}_A$.

Suppose that $A=\emptyset$. By LEMMA 5, $\emptyset$ is connected; and $\emptyset$ is the only maximal element in $(\mathcal{C}_A, \subseteq)$. Hence, $\mathcal{J}_A=\{\emptyset\}$.

Clearly, if $\mathcal{J}_A=\{\emptyset\}$, then $\emptyset \in \mathcal{J}_A$.

(ii) Let $A_1, A_2 \in \mathcal{J}_A$ and suppose that $A_1 \neq A_2$. Hence, $\emptyset \notin \mathcal{J}_A$ by (i) since $A \neq \emptyset$. Thus, $A_1 \neq \emptyset$ and $A_2 \neq \emptyset$.

A_1 and A_2 are connected since both are elements of $\mathcal{J}_A$.

Now $A_1 \cup A_2 \subseteq A$ since A_1 and A_2 are subsets of A, and $A_1 \neq A_2$; so $A_1 \cup A_2$ properly includes A_1. But A_1 is a maximal connected subset of A; so $A_1 \cup A_2$ is disconnected.

Thus, $A_1 \not\emptyset A_2$ by LEMMA 6. Hence, $\mathcal{J}_A$ is separated.

(iii) To show that $A \subseteq \cup \mathcal{J}_A$: Let $t \in A$. By LEMMA 5, $\{t\}$ is a connected subset of A; so choose $A_1 \in \mathcal{J}_A$ so that $\{t\} \subseteq A_1$. Then $t \in A_1 \subseteq \cup \mathcal{J}_A$.

To show that $\cup \mathcal{J}_A \subseteq A$: For each $A_1 \in \mathcal{J}_A$, $A_1 \subseteq A$ since

A_1 is an A-component; so $\bigcup \mathcal{J}_A \subseteq A$.

(iv) If $A=\emptyset$, then $\emptyset \in \mathcal{J}_A$ by (i); so $\mathcal{J}_A$ is not a decomposition of A.

Conversely, suppose that $A \neq \emptyset$. By (i), $\emptyset \notin \mathcal{J}_A$. $\mathcal{J}_A$ is separated by (ii); so $\mathcal{J}_A$ is disjointed by LEMMA 4-ii. Finally, $A=\bigcup \mathcal{J}_A$ by (iii); so $\mathcal{J}_A$ is a decomposition of A.

(v) Let A_1 and A_2 be distinct A-components. By (ii) $\mathcal{J}_A$ is separated; so $A_1 \emptyset A_2$. Hence, $M(a_1,a_2)=0$ for each $a_1 \in A_1$ and each $a_2 \in A_2$.//

THEOREM 4: Let $A \in \mathcal{M}_C$, let $C \in \mathcal{P}(T)$, and suppose that $A \subseteq C$. Let B be a C-component. Then:

(i) $B \emptyset A-B$.

(ii) For each $t \in B$, $M_t(A)=M_t(A \cap B)$.

Proof:

(i) Let $\mathcal{J}_C$ be the class of all C-components. Then $B \in \mathcal{J}_C$; and by THEOREM 3-iii, $C=\bigcup \mathcal{J}_C$. By THEOREM 3-ii, $\mathcal{J}_C$ is separated; so $B \emptyset \bigcup(\mathcal{J}_C-\{B\})$ by LEMMA 4-iii since $\{B\} \cap (\mathcal{J}_C-\{B\}) \neq \emptyset$.

Now $\mathcal{J}_C$ is disjointed by LEMMA 4-ii since $\mathcal{J}_C$ is separated; so $\bigcup(\mathcal{J}_C-\{B\})=\bigcup \mathcal{J}_C-B$. Hence, $\bigcup(\mathcal{J}_C-B)=C-B$ since $C=\bigcup \mathcal{J}_C$; so $B \emptyset C-B$. But $A-B \subseteq C-B$ since $A \subseteq C$; so by LEMMA 3-iv, $B \emptyset A-B$.

(ii) A is countable since $A \in \mathcal{M}_C$; so A—B is countable. Hence $A-B \in \mathcal{M}_C$. Let $t \in B$. By (i), $B \emptyset A-B$; so by LEMMA 3-iv, $\{t\} \emptyset A-B$. Thus, $M_t(A-B)=0$ by THEOREM 2-i. Then:

$M_t(A)=M_t(A \cap B)+M_t(A-B)$ (by LEMMA 1-ii and THEOREM 1)

$=M_t(A \cap B)$ (since $M_t(A-B)=0$).//

DEF'N. 10: For each $t\in T$, the adherency of t, written as K_t, is defined as $\{t'\in T: M(t,t')>0\}$.

LEMMA 7: Let $t\in T$. Then:

(i) For each $A\in\mathcal{M}_c$, $M_t(A)=M_t(K_t\cap A)$.

(ii) For each $A\in\mathcal{M}_c$, $M_t(A)\leq M_t(K_t)$ if $K_t\in\mathcal{M}_c$.

(iii) If K_t is uncountable, then $M_t(A)=\infty$ for some $A\in\mathcal{M}_c$.

Proof:

(i) Let $A\in\mathcal{M}_c$. Then for each $t'\in A-K_t$, $M(t,t')=0$; so $M_t(A-K_t)=\sum\{M(t,t'): t'\in A-K_t\}=0$. Thus:

$M_t(A)=M_t(A\cap K_t)+M_t(A-K_t)$ (by LEMMA 1-ii and THEOREM 1)

$=M_t(A\cap K_t)$ (since $M_t(A-K_t)=0$).

(ii) Let $A\in\mathcal{M}_c$; and suppose that $K_t\in\mathcal{M}_c$. Then:

$M_t(A)=M_t(A\cap K_t)$ (by (i))

$\leq M_t(K_t)$ (by LEMMA 1-iv and THEOREM 1, since $K_t\in\mathcal{M}_c$).

(iii) Suppose that K_t is uncountable. For each $n\in\mathbb{N}$: if $n=1$, let $I_n=I_1=\{x\in\mathbb{R}^*: 2^{-1}<x\leq\infty\}$; and if $n\geq 2$, let $I_n=\{x\ \mathbb{R}: 2^{-n}<x\leq 2^{-n+1}\}$.

Then $\{I_n: n\in\mathbb{N}\}$ is a countable decomposition of $\mathbb{R}^*-\{0\}$, and for each $t'\in T$, $t'\in K_t$ iff there is a unique $n\in\mathbb{N}$ for which $M(t,t')\in I_n$; so for each $n\in\mathbb{N}$, let $K_t(n)=\{t'\in K_t: M(t,t')\in I_n\}$. Hence, $\{K_t(n): n\in\mathbb{N}\}-\{\emptyset\}$ is clearly a countable decomposition of K_t. But K_t is uncountable; so choose $m\in\mathbb{N}$ for which $K_t(m)$ is uncountable. Then for each $t'\in K_t(m)$, $M(t,t')\geq 2^{-m}$ since $M(t,t')\in I_m$.

Choose A to be any countably infinite subset of $K_t(m)$. Then for each $t'\in A$, $M(t,t')\geq 2^{-m}$; so $M_t(A)=\sum\{M(t,t'): t'\in A\}\geq\sum_{n=1}^{\infty}2^{-m}=\infty$. Thus, $M_t(A)=\infty$.//

THEOREM 5: For each $t \in T$, there is some $A \in \mathcal{M}_c$ such that $t \in A$ and such that $M_t(B—A) \leqslant M_t(A)$ for each $B \in \mathcal{M}_c$.

Proof: Let $t \in T$. If K_t is uncountable, choose $C \in \mathcal{M}_c$ so that $M_t(C) = \infty$ by LEMMA 7-iii. Let $A = C \cup \{t\}$. Then $M_t(A) = \infty$; so $M_t(B—A) \leqslant \infty = M_t(A)$ for each $B \in \mathcal{M}_c$. On the other hand, if K_t is countable, let $A = K_t$. Then $t \in A$. Let $B \in \mathcal{M}_c$. For each $t' \in B—A$, $t' \notin K_t$; so $M(t,t') = 0$. Thus, $M_t(B—A) = 0 \leqslant M_t(A)$.//

Section 1.5: R-clumps and Strong Clumps.

Using the results of the two previous sections, it is now possible to introduce and develop the notions which are central to the theory of clumps. The notion of an R-clump was first introduced in a less general form by R. M. Needham in [5] and [4] and was modified in [1] by others. A non-empty measurable set A (i.e., a non-empty set $A \in \mathcal{M}_c$) is an R-clump iff for each $B \in \mathcal{M}_c$ and for each $t \in A$, $M_t(B—A) \leqslant M_t(A)$. In terms of the structure (T,M), if A is an R-clump and if $t \in A$, then t touches A more strongly than any measurable relative complement of A. In a sense, A "clumps" together; i.e., coheres internally more strongly than any of its points adheres externally to any measurable relative complement. The existence of an R-clump cover of T is guaranteed by THEOREM 8.

THEOREM 9 shows that any union of R-clumps is also an R-clump if the union is measurable (i.e., countable), and if the union over a separated class is an R-clump then each set in that class is an R-clump. An atomic R-clump is one

which is not expressible as the union of two distinct properly included R-clumps, while a non-atomic R-clump is one which is so expressible. The notions of atomic and non-atomic R-clumps comprise the basis for generating all the R-clumps provided that M_t is bounded for each $t\in T$. This fundamental result is given in THEOREM 13. Although every minimal R-clump is atomic, the converse is not true in general. Also, no general relationship holds between the boundedness of each M_t for $t\in T$ and the boundedness of the function M on $T\times T$.

A strong clump is a connected R-clump. It is this notion which is of primary interest in Chapter 2.

DEF'N. 11: Let $A\in\mathcal{M}_c$ be non-empty. Then A is an R-clump iff for each $B\in\mathcal{M}_c$ and for each $t\in A$, $M_t(B-A)\leqslant M_t(A)$.

Let $\mathcal{R}$ be the class of all R-clumps in T.

THEOREM 6: Let $A\in\mathcal{M}_c$ be non-empty. Then A is an R-clump iff for each $B\in\mathcal{M}_c$ and for each $\tau\in V(A)$, $M_\tau(B-A)\leqslant M(A)$.

Proof: Suppose that A is an R-clump. Let $B\in\mathcal{M}_c$ and let $\tau\in V(A)$, say $\tau=\sum_{k=1}^{m} c_k t_k$ where $\{c_1, c_2,\ldots, c_m\}\subseteq\mathbb{R}^*$ and $\{t_1, t_2,\ldots, t_m\}\subseteq A$. For $k=1,2,\ldots,m$, $t_k\in A$; so $M_{t_k}(B-A)\leqslant M_{t_k}(A)$ since A is an R-clump. Since each $c_k\geqslant 0$, then $c_k M_{t_k}(B-A)\leqslant c_k M_{t_k}(A)$; so $\sum_{k=1}^{m} c_k M_{t_k}(B-A)\leqslant\sum_{k=1}^{m} c_k M_{t_k}(A)$. Thus, $M_\tau(B-A)\leqslant M_\tau(A)$.

Conversely, let $B\in\mathcal{M}_c$ and suppose that $M_\tau(B-A)\leqslant M_\tau(A)$ for each $\tau\in V(A)$. Let $t\in A$. Then for $m=1$, $c_1=1$, and $t_1=t$, if $\tau=\sum_{k=1}^{m} c_k t_k$, then $M_t(B-A)=M_\tau(B-A)\leqslant M_\tau(A)=M_t(A)$. Thus, A is an R-clump.//

THEOREM 7: Let $A \in \mathcal{M}_c$ be non-empty, and suppose that T is countable. Then the following conditions are equivalent:

(i) A is an R-clump.

(ii) For each $\tau \in V(A)$, $M_\tau(T—A) \leq M_\tau(A)$.

(iii) For each $t \in A$, $M_t(T—A) \leq M_t(A)$.

Proof:

To show that (i) implies (ii): Suppose that A is an R-clump. Since T is countable, $T \in \mathcal{M}_c$; so by THEOREM 6, $M_\tau(T—A) \leq M_\tau(A)$ for each $\tau \in V(A)$.

To show that (ii) implies (iii): Suppose that for each $\tau \in V(A)$, $M_\tau(T—A) \leq M_\tau(A)$. Let $t \in A$. Then for $m=1$, $c_1=1$, and $t_1=t$, if $\tau = \sum_{k=1}^{m} c_k t_k$, $M_t(T—A)=M_\tau(T—A) \leq M_\tau(A)=M_t(A)$.

To show that (iii) implies (i): Let $t \in T$, and suppose that $M_t(T—A) \leq M_t(A)$. Let $B \in \mathcal{M}_c$. Since $B \subseteq T$, $B—A \subseteq T—A$; so $M_t(B—A) \leq M_t(T—A)$ by LEMMA 1-iv and THEOREM 1. Hence, $M_t(B—A) \leq M_t(A)$; so A is an R-clump.//

THEOREM 8: For each $a \in T$, there is some $A \in \mathcal{R}$ for which $a \in A$.

Proof: Let $a \in T$ be fixed. Let $A_0=\{a\}$. A sequence $\langle A_n \rangle$ of subsets of T is constructed so that, $n \in \mathbb{N}$:

(i) $A_{n-1} \subseteq A_n$;

(ii) $A_n \in \mathcal{M}_c$; and

(iii) For each $t \in A_{n-1}$ and each $B \in \mathcal{M}_c$, $M_t(B—A_n) \leq M_t(A_n)$.

This is done by induction on $\mathbb{N}$ as follows:

By THEOREM 5, choose $A_1 \in \mathcal{M}_c$ so that $a \in A_1$ and so that $M_a(B—A_1) \leq M_a(A_1)$ for each $B \in \mathcal{M}_c$. Since $A_0=\{a\}$, conditions (i), (ii), and (iii) hold when $n=1$.

Let $n \in \mathbb{N}$; and suppose that A_n is chosen so that (i),

(ii), and (iii) hold. By THEOREM 5, for each $t\in A_n$, choose $A_{n+1}(t)\in\mathcal{M}_c$ so that $t\in A_{n+1}(t)$ and so that $M_t(B-A_{n+1}(t)) \leqslant M_t(A_{n+1}(t))$ for each $B\in\mathcal{M}_c$. Let $A_{n+1}=\bigcup\{A_{n+1}(t): t\in A_n\}$.

For each $t\in A_n$, $t\in A_{n+1}(t)\subseteq A_{n+1}$, so $A_n\subseteq A_{n+1}$; and (i) holds.

For each $t\in A_n$, $A_{n+1}(t)$ is countable by the choice of $A_{n+1}(t)$, and A_n is countable since $A_n\in\mathcal{M}_c$ by the induction hypothesis; so A_{n+1} is a countable union of countable sets. Thus, A_{n+1} is countable, so $A_{n+1}\in\mathcal{M}_c$; and (ii) holds.

Let $t\in A_n$ and let $B\in\mathcal{M}_c$. $A_{n+1}(t)\subseteq\bigcup\{A_{n+1}(t'): t'\in A_n\}$ $=A_{n+1}$; so $B-A_{n+1}\subseteq B-A_{n+1}(t)$. Hence, $M_t(B-A_{n+1})$ $\leqslant M_t(B-A_{n+1}(t))$ and $M_t(A_{n+1}(t))\leqslant M_t(A_{n+1})$ by LEMMA 1-iv and THEOREM 1. But by the choice of $A_{n+1}(t)$, $M_t(B-A_{n+1}(t))$ $\leqslant M_t(A_{n+1}(t))$; so $M_t(B-A_{n+1})\leqslant M_t(A_{n+1})$, and (iii) holds.

Now let $A=\bigcup_{n=1}^{\infty}A_n$. By (ii), each $A_n\in\mathcal{M}_c$; so $A\in\mathcal{M}_c$. Also, $a\in A_0\subseteq A_1\subseteq A$. Let $t\in A$ and let $B\in\mathcal{M}_c$. Choose $k\in\mathbb{N}$ so that $t\in A_{k-1}$. By (iii), $M_t(B-A_k)\leqslant M_t(A_k)$. By (i), $\langle A_n\rangle$ is an increasing sequence of sets; so $A_k\subseteq\bigcup_{n=1}^{\infty}A_n=A$ and $B-A\subseteq B-A_k$. Hence, $M_t(B-A)\leqslant M_t(B-A_k)$ and $M_t(A_k)\leqslant M_t(A)$ by LEMMA 1-iv and THEOREM 1. Thus, $M_t(B-A)\leqslant M_t(A)$; so A is an R-clump for which $a\in A$.//

THEOREM 9: Suppose that $\mathcal{B}$ is a class of non-empty sets in $\mathcal{M}_c$. Then:

(i) If $\mathcal{B}\subseteq\mathcal{R}$ and $\cup\mathcal{B}\in\mathcal{M}_c$, then $\cup\mathcal{B}\in\mathcal{R}$.

(ii) If $\mathcal{B}$ is separated and $\cup\mathcal{B}\in\mathcal{R}$, then $\mathcal{B}\subseteq\mathcal{R}$.

Proof:

(i) Suppose that $\mathcal{B}\subseteq\mathcal{R}$ and $\cup\mathcal{B}\in\mathcal{M}_c$. Let $t\in\cup\mathcal{B}$ and

let $B \in \mathcal{M}_c$. Choose $B' \in \mathcal{B}$ so that $t \in B'$. Then $B' \subseteq \cup\mathcal{B}$ and $B - \cup\mathcal{B} \subseteq B - B'$; so $M_t(B - \cup\mathcal{B}) \leq M_t(B - B')$ and $M_t(B') \leq M_t(\cup\mathcal{B})$ by LEMMA 1-iv and THEOREM 1 since $\cup\mathcal{B} \in \mathcal{M}_c$. But B' is an R-clump since $B' \in \mathcal{B} \subseteq \mathcal{R}$; so $M_t(B - B') \leq M_t(B')$. Thus, $M_t(B - \cup\mathcal{B}) \leq M_t(\cup\mathcal{B})$; so $\cup\mathcal{B} \in \mathcal{R}$.

(ii) Suppose that $\mathcal{B}$ is separated and that $\cup\mathcal{B} \in \mathcal{R}$. Let $B' \in \mathcal{B}$ be fixed. Let $t \in B'$ and let $B \in \mathcal{M}_c$. Since $t \in B' \subseteq \cup\mathcal{B}$ and since $\cup\mathcal{B}$ is an R-clump, $M_t(B - \cup\mathcal{B}) \leq M_t(\cup\mathcal{B})$. Now $\cup\mathcal{B} \in \mathcal{M}_c$ since $\cup\mathcal{B} \in \mathcal{R}$; so by THEOREM 2-iv and THEOREM 2-v, $M_t(\cup\mathcal{B}) = M_t(B')$ and $M_t(B - \cup\mathcal{B}) = M_t(B - B')$. Thus, $M_t(B - B') \leq M_t(B')$; so B' is an R-clump. But $B' \in \mathcal{B}$ is arbitrary; so $\mathcal{B} \subseteq \mathcal{R}$. //

<u>THEOREM 10</u>: Let $A \in \mathcal{R}$, let $t \in T - A$, and suppose that $\{t\} \notin \mathcal{R}$.

(i) If $M_t(B - (A \cup \{t\})) \leq M_t(A \cup \{t\})$ for each $B \in \mathcal{M}_c$, then $A \cup \{t\} \in \mathcal{R}$.

(ii) If $A \cup \{t\} \in \mathcal{R}$, then $A \theta \{t\}$.

<u>Proof</u>:

(i) Let $B \in \mathcal{M}_c$, and suppose that $M_t(B - (A \cup \{t\})) \leq M_t(A \cup \{t\})$. Let $a \in A$. Then $A \subseteq A \cup \{t\}$ and $B - (A \cup \{t\}) \subseteq B - A$; so by LEMMA 1-iv and THEOREM 1, $M_a(B - (A \cup \{t\})) \leq M_a(B - A)$ and $M_a(A) \leq M_a(A \cup \{t\})$. Now A is an R-clump, so $M_a(B - A) \leq M_a(A)$. Hence, $M_a(B - (A \cup \{t\})) \leq M_a(A \cup \{t\})$ for each $a \in A$. It follows that $A \cup \{t\}$ is an R-clump.

(ii) Suppose that $A \cup \{t\} \in \mathcal{R}$, and assume that $A \not\theta \{t\}$. Then $\{A, \{t\}\}$ is separated; so $\{t\} \in \mathcal{R}$ by THEOREM 9-ii. But this contradicts the hypothesis that $\{t\} \notin \mathcal{R}$; so it follows that $A \theta \{t\}$. //

DEF'N. 12: Let $A\in\mathcal{R}$. Then A is an atomic R-clump iff for each $A_1, A_2\in\mathcal{R}$, if A_1 and A_2 are distinct from A and $A_1\neq A_2$, then $A\neq A_1\cup A_2$; and A is a non-atomic R-clump iff A is not atomic.

If $C\in\mathcal{P}(T)$, let $\mathcal{A}_C$ be the class of all atomic R-clumps included in C.

Let $\mathcal{A}=\mathcal{A}_T$ (i.e., the class of all atomic R-clumps in T).

THEOREM 11: If A is an atomic R-clump, then A is connected.

Proof: Suppose that A is a disconnected R-clump. Since A is disconnected, choose a decomposition $\mathcal{B}$ of A so that $\mathcal{B}$ is separated and $\|\mathcal{B}\|\geq 2$. But $\cup\mathcal{B}=A\in\mathcal{R}$; so by THEOREM 9-ii, $\mathcal{B}\subseteq\mathcal{R}$. Since $\|\mathcal{B}\|\geq 2$, choose $B\in\mathcal{B}$. Then $\|\mathcal{B}-\{B\}\|\geq 1$. But $\mathcal{B}$ is a decomposition of A; so $\mathcal{B}-\{B\}$ contains only non-empty sets. Since $\mathcal{B}-\{B\}\subseteq\mathcal{B}\subseteq\mathcal{R}$ and $\cup(\mathcal{B}-\{B\})\in\mathcal{M}_c$, it follows by THEOREM 9-i that $\cup(\mathcal{B}-\{B\})\in\mathcal{R}$. Also $B\in\mathcal{B}\subseteq\mathcal{R}$. But $\cup(\mathcal{B}-\{B\})$ and B are distinct sets, since $B\in\mathcal{B}$ and since $\mathcal{B}$ is a decomposition of A; so $A=B\cup\cup(\mathcal{B}-\{B\})$ is clearly a non-atomic R-clump.//

LEMMA 8: If A is a countable component, then A is an R-clump.

Proof: Suppose that A is a countable component. Let $B\in\mathcal{M}_c$, and let $t\in A$. Now A is a T-component and $A\subseteq T$; so $M_t(B-A)=M_t((B-A)\cap A)$ by THEOREM 4-ii. But $(B-A)\cap A=\emptyset$; so it follows that $M_t(B-A)=M_t(\emptyset)=0\leq M_t(A)$. Thus, A is an R-clump.//

LEMMA 9: Let B be a component, and let $t\in B$. Then $K_t\subseteq B$.

Proof: Assume that $K_t\not\subseteq B$. Choose $a\in K_t$ so that $a\notin B$. Then by THEOREM 4-ii, $M_t(\{a\})=M_t(\{a\}\cap B)$. But $\{a\}\cap B=\emptyset$ since $a\notin B$; so $M(t,a)=M_t(\emptyset)=0$. Hence $a\notin K_t$, which is a contradiction.//

THEOREM 12: Suppose that M_t is bounded for each $t\in T$. Then every component is countable.

Proof: Suppose that B is any component. $T\neq\emptyset$ by (M-i); so $\emptyset$ is not a component by THEOREM 3-i. Hence, $B\neq\emptyset$; so choose $a\in B$. By LEMMA 7-iii for each $t\in T$, K_t is countable since M_t is bounded.

Let $A_1=K_a$. Then $A_1\subseteq B$ by LEMMA 9 since $a\in B$; and since $A_1=K_a$ is countable, $A_1\in\mathcal{M}_c$.

Let $n\in\mathbb{N}$; and suppose that A_n has been chosen so that $A_n\subseteq B$ and $A_n\in\mathcal{M}_c$. Let $A_{n+1}=\bigcup\{K_t: t\in A_n\}$. Now for each $t\in A_n\subseteq B$, $K_t\subseteq B$ by LEMMA 9; so $A_{n+1}\subseteq B$. For each $t\in A_n$, K_t is countable; and A_n is countable since $A\in\mathcal{M}_c$. Hence, A_{n+1} is a countable union of countable sets; so $A_{n+1}\in\mathcal{M}_c$.

Let $A=\bigcup_{n=1}^{\infty}A_n$. Then $A\subseteq B$ and $A\in\mathcal{M}_c$.

Assume that $B\neq A$. Since $A\subseteq B$, $B-A\neq\emptyset$; and since $a\in K_a\subseteq A_1\subseteq\bigcup_{n=1}^{\infty}A_n\subseteq A$, $A\neq\emptyset$. Hence, $\{A, B-A\}$ is a decomposition of B. Then $\{A, B-A\}$ is not separated since B is connected; so $A\theta B-A$. Choose $t_1\in A$ and $t_2\in B-A$ so that $M(t_1,t_2)>0$. Then choose $k\in\mathbb{N}$ so that $t_1\in A_k$. Then $t_2\in K_{t_1}\subseteq A_{n+1}\subseteq A$. But $t_2\notin A$ since $t_2\in B-A$, which is a contradiction. Hence, $B=A$. But $A\in\mathcal{M}_c$; so B is countable.//

LEMMA 10: Suppose that M_t is bounded for each $t\in T$. Suppose that $\langle C_n\rangle$ is a decreasing sequence of R-clumps. Let $C=\bigcap_{n=1}^{\infty}C_n$, and suppose that $C\neq\emptyset$. Then C is an R-clump.

Proof: Since M_t is a bounded function for each $t\in T$, it is clear from THEOREM 12 that one may assume, without loss of generality, that $T\in\mathcal{M}_c$.

Let $t \in T$ be fixed. Since M_t is a bounded function, $M_t(C_1) < \infty$, and $\langle C_n \rangle$ is a decreasing sequence of elements of $\mathcal{M}_c$; so by LEMMA 1-vii and THEOREM 1, $\langle M_t(C_n) \rangle$ converges monotonely from above to $M_t(C)$.

Also, $\langle T - C_n \rangle$ is an increasing sequence of elements of $\mathcal{M}_c$ since $T \in \mathcal{M}_c$; so by LEMMA 1-vi and THEOREM 1, $\langle M_t(T - C_n) \rangle$ converges monotonely from below to $M_t(T - C)$.

But for each $n \in \mathbb{N}$, $M_t(T - C_n) \leqslant M_t(C_n)$ by THEOREM 7 since each C_n is an R-clump and since T is countable; so by the "sandwiching" property of convergent sequences, $M_t(T - C) \leqslant M_t(C)$. By THEOREM 7 and since $C \neq \emptyset$, C is an R-clump.//

THEOREM 13: Suppose that M_t is bounded for each $t \in T$. Then every R-clump is a union of atomic R-clumps.

Proof: As in the proof of LEMMA 10, assume without loss of generality that T is countable.

Assume for some $C \in \mathcal{R}$ that C is not a union of atomic R-clumps. Then $\bigcup \mathcal{A}_C$ is properly included in C; so choose $t \in C$ so that $t \notin \bigcup \mathcal{A}_C$. Thus, t is a member of no atomic R-clump included in C. Now proceed by ordinal induction:

Define $C_0 = C$. Then $t \in C_0$ and C_0 is a non-atomic R-clump.

Let $\nu > 0$ be any countable ordinal; and suppose that $C_{\nu'}$ is defined so that $C_{\nu'}$ properly includes C_ν for each $\nu' < \nu$, $t \in C_\nu$, and C_ν is a non-atomic R-clump. Since C_ν is a non-atomic R-clump, choose $A_1, A_2 \in \mathcal{R}$ so that A_1 and A_2 are distinct from C_ν and $C_\nu = A_1 \cup A_2$. If $t \in A_1$, let $C_{\nu+1} = A_1$; and if $t \in A_2 - A_1$, let $C_{\nu+1} = A_2$. By the choice of A_1 and A_2,

C_ν properly includes $C_{\nu+1}$; and so $C_{\nu'}$ properly includes $C_{\nu+1}$ for each $\nu'<\nu+1$. Also $t\in C_{\nu+1}$; and by the choice of t, $C_{\nu+1}$ is hence a non-atomic R-clump.

Now let λ be any countable limit ordinal; and suppose that C_ν is defined for each ordinal $\nu<\lambda$ so that $C_{\nu'}$ properly includes C_ν for each $\nu'<\nu$, $t\in C_\nu$, and C_ν is a non-atomic R-clump. Since λ is a countable limit ordinal, $\lambda=\lim\{\nu_k : k<\omega\}$ where $\{\nu_k : k<\omega\}$ is some increasing sequence of ordinals less than λ. Let $C_\lambda=\bigcap\{C_{\nu_k} : k<\omega\}$. Now $\{C_{\nu_k} : k<\omega\}$ is a countable decreasing sequence of non-atomic R-clumps, and $t\in C_\lambda$ since $t\in C_{\nu_k}$ for each $k<\omega$; so by LEMMA 10, C_λ is an R-clump. Now $t\in C_\lambda$; so by the choice of t, C_λ is a non-atomic R-clump. Finally, let $\nu<\lambda$. Since $\lambda=\lim\{\nu_k : k<\omega\}$, choose $n<\omega$ so that $\nu<\nu_n$. Hence, C_ν properly includes C_{ν_n}. But $C_\lambda\subseteq C_{\nu_n}$; so C_ν properly includes C_λ.

Therefore, for every countable ordinal ν, C_ν is defined so that $C_{\nu'}$ properly includes C_ν for each $\nu'<\nu$, so that $t\in C_\nu$, and so that C_ν is a non-atomic R-clump. But there are uncountably many countable ordinals; so there are uncountably many elements in the original R-clump C. But C is countable since $C\subseteq T$ and T is countable, so a contradiction arises. Thus, every R-clump is a union of non-atomic R-clumps.//

THEOREM:14: Every minimal R-clump is an atomic R-clump.

Proof: If A is non-atomic, choose $A_1, A_2\in\mathcal{R}$ so that A_1 and A_2 are distinct from A and $A=A_1\cup A_2$. Then A properly

includes A_1; so A is not minimal in the partial ordering $(\mathcal{R},\subseteq)$, i.e., A is not a minimal R-clump.//

REMARK: The following examples show that the condition "M_t is bounded for each $t\in T$" and the condition "M is bounded" are not related in general. It is also shown that the converse of THEOREM 14 does not hold in general.

(i) Let A be uncountable. For some $x\notin A$, let $T=A\cup\{x\}$. M is defined as follows:

$M(a,x)=M(x,a)=2$; for each $a\in A$.

$M(a,a)=1$; for each $a\in A$.

$M(x,x)=1$.

$M(a,b)=0$; otherwise.

T is an uncountable component; so by THEOREM 12, it is not the case that M_t is bounded for each $t\in T$. However, M is bounded. For each $C\in\mathcal{P}(T)$, C is an R-clump iff $x\in C$ and C is countably infinite. No atomic R-clumps exist; so the atomic R-clumps clearly do not generate the R-clumps.

(ii) Let T be countably infinite, and let $\langle t_n\rangle$ be an enumeration of T. M is defined as follows:

$M(t_n,t_{n+1})=M(t_{n+1},t_n)=n+1$; for each $n\in\mathbb{N}$.

$M(t_n,t_n)=1$; for each $n\in\mathbb{N}$.

$M(t_m,t_n)=0$; for $m,n\in\mathbb{N}$, $m\neq n+1$, and $n\neq m+1$.

M_{t_n} is bounded for each $t_n\in T$ with upper bound n+2; but M is not bounded. $\mathcal{R}=\mathcal{A}=\{\{t_n: n\geq k\}: k\in\mathbb{N}\}$. Hence, the atomic R-clumps trivially generate the R-clumps. No minimal R-clumps exist, illustrating that the converse of THEOREM 14 does not hold in general.

(iii) Let $T=\{a, b\}$. M is defined as follows: $M(a,a)=3$; $M(a,b)=M(b,a)=2$; and $M(b,b)=1$. Then $\mathcal{R}=\mathcal{A}=\{\{a\},\{a, b\}\}$. Then $\{a, b\}$ is an atomic R-clump which is not a minimal R-clump; while $\{a\}$ is a minimal and, therefore, an atomic R-clump.

DEF'N. 13: Let $A\in\mathcal{P}(T)$. Then A is a **strong clump** iff A is a connected R-clump.

Let $\mathcal{S}$ be the class of all strong clumps in T.

REMARK: Clearly, $\mathcal{S}=\mathcal{R}\cap\mathcal{C}_T$. Also by THEOREM 11, $\mathcal{A}\subseteq\mathcal{S}$; i.e., every atomic R-clump is connected and, therefore, is a strong clump.

DEF'N. 14: Let $\mathcal{B}\subseteq\mathcal{P}(T)$. Let $n\in\mathbb{N}$. Then:

(i) $\mathcal{B}^n$ is defined as $\{B\in\mathcal{B}: \|B\|=n\}$.

(ii) $n(\mathcal{B})$ is defined as $\{\cup\mathcal{B}_o: \mathcal{B}_o\subseteq\bigcup_{k=1}^{n}\mathcal{B}^k, \emptyset\notin\mathcal{B}_o\}$.

(iii) $\overline{\mathcal{B}}$ is defined to be $\mathcal{P}(T)-\mathcal{B}$.

CHAPTER TWO: THE APPLICATION OF THE THEORY OF CLUMPS

Section 2.1: An Interpretation of the Structure (T,M).

It is now possible to interpret (T,M) as a retrieval structure whenever the following conditions are imposed in addition to (M-i) through (M-iv) as given in Section 1.2.

(M-v) T is a finite set.

(M-vi) For each $t_1, t_2 \in T$, $M(t_1, t_2) < \infty$.

Intuitively, T is a finite set of terms used to characterize a set of documents, and M is a term-term association matrix which is derived in some way from such a characterization so that the values of M indicate the degree to which term pairs are associated. Condition (M-vi) is a simplifying assumption which may be interpreted as asserting that no two distinct vocabulary terms are "inseparably" associated in the characterization process. This associative retrieval structure when properly implemented in an information retrieval system gives the user of such a system the capability of receiving as output a weighted list of terms in response to an input consisting of a weighted list of terms. The user's input list of weighted terms corresponds to an A-combination (see Section 1.3) where A, the set of listed terms, reflects the user's subject interests and where the "co-efficients" of the terms reflect the user's assignment of priorities to his term

list. The user selects a set $A \subseteq T$ and chooses some $\tau \in V(A)$. The system then computes the value of the function F which maps V(A) into V(T) so that $F(\tau) = \sum \{M_\tau(\{t\})\, t: \ t \in T\}$. For each $t \in T$, $M_\tau(\{t\})$ is the "co-efficient" of t in $F(\tau)$.

In practice, the system presents as output weighted terms in decreasing order of their $F(\tau)$-weights until the user decides he wishes to see no more terms. The system's partial or complete listing of $F(\tau)$ reflects the system's assessment of priorities to the vocabulary terms which are most highly associated with τ, the user's assessment of his subject interests and priorities. The value of such a process as an aid to request reformulation by the user in terms of the system's characterization of its documents is well-known in the field of information retrieval. Moreover, with respect to the structure (T,M) interpreted as a retrieval structure, the system's role in such a process is described precisely.

Section 2.2: An Example.

The retrieval structure described in Section 2.1 has been proposed in [3] as part of an information retrieval system which is presently being implemented as the LEADER system by the Center for the Information Sciences at Lehigh University. For the purposes of this paper, it is not necessary to pursue the relationship of this retrieval structure to the rest of the system beyond the brief remarks made in Section 2.1. However, before discussing

applications of the theory of clumps to such a structure, it is illuminating to first present a concrete example of a structure (T,M) and $\mathcal{S}$, the class of all strong clumps. Figure 1 is a graph of a genus (or component) of terms also used for illustrative purposes on page 29 of [3]. In this example, $T=\{a,c,d,g,h,i,j,k,l,n\}$, and the M-value of a pair of terms is given by the number on the edge connecting them.

In Figure 1, $\mathcal{A}$, the class of all atomic R-clumps is $\{\{k\},\{a,h\},\{a,k\},\{c,g\},\{d,g\},\{g,j\},\{h,n\},\{i,l\},\{i,n\},\{l,n\},\{a,c,j\},\{a,d,j\},\{a,g,j\},\{a,h,j\},\{c,d,j\},\{h,i,l\},\{c,h,j,n\},\{d,h,j,n\},\{c,h,i,j,l\},\{d,h,i,j,l\}\}$. In Figure 1 there are around 430 R-clumps; and the class of all strong clumps is listed in tabular form in Table 1.

Figure 1: Graph of a typical genus of terms.

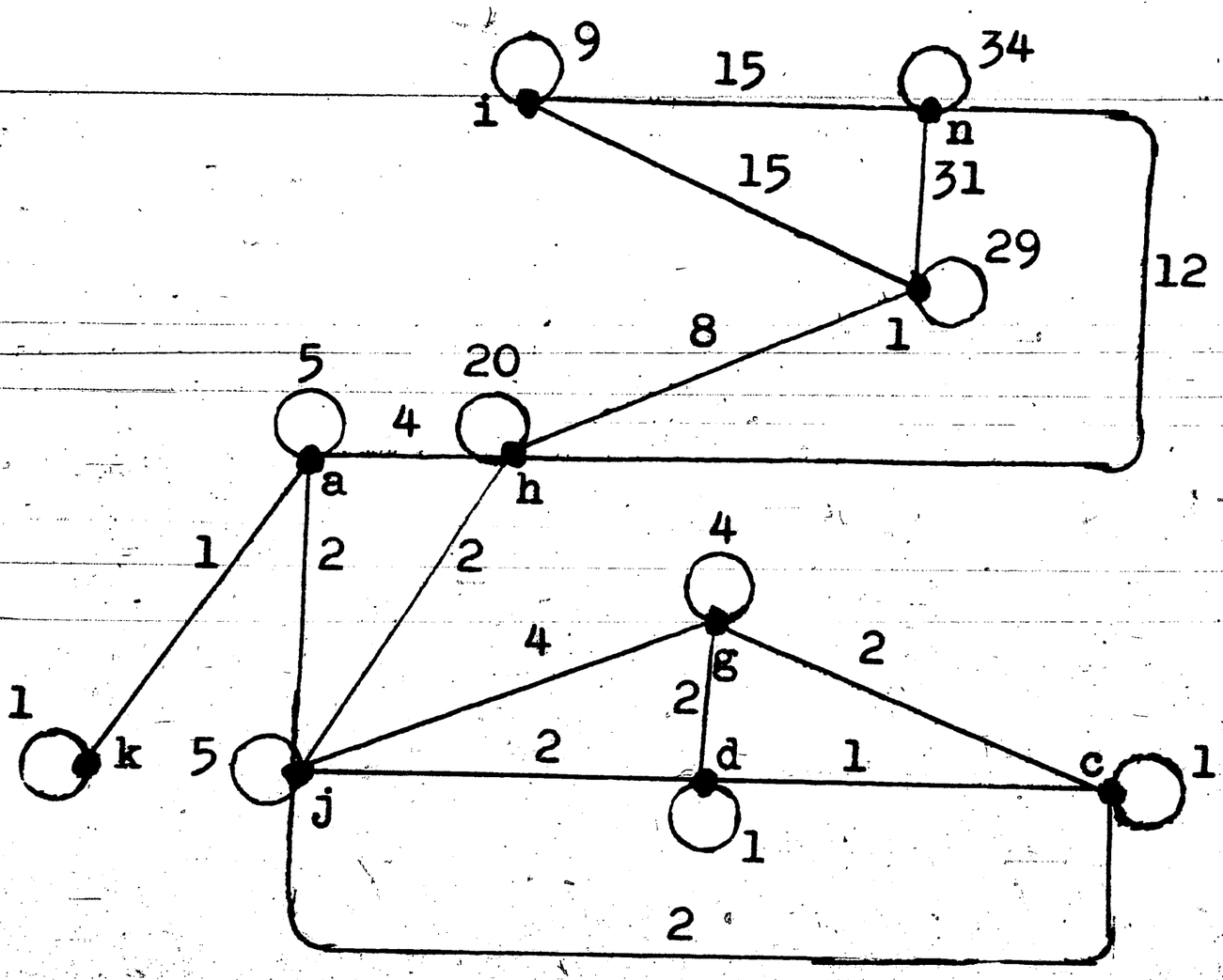


Table 1: The strong clumps of the genus in Figure 1.

{k}	{a,h,i,j,l}	{a,h,i,k,l,n}	{c,d,g,h,i,j,n}
{a,k}	{a,h,i,j,n}	{a,h,j,k,l,n}	{c,d,g,h,j,l,n}
{c,g}	{a,h,i,k,l}	{c,d,h,i,j,l}	{c,d,h,i,j,l,n}
{d,g}	{a,h,i,k,n}	{c,d,h,i,j,n}	{c,g,h,i,j,l,n}
{g,j}	{a,h,j,l,n}	{c,d,h,j,l,n}	{d,g,h,i,j,l,n}
{i,l}	{a,h,k,l,n}	{c,g,h,i,j,l}	{a,c,d,g,h,i,j,l}
{i,n}	{c,d,h,j,n}	{c,g,h,i,j,n}	{a,c,d,g,h,i,j,n}
{l,n}	{c,g,h,j,n}	{c,g,h,j,l,n}	{a,c,d,g,h,j,k,n}
{a,g,j}	{c,h,i,j,l}	{d,g,h,i,j,l}	{a,c,d,g,h,j,l,n}
{c,d,g}	{c,h,i,j,n}	{d,g,h,i,j,n}	{a,c,d,h,i,j,k,l}
{c,d,j}	{c,h,j,l,n}	{d,g,h,j,l,n}	{a,c,d,h,i,j,k,n}
{c,g,j}	{d,g,h,j,n}	{a,c,d,g,h,j,k}	{a,c,d,h,i,j,l,n}
{d,g,j}	{d,h,i,j,l}	{a,c,d,h,i,j,l}	{a,c,d,h,j,k,l,n}
{h,i,l}	{d,h,i,j,n}	{a,c,d,h,i,j,n}	{a,c,g,h,i,j,k,l}
{h,i,n}	{d,h,j,l,n}	{a,c,d,h,j,k,n}	{a,c,g,h,i,j,k,n}
{h,l,n}	{g,h,i,j,l}	{a,c,d,h,j,l,n}	{a,c,g,h,i,j,l,n}
{i,l,n}	{g,h,i,j,n}	{a,c,g,h,i,j,l}	{a,c,g,h,j,k,l,n}
{a,c,d,j}	{g,h,j,l,n}	{a,c,g,h,i,j,n}	{a,c,h,i,j,k,l,n}
{a,c,g,j}	{a,c,d,g,j,k}	{a,c,g,h,j,k,n}	{a,d,g,h,i,j,k,l}
{a,d,g,j}	{a,c,d,h,j,k}	{a,c,g,h,j,l,n}	{a,d,g,h,i,j,k,n}
{a,g,h,j}	{a,c,d,h,j,n}	{a,c,h,i,j,k,l}	{a,d,g,h,i,j,l,n}
{a,g,j,k}	{a,c,g,h,j,k}	{a,c,h,i,j,k,n}	{a,d,g,h,j,k,l,n}
{a,h,i,l}	{a,c,g,h,j,n}	{a,c,h,j,k,l,n}	{a,d,h,i,j,k,l,n}
{a,h,i,n}	{a,c,h,i,j,l}	{a,d,g,h,i,j,l}	{a,g,h,i,j,k,l,n}
{a,h,l,n}	{a,c,h,i,j,n}	{a,d,g,h,i,j,n}	{c,d,g,h,i,j,l,n}
{c,d,g,j}	{a,c,h,j,l,n}	{a,d,g,h,j,k,n}	{a,c,d,g,h,i,j,k,l}
{g,h,j,n}	{a,d,g,h,j,k}	{a,d,g,h,j,l,n}	{a,c,d,g,h,i,j,k,n}
{h,i,l,n}	{a,d,g,h,j,n}	{a,d,h,i,j,k,l}	{a,c,d,g,h,i,j,l,n}
{a,c,d,h,j}	{a,d,h,i,j,l}	{a,d,h,i,j,k,n}	{a,c,d,g,h,j,k,l,n}
{a,c,d,j,k}	{a,d,h,i,j,n}	{a,d,h,j,k,l,n}	{a,c,d,h,i,j,k,l,n}
{a,c,g,h,j}	{a,d,h,j,l,n}	{a,g,h,i,j,k,l}	{a,c,g,h,i,j,k,l,n}
{a,c,g,j,k}	{a,g,h,i,j,l}	{a,g,h,i,j,k,n}	{a,d,g,h,i,j,k,l,n}
{a,d,g,h,j}	{a,g,h,i,j,n}	{a,g,h,i,j,l,n}	{a,c,d,g,h,i,j,k,l,n}
{a,d,g,j,k}	{a,g,h,j,k,n}	{a,g,h,j,k,l,n}	
{a,g,h,j,k}	{a,g,h,j,l,n}	{a,h,i,j,k,l,n}	
{a,g,h,j,n}	{a,h,i,j,k,n}	{c,d,g,h,i,j,l}	

Section 2.3: The Applications of Strong Clumps in a Retrieval Structure.

Strong clumps may be used in the retrieval structure of an information retrieval system such as LEADER in three interesting ways: (i) aiding the request reformulation process during man-machine negotiation, (ii) affording a means for defining informal control methods to be applied to the set of vocabulary terms, and (iii) affording a means for defining formal vocabulary control methods to be applied semi-automatically. It should be remarked that in all three applications, the strong clumps should be computed when the retrieval structure is initially generated and not during man-machine negotiation. Efficiency during the request reformulation phase of man-machine negotiation is of primary importance in any real-time interactive retrieval system. If strong clumps are to be employed in negotiation, it must be decided whether or not the cost factor of strong clump storage outweighs the efficiency factor of quick system response. Perhaps the strong clump method described below and the method of partial computation of $F(\mathcal{T})$ described in Section 2.1 could both be implemented in a pilot system such as LEADER, and a comparative evaluation of the two methods could be carried out in terms of such criteria as cost, efficiency, and effectiveness. With respect to the application of strong clumps in the area of vocabulary control, strong clumps are computed for the purpose of potentially altering the retrieval structure.

Such activities are, of course, conducted by subject experts and system managers independently of man-machine negotiation.

First consider the application of strong clumps in negotiation. On the basis of the previous paragraph, it is assumed that strong clumps are stored in the system. Suppose that a user presents the system with a set A of terms during negotiation. The system then locates by some matching procedure a strong clump which includes A. The terms in such a strong clump are more highly associated with each other than with any set of terms not in the clump. If not all the terms of A are in the same genus or component, no R-clump including A is connected and, hence, no strong clump includes A. For this reason it is assumed that the user has already narrowed his term set A so that A is included in a single genus.

Now assume that the user specifies that he wishes to see at least k but no more than n new terms where $k \leq n$. The retrieval system scans $\mathcal{S}^{\|A\|+k}, \mathcal{S}^{\|A\|+k+1}, \ldots, \mathcal{S}^{\|A\|+n}$ in that order and selects the first strong clump (or the first few strong clumps) which includes A. The new terms in the selected clump or clumps are presented to the user for reformulation purposes. If no such clump is found, the user's request is still too broad to be handled by the system. In this case, the request is handled in the same manner as a request whose terms are not in a single genus. Thus, clump-matching may be used as a more sensitive

criterion than genus-sharing for determining that a term set is too broad for document retrieval purposes and that further refinement and reformulation is desirable.

Consider now the applications of strong clumps in vocabulary control. Such applications fall into the categories of informal and formal procedures. As an example of informal control, subject experts and system managers examine each strong clump of sufficiently small size (or perhaps only each atomic R-clump), attempt to assign meaningful phrases to represent those strong clumps which seem to have intuitive as well as formal coherence, and remove those strong clumps which have absolutely no meaningful intuitive coherence. Such meaningful phrases as are assigned would then be used in the negotiation process in place of the clumps they represent.

In the third and final category of formal vocabulary control, one attempts to formally relate the additional structure of strong clumps to the original connectivity structure (T,M) as proposed in [3]. Strong clumps enable one to estimate the "articulation values" of terms. Intuitively, a term of a genus has a high articulation value whenever its removal would break or, in less extreme situations, weaken the connectedness of a genus; i.e., whenever the term is or "almost is" an articulation point of the genus.

The so-called unique probability vector of a genus has been proposed in [3] as an estimate of a term's

"centrality" or importance in the genus. In the example found in Section 2.2, the unique probability vector values for a, c, d, g, h, i, j, k, l, and n are, respectively, $\frac{12}{315}$, $\frac{6}{315}$, $\frac{6}{315}$, $\frac{12}{315}$, $\frac{46}{315}$, $\frac{39}{315}$, $\frac{17}{315}$, $\frac{2}{315}$, $\frac{83}{315}$, and $\frac{92}{315}$.

The concepts "articulation point" and "central point" are different, as is made clear by the following discussion from page 37 of [3]:

> In addition, there are several characteristic-vertices [terms] in the graph of the genus [in Figure 1 of this paper] whose removal will leave the graph unconnected. The characteristics j, a, and h have this property. Such vertices are called <u>articulation points</u> of the first level genus. Even though these characteristics are not all heavily weighted in the unique probability vector..., they are nevertheless important in assuring first level connections between subgraphs.

Thus, it has been recognized that unique probability vector values are not really appropriate for estimating articulation values.

For each $t \in T$ and for each $n \in \mathbb{N}$, let the <u>n^{th} articulation value</u> of t, written as $a_n(t)$, be the number of strong clumps of cardinality n which contain term t; i.e., $a_n(t) = \| \{C \in \mathcal{S}^n : t \in C\} \|$. The articulation values for the terms in the genus in Figure 1 may be directly computed by observing Table 1; and these values are presented in Table 2 as a function of $t \in T$ and $n \in \mathbb{N}$ (where $n \leq \|T\|$). For each $t \in T$, let the <u>articulation value</u> of t, written as a(t), be $\sum_{n=1}^{\|T\|} a_n(t)$. These values may be found by summing the rows of Table 2 for each $t \in T$. Observing Table 2, it is clear that the

Table 2: Values of $a_n(t)$ as a function of $t \in T$ and $n \in \mathbb{N}$ $(n \leq \|T\|)$ for the genus in Figure 1.

t \ n	1	2	3	4	5	6	7	8	9	10
a	0	1	1	8	14	20	24	19	7	1
c	0	1	3	3	9	14	17	14	6	1
d	0	1	3	3	9	14	17	14	6	1
g	0	3	4	6	11	15	18	14	6	1
h	0	0	3	6	23	28	30	20	7	1
i	0	2	3	3	10	14	19	15	6	1
j	0	1	4	7	23	28	30	20	7	1
k	1	1	0	1	7	8	14	13	6	1
l	0	2	3	2	10	14	19	15	6	1
n	0	2	3	4	14	19	22	16	6	1

articulation values for a, c, d, g, h, i, j, k, l, and n are, respectively, 95, 68, 68, 78, 118, 73, 121, 52, 73, and 87. The fact that j, a, and h have the highest articulation values agrees with the fact that j, a, and h are the articulation points of the genus in Figure 1.

It is conjectured that the articulation values as defined above are appropriate for determining the articulation points of a genus. How then are the articulation points thus found of any use to a system manager interested in vocabulary control? If articulation points are terms which are, in fact, "junk" terms adjudged to be useless in actual retrieval and negotiation operations, such terms are removed from the genus, and the genus may then "fall apart"

into several genera. At the very least, associations among "useful" terms will stand out more clearly when associations with spurious terms are eliminated. In this way, the formal relationships between strong clumps and the original retrieval structure may be exploited by developing semi-automatic vocabulary control procedures.

CHAPTER THREE: THE COMPUTATUION OF STRONG CLUMPS

Section 3.1: Basic Operations in Computing Strong Clumps.

Now that the theory of clumps and its applications have been presented, it remains only to provide an algorithm for computing the class of all strong clumps in a genus of terms. This algorithm is described in such a way that a computer program may easily be written to perform the computation. The computation procedure has two main components which are called the analytic component and the generative component. Let $n \in \mathbb{N}$, $n \leq \|T\| - 1$, be fixed for the remainder of this chapter. First the n^{th} stage of the analytic component is described and then the n^{th} stage of the generative component is described, since this is the order in which these stages occur during computation. The notation from Chapter 1 is followed, and results from several theorems are quoted as procedural justification.

There are five basic types of operations involved in these components: STORE, RENAME, COMPUTE, TEST, and GEN. The first four of these are rather simple to describe, while the last is somewhat complex. (the "operation" BEGIN is not really an operation, but merely a notational device in a diagram to indicate which component precedes or follows the operations in that particular diagram.) The five basic operations are explained as follows:

1. STORE: Several storage areas are created for the purpose of keeping a list of sets in a particular class of sets. The STORE operation directs that a set be stored in the storage area designated by the symbol for the corresponding class. So, for example, "STORE: $X \in a^n$" means that the set X is to be placed in the storage area in which the elements of the class a^n are stored. If the set X is to be used later, the instruction might instead be "STORE: copy of $X \in a^n$", which means that a reproduction of X is created and stored while the original X remains in the work space. Storage areas may be empty while the classes used to designate them are not. For example, "$a^n = \emptyset$" as part of an operation means not that the class a^n is equal to the empty set, but that the storage area designated by a^n is empty of elements at that particular time.

2. RENAME: The RENAME operation is a device which changes labels on a storage area. For example, the instruction "RENAME: (n-1)(a) as n(a)" removes the label from the storage area it designates and attaches another label, namely, n(a).

3. COMPUTE: The COMPUTE operation allows one to compute unions over classes of sets. Storage areas are created by this operation. For example, the instruction "COMPUTE: $\mathcal{F}(n) = \{\cup \mathcal{B} : \mathcal{B} \subseteq a^n\}$" directs that a storage area designated by $\mathcal{F}(n)$ be created by means of removing copies of several sets from a^n, putting them in the work space, and storing in $\mathcal{F}(n)$ all the distinct sets which

result from taking all possible unions of these input sets from a^n.

4. TEST: The TEST operation (always followed by a "?" and numbered for convenience) is really a blanket operation which covers several different kinds of operations. This operation always has the effect of testing whether or not some entity has a particular property, either by looking at one of the storage areas designated by a class symbol or by performing operations involving the matrix M. The result of a TEST operation is either a "YES" or a "NO" depending on whether or not the entity has the property. For example, "TEST 1: $a^n=\emptyset$?" is an instruction to examine the storage area to see whether or not it contains any sets at that time; and given a set X in the work space, "TEST 2: X $(\mathcal{S}-a)^n$?" is an instruction to examine the storage area $(\mathcal{S}-a)^n$ to see whether or not it contains a copy of X at that time. As an example of the kind of TEST involving M, "TEST 4: X connected?" is an instruction to see if there is a connected path between all terms in X. This paper, however, is not concerned with the programming involved in the internal structure of such a TEST, since such programming depends upon the organization of the file in which a record of the structure (T,M) is kept.

5. GEN: The GEN operation is a bit more complicated than the previous operations. GEN is an operation which involves two variables: the first variable, called the determiner, is a set which is fixed at the beginning of a

multiple cyclic use of the GEN instruction; and the second variable, called the variant, ranges over the elements of the determiner and changes after each single cycle through the GEN instruction. If the determiner is a set of terms, then the variant ranges over the terms of that set; and if the determiner is a class of sets, then the variant ranges over the sets in that class. As long as new variants are available in the determiner set, a "YES" results from the GEN instruction; but as soon as they are exhausted, the GEN instruction results in a "NO". There are three situations in which the GEN operation occurs:

(i) If the determiner is a set X in the work space, "GEN: X;x" generates elements of X. Moreover, the order in which the variant x takes on new values of X is the "alphabetical order" which is imposed on T throughout all computations. At the beginning of each single GEN cycle, either "GEN: X;x" results in a "YES" if a new element of X is available, whereupon this element is put in the work space, or else "GEN: X;x" results in a "NO" if all the elements of X have already been used and no new elements are available, whereupon the multiple GEN cycle terminates.

(ii) If the determiner is the class $\mathcal{P}(T)^n$, then "GEN: $\mathcal{P}(T)$;X" generates sets in this class. Moreover, the order in which the variant X takes on new values of $\mathcal{P}(T)^n$ is the "alphabetical order" on T as naturally induced on all sets of cardinality n. (for example, the sets in Table 2 are listed in this induced order.) As

before, when a "YES" results, the new set in $\mathcal{P}(T)^n$ is put in the work space; and when a "NO" results, $\mathcal{P}(T)^n$ has been exhausted and the multiple GEN cycle terminates.

(iii) If the determiner is a storage area, the sets in the storage area (i.e., the variants) are put in the work space one by one in the order in which they are stored as long as the GEN instruction results in a "YES". As soon as a "NO" results, the storage area has been completely emptied, and the GEN cycle terminates.

Section 3.2: The n^{th} Stage of the Analytic Component.

Recall the assumption in the previous section that T is a genus and that $n \in \mathbb{N}$, $n \leq \|T\|-1$, is fixed. The following is a discussion of the n^{th} analytic component; i.e., the n^{th} stage of the analytic component. The n^{th} analytic stage follows the $(n-1)^{th}$ generative stage if $n \geq 2$ and precedes the n^{th} generative stage. Before the n^{th} analytic stage begins, the storage areas $(\mathcal{R}-\mathcal{S})^n$ and $(\mathcal{S}-\mathcal{A})^n$ have been completely computed (i.e., the storage area designated by $(\mathcal{R}-\mathcal{S})^n$ contains all and only those sets in the class $(\mathcal{R}-\mathcal{S})^n$; likewise for $(\mathcal{S}-\mathcal{A})^n$) and the storage area $\overline{\mathcal{R}}^n$ has been partially computed (i.e., the storage area designated by $\overline{\mathcal{R}}^n$ contains only but not necessarily all the sets in the class $\overline{\mathcal{R}}^n$). These complete and partial computations were performed by the $(n-1)^{th}$ generative stage. The purpose of the n^{th} analytic stage is to complete the computation of $\overline{\mathcal{R}}^n$ and to completely compute

before, when a "YES" results, the new set in $\mathcal{P}(T)^n$ is put in the work space; and when a "NO" results, $\mathcal{P}(T)^n$ has been exhausted and the multiple GEN cycle terminates.

(iii) If the determiner is a storage area, the sets in the storage area (i.e., the variants) are put in the work space one by one in the order in which they are stored as long as the GEN instruction results in a "YES". As soon as a "NO" results, the storage area has been completely emptied, and the GEN cycle terminates.

Section 3.2: The n^{th} Stage of the Analytic Component.

Recall the assumption in the previous section that T is a genus and that $n \in \mathbb{N}$, $n \leq \|T\| - 1$, is fixed. The following is a discussion of the n^{th} analytic component; i.e., the n^{th} stage of the analytic component. The n^{th} analytic stage follows the $(n-1)^{th}$ generative stage if $n \geq 2$ and precedes the n^{th} generative stage. Before the n^{th} analytic stage begins, the storage areas $(\mathcal{R} - \mathcal{S})^n$ and $(\mathcal{S} - \mathcal{A})^n$ have been completely computed (i.e., the storage area designated by $(\mathcal{R} - \mathcal{S})^n$ contains all and only those sets in the class $(\mathcal{R} - \mathcal{S})^n$; likewise for $(\mathcal{S} - \mathcal{A})^n$) and the storage area $\overline{\mathcal{R}}^n$ has been partially computed (i.e., the storage area designated by $\overline{\mathcal{R}}^n$ contains only but not necessarily all the sets in the class $\overline{\mathcal{R}}^n$). These complete and partial computations were performed by the $(n-1)^{th}$ generative stage. The purpose of the n^{th} analytic stage is to complete the computation of $\overline{\mathcal{R}}^n$ and to completely compute

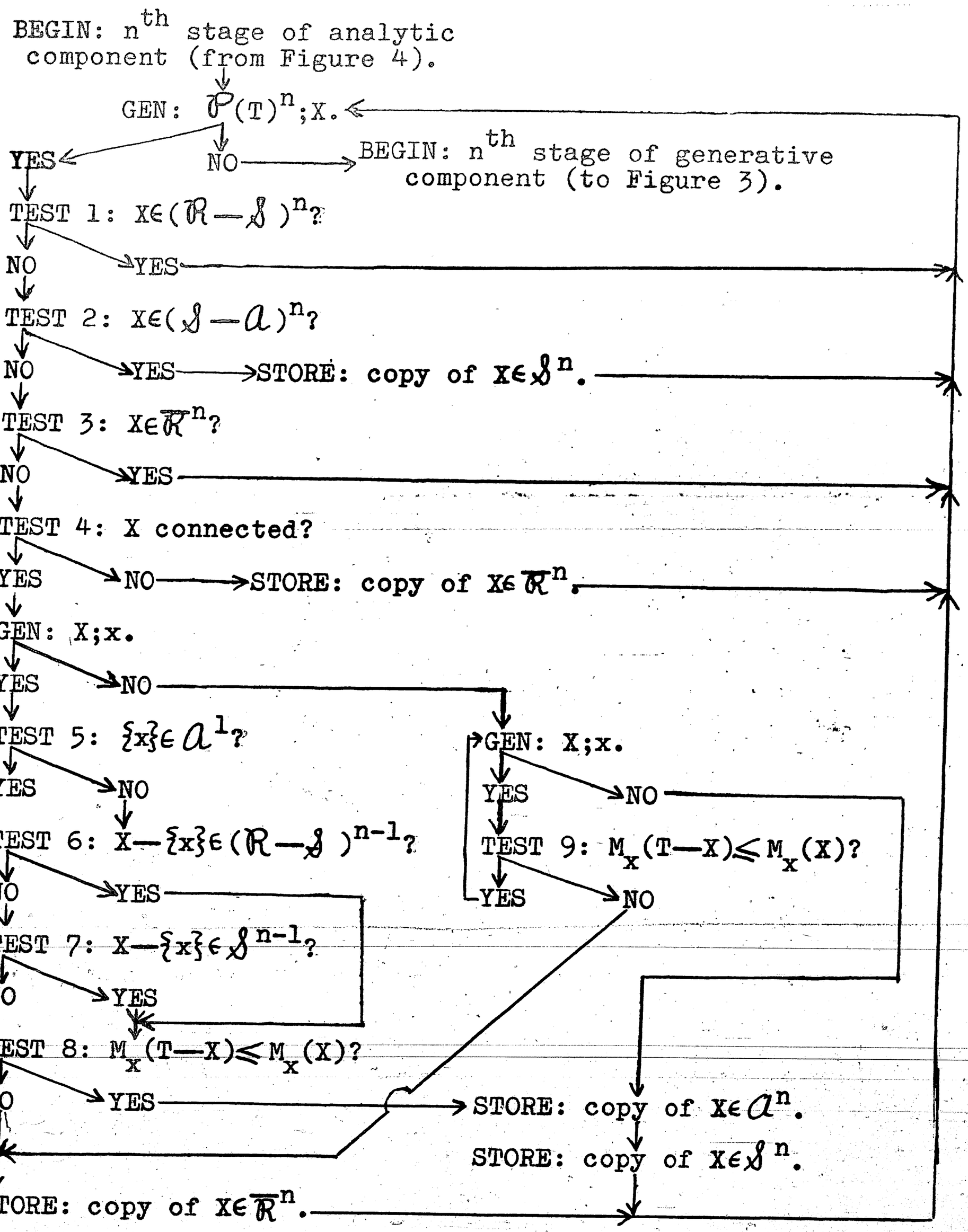
the storage areas $\mathcal{A}^n$ and $\mathcal{S}^n$. A diagram of the order of operations in the n^{th} analytic stage appears in Figure 2. An explanation of this diagram constitutes the remainder of this section.

It should be remarked that the first analytic stage (n=1) is the beginning of the overall clumping procedure; and all the storage areas are initially empty. However, $(\mathcal{R}-\mathcal{S})^1$ and $(\mathcal{S}-\mathcal{A})^1$ are still completely computed at this point since these classes are empty. $(\mathcal{R}-\mathcal{S})^1$ is empty because every singleton is connected (by LEMMA 5) and, hence, is a strong clump if it is an R-clump. Also, $(\mathcal{S}-\mathcal{A})^1$ is empty because every singleton which is a strong clump and, hence, an R-clump must be atomic.

As soon as the n^{th} stage of the analytic component begins, elements of $\mathcal{P}(T)^n$ are generated in the appropriate order until $\mathcal{P}(T)^n$ is exhausted, at which time the n^{th} analytic stage terminates and the n^{th} generative stage begins. A typical set $X \in \mathcal{P}(T)^n$ is now traced through the GEN cycle comprising the analytic component's n^{th} stage.

TEST 1: $X (\mathcal{R}-\mathcal{S})^n$? If a "YES" results, it is clear that X is disconnected and is neither an atomic R-clump (by THEOREM 11), a strong clump (by definition), nor a non-R-clump. In this case, $X \notin \mathcal{A}^n$, $X \notin \mathcal{S}^n$, and $X \notin \mathcal{R}^n$; so the process returns to the "GEN: $\mathcal{P}(T)^n$;X" instruction, X is discarded, and a new set of cardinality n (if it exists) is generated. If a "NO" results from this TEST, operation, further tests are performed.

Figure 2: Diagram of operations in the n^{th} stage of the analytic component.



TEST 2: $X\in(\mathcal{S}-\mathcal{A})^n$? If a "YES" results, then $X\notin\mathcal{A}^n$, $X\in\mathcal{S}^n$, and $X\notin\mathcal{R}^n$; so a copy of X is stored in storage area $\mathcal{S}^n$, and the process returns to "GEN: $\mathcal{P}(T)^n$;X". If a "NO" results, further tests are performed. In all such further tests, it is known that either $X\in\mathcal{A}^n$ or $X\in\mathcal{R}^n$ since the storage areas $(\mathcal{R}-\mathcal{S})^n$ and $(\mathcal{S}-\mathcal{A})^n$ are completely computed before the n^{th} stage of the analytic component begins.

TEST 3: $X\in\mathcal{R}^n$? If a "YES" results, X is already in $\mathcal{R}^n$, $X\notin\mathcal{A}^n$, and $X\notin\mathcal{S}^n$; so the process returns to the "GEN: $\mathcal{P}(T)^n$;X" instruction. If a "NO" results, X may or may not be a member of the class $\mathcal{R}^n$ since the corresponding storage area is only partially computed at this time.

TEST 4: X connected? If a "NO" results, $X\notin\mathcal{A}^n$ (by THEOREM 11); so by the remark at the end of TEST 2, $X\in\mathcal{R}^n$. A copy of X is stored in $\mathcal{R}^n$, and the process returns to "GEN: $\mathcal{P}(T)^n$;X". If a "YES" results, further tests are performed.

TESTS 5, 6, 7, and 8: This group of four tests taken together constitutes the "short check" for clumps and uses THEOREM 10. Elements x of X are generated (in alphabetical order) until X is exhausted or until the hypothesis of THEOREM 10 is satisfied with $A=X-\{x\}$ and $t=x$ for some $x\in X$ (i.e., for some $x\in X$, TEST 5 results in a "NO" and either TEST 6 or TEST 7 results in a "YES"). This cycle of tests is bypassed in the first analytic stage. If TEST 5 results in a "NO", $\{x\}\notin\mathcal{A}^1$ since the storage area is completely computed at the beginning of the n^{th} analytic stage, $n\geq 2$;

so $\{x\} \notin \mathcal{R}$ since all singleton R-clumps must be atomic. If either TEST 6 or TEST 7 results in a "YES", $X - \{x\} \in \mathcal{R}$. By THEOREM 10 and THEOREM 7, if $M_X(T-X) \leqslant M_X(X)$, then X is an R-clump and, by the remark at the end of TEST 2, X is also in $\mathcal{A}^n$. So if a "YES" results from TEST 8, copies of X are stored in $\mathcal{A}^n$ and $\mathcal{S}^n$. On the other hand, if a "NO" results from TEST 8, X does not satisfy the definition of an R-clump; so a copy of X is stored in $\mathcal{R}^n$. Note that TEST 8, if performed, is performed exactly once for just one element x of X; and this is sufficient to determine whether or not X is an R-clump. After TEST 8 is performed and the appropriate storage operations are carried out, the process returns to "GEN: $\mathcal{P}(T)^n$;X". If TEST 5 always results in a "YES" or if both TEST 6 and TEST 7 always result in a "NO" whenever TEST 5 results in a "NO", no $x \in X$ ever satisfies the property that $X - \{x\} \in \mathcal{R}$ and $\{x\} \notin \mathcal{R}$. In this case, THEOREM 10 and the "short check" for clumps fails to be of use; so after the "GEN: X;x" cycle terminates, further tests are performed involving the "long check" for clumps.

TEST 9: This "long check" for clumps is so called because $M_X(T-X)$ and $M_X(X)$ are computed and compared for each $x \in X$ until either a "NO" results for some $x \in X$, in which case X is not an R-clump, or until a "YES" results for each $x \in X$, in which case X is an R-clump. If X is not an R-clump, a copy of X is stored in $\mathcal{R}^n$. If X is an R-clump, X is atomic by the remark at the end of TEST 2; so copies of X are

stored in $\mathcal{a}^n$ and $\mathcal{s}^n$. In either case, the process returns to "GEN: $\mathcal{P}(T)^n$;X" after the decision is made and the set is stored appropriately.

This entire procedure continues until all term sets of cardinality n have been inspected and $\mathcal{P}(T)^n$ is exhausted. "GEN: $\mathcal{P}(T)^n$;X" then results in a "NO" and the n^{th} stage of the analytic component terminates. From the previous discussion, it is clear that the storage areas $\mathcal{R}^n$, $\mathcal{a}^n$, and $\mathcal{s}^n$ are completely computed at this point; so the n^{th} stage of the generative component begins.

Section 3.3: Phase 1 of the n^{th} Stage of the Generative Component.

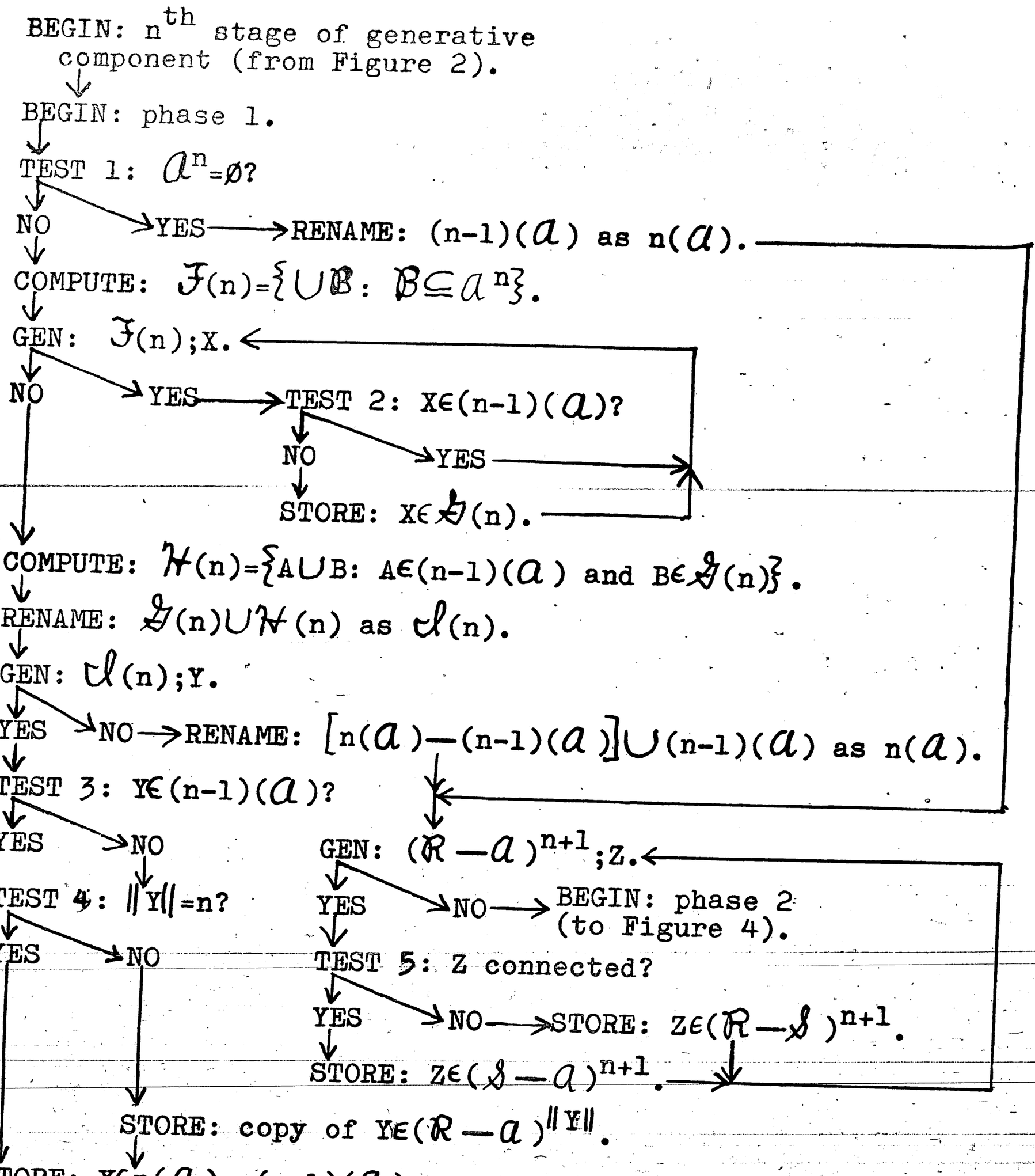
The following is a discussion of the n^{th} generative component; i.e., the n^{th} stage of the generative component. The n^{th} generative stage follows the n^{th} analytic stage and precedes the $(n+1)^{th}$ analytic stage if $n \leq \|T\|-2$. The storage areas $\mathcal{R}^n$, $\mathcal{a}^n$, and $\mathcal{s}^n$ have been completely computed by the n^{th} analytic stage; and the storage area $(n-1)(\mathcal{a})$ has been completely computed by the $(n-1)^{th}$ generative stage. (By DEF'N. 14, if n=1, then $(n-1)(\mathcal{a})$ $=0(\mathcal{a})=\{\cup \mathcal{B}_0 : \mathcal{B}_0 \subseteq \bigcup_{k=1}^{0} \mathcal{a}^k$ and $\emptyset \neq \mathcal{B}_0\}=\emptyset$; so $(n-1)(\mathcal{a})$ is still completely computed at this point.) The storage area $\mathcal{s}^n$ is not used in the n^{th} generative component, which consists of two phases. Storage areas $(n-1)(\mathcal{a})$ and $\mathcal{a}^n$ are used in phase 1 for the purpose of completely computing $n(\mathcal{a})$ and $(\mathcal{R}-\mathcal{a})^{n+1}$ and of partially computing

$(\mathcal{R}-\mathcal{A})^k$, where $n+2 \leqslant k \leqslant \|T\|$. Later in phase 1, the completely computed storage area $(\mathcal{R}-\mathcal{A})^{n+1}$ is used for the purpose of completely computing $(\mathcal{R}-\mathcal{S})^{n+1}$ and $(\mathcal{S}-\mathcal{A})^{n+1}$. Storage area $\mathcal{R}^n$ is used in phase 2 for the purpose of partially computing $\mathcal{R}^{n+1}$. A diagram of the order of operations in phase 1 of the n^{th} generative stage appears in Figure 3. An explanation of this diagram constitutes the remainder of this section, while phase 2 of the n^{th} generative component is discussed in the next section.

As soon as the n^{th} stage of the generative component begins, phase 1 begins; and TEST 1 is performed. The storage area $\mathcal{A}^n$ is completely computed at this time; so if it is empty at this time, no atomic R-clump of cardinality n exists. Hence, if a "YES" results from TEST 1, $(n-1)(\mathcal{A})=n(\mathcal{A})$; so the storage area designated by $(n-1)(\mathcal{A})$ is appropriately renamed $n(\mathcal{A})$, and the process moves to the instruction "GEN: $(\mathcal{R}-\mathcal{A})^{n+1}$;Z". On the other hand, if a "NO" results from TEST 1, the storage area $\mathcal{F}(n)$ is created and $\mathcal{F}(n)=\{\cup\mathcal{B}: \mathcal{B}\subseteq\mathcal{A}^n\}$ is computed.

As soon as $\mathcal{F}(n)$ is completely computed, the sets in $\mathcal{F}(n)$ are put in the work space one by one until $\mathcal{F}(n)$ is emptied. Every such input set undergoes TEST 2, which has the effect of discarding all members of $\mathcal{F}(n)$ which already are stored in $(n-1)(\mathcal{A})$ and of storing the rest in the storage area $\mathcal{G}(n)$. After each set in $\mathcal{F}(n)$ has been inspected and $\mathcal{F}(n)$ is exhausted of sets, the storage area $\mathcal{G}(n)$ is completely computed and consists of all R-clumps

Figure 3: Diagram of operations in phase 1 of the n^{th} stage of the generative component.



which are generable by unions from $\mathcal{A}^n$ and which are not generable by unions from $\bigcup_{k=1}^{n-1}\mathcal{A}^k$ (i.e., which do not belong to $(n-1)(\mathcal{A})$).

After $\mathcal{G}(n)$ is completely computed, a new storage area $\mathcal{H}(n)$ is created and $\mathcal{H}(n)=\{A\cup B: A\in(n-1)(\mathcal{A})$ and $B\in\mathcal{G}(n)\}$ is computed. As soon as $\mathcal{H}(n)$ is completely computed, the storage areas $\mathcal{G}(n)$ and $\mathcal{H}(n)$ are merged into a storage area which is renamed $\mathcal{CL}(n)$. Hence, $\mathcal{CL}(n)=\mathcal{G}(n)\cup\mathcal{H}(n)$. $\mathcal{CL}(n)$ consists of those R-clumps which are generable by unions from $\bigcup_{k=1}^{n}\mathcal{A}^k$ except possibly for some which are generable exclusively from $\bigcup_{k=1}^{n-1}\mathcal{A}^k$. To see this, it suffices to show that $n(\mathcal{A})-(n-1)(\mathcal{A})\subseteq\mathcal{CL}(n)\subseteq n(\mathcal{A})$. But clearly $\mathcal{CL}(n)\subseteq n(\mathcal{A})$; so to show that $n(\mathcal{A})-(n-1)(\mathcal{A})\subseteq\mathcal{CL}(n)$, let $C\in n(\mathcal{A})-(n-1)(\mathcal{A})$. Since $C\in n(\mathcal{A})$, choose $\mathcal{B}\subseteq\bigcup_{k=1}^{n}\mathcal{A}^k$ so that $C=\cup\mathcal{B}$. Let $\mathcal{B}_1=\mathcal{B}\cap(n-1)(\mathcal{A})$ and let $\mathcal{B}_2=\mathcal{B}\cap\mathcal{G}(n)$. Then the following assertions hold:

(i) $\mathcal{B}=\mathcal{B}_1\cup\mathcal{B}_2$: Clearly $\mathcal{B}_1\cup\mathcal{B}_2\subseteq\mathcal{B}$; so to show that $\mathcal{B}\subseteq\mathcal{B}_1\cup\mathcal{B}_2$, let $B\in\mathcal{B}$. Then choose $k\leq n$ so that $B\in\mathcal{A}^k$. If $k<n$, then $B\in\mathcal{A}^k\subseteq(n-1)(\mathcal{A})$; so $B\in\mathcal{B}_1$. If $k=n$, then $B\in\mathcal{A}^n\subseteq\mathcal{G}(n)$; so $B\in\mathcal{B}_2$. Hence, $B\in\mathcal{B}_1\cup\mathcal{B}_2$.

(ii) $C=(\cup\mathcal{B}_1)\cup(\cup\mathcal{B}_2)$: This follows because:

$C=\cup\mathcal{B}$ (by the choice of $\mathcal{B}$)

$=\cup(\mathcal{B}_1\cup\mathcal{B}_2)$ (by (i))

$=(\cup\mathcal{B}_1)\cup(\cup\mathcal{B}_2)$.

(iii) $\cup\mathcal{B}_2 \in \mathcal{G}(n)$: Now $\mathcal{B}_2 \subseteq \mathcal{G}(n) \subseteq \mathcal{F}(n)$; so by the definition of $\mathcal{F}(n)$, it follows that $\cup\mathcal{B}_2 \in \mathcal{F}(n)$. It is clear from TEST 2, which is in the "GEN: $\mathcal{F}(n)$;X" cycle, that $\cup\mathcal{B}_2 \in (n-1)(\mathcal{A}) \cup \mathcal{G}(n)$. So assume that $\cup\mathcal{B}_2 \in (n-1)(\mathcal{A})$. Then $(\cup\mathcal{B}_1) \cup (\cup\mathcal{B}_2) \in (n-1)(\mathcal{A})$, since $(n-1)(\mathcal{A})$ is closed under union. Hence, by (ii) it follows that $C \in (n-1)(\mathcal{A})$, which contradicts the choice of $C \in n(\mathcal{A}) - (n-1)(\mathcal{A})$. Thus, $\cup\mathcal{B}_2 \notin (n-1)(\mathcal{A})$. However, $\cup\mathcal{B}_2 \in (n-1)(\mathcal{A}) \cup \mathcal{G}(n)$; so $\cup\mathcal{B}_2 \in \mathcal{G}(n)$.

(iv) $C \in \mathcal{C}\ell(n)$: There are two possible cases.

If $\cup\mathcal{B}_1 = \emptyset$, then $C = \cup\mathcal{B}_2$ by (ii). Then by (iii), $C = \cup\mathcal{B}_2 \in \mathcal{G}(n) \subseteq \mathcal{G}(n) \cup \mathcal{H}(n) \subseteq \mathcal{C}\ell(n)$.

If $\cup\mathcal{B}_1 \neq \emptyset$, then $\mathcal{B}_1 \neq \emptyset$; so $\cup\mathcal{B}_1 \in (n-1)(\mathcal{A})$. Thus, by (ii) and (iii), it is clear that C is a union of a set in $(n-1)(\mathcal{A})$ and a set in $\mathcal{G}(n)$; so $C \in \mathcal{H}(n) \subseteq \mathcal{G}(n) \cup \mathcal{H}(n) \subseteq \mathcal{C}\ell(n)$.

In either case, $C \in \mathcal{C}\ell(n)$.

Summarizing assertions (i) through (iv), any set C which belongs to $n(\mathcal{A}) - (n-1)(\mathcal{A})$ also belongs to $\mathcal{C}\ell(n)$; so $n(\mathcal{A}) - (n-1)(\mathcal{A}) \subseteq \mathcal{C}\ell(n)$.

Now since it has been assumed that TEST 1 results in a "NO", the class $\mathcal{A}^n \neq \emptyset$. But $\mathcal{A}^n \subseteq \mathcal{G}(n) \subseteq \mathcal{C}\ell(n)$; so the storage area $\mathcal{C}\ell(n)$ is initially non-empty. Sets from the storage area $\mathcal{C}\ell(n)$ are put in the work space one by one until eventually $\mathcal{C}\ell(n)$ is exhausted. Suppose that Y is

a typical input set from $\mathcal{C}(n)$. If TEST 3 results in a "YES", Y is discarded since Y is already in the completely computed storage area $(n-1)(\mathcal{A})$; and if TEST 3 results in a "NO", TEST 4 is performed on Y. If TEST 4 results in a "YES", then $\|Y\|=n$ and Y is hence an atomic R-clump since it is not expressible as a union of smaller R-clumps (i.e., since it is not a member of $(n-1)(\mathcal{A})$). If TEST 4 results in a "NO", then $\|Y\|\geq n+1$ and Y is hence a non-atomic R-clump since $Y\in\mathcal{C}(n)\subseteq n(\mathcal{A})$. Thus, if a "NO" results, a copy of Y is stored in $(\mathcal{R}-\mathcal{A})^{\|Y\|}$. In either case, Y is stored in $n(\mathcal{A})-(n-1)(\mathcal{A})$ since $Y\in\mathcal{C}(n)\subseteq n(\mathcal{A})$ and since $Y\notin(n-1)(\mathcal{A})$ by TEST 3.

Sets from the storage area $\mathcal{C}(n)$ are put in the work space one by one until $\mathcal{C}(n)$ is exhausted. But since $n(\mathcal{A})-(n-1)(\mathcal{A})\subseteq\mathcal{C}(n)$, the storage area $n(\mathcal{A})-(n-1)(\mathcal{A})$ is completely computed by the time $\mathcal{C}(n)$ is exhausted. Since $(n-1)(\mathcal{A})$ is completely computed, the merged storage area $[n(\mathcal{A})-(n-1)(\mathcal{A})]\cup(n-1)(\mathcal{A})$ is also completely computed; so $[n(\mathcal{A})-(n-1)(\mathcal{A})]\cup(n-1)(\mathcal{A})$ is appropriately renamed $n(\mathcal{A})$; and the process moves to the instruction "GEN: $(\mathcal{R}-\mathcal{A})^{n+1}$;X".

It is clear that as this GEN cycle begins, $n(\mathcal{A})$ is completely computed. It is also the case at this time that $(\mathcal{R}-\mathcal{A})^{n+1}$ is completely computed while $(\mathcal{R}-\mathcal{A})^{k}$ is partially computed for $n+2\leq k\leq\|T\|$. By the time this GEN cycle begins, all the sets in $n(\mathcal{A})$ have been inspected to see whether they are atomic or non-atomic R-clumps.

Therefore, to show that $(\mathcal{R}-\mathcal{A})^{n+1}$ is completely computed, it suffices to show that $(\mathcal{R}-\mathcal{A})^{n+1}\subseteq n(\mathcal{A})$.

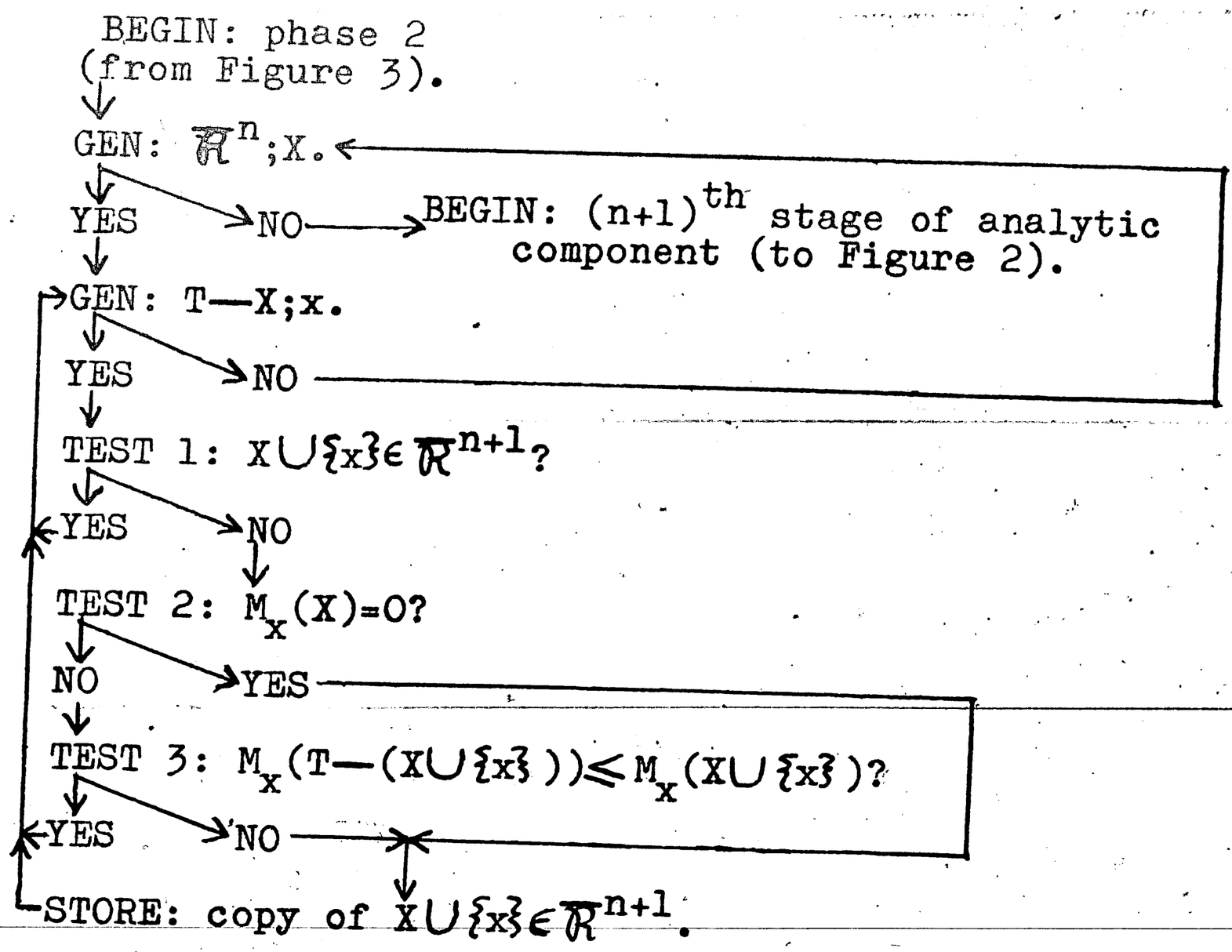
Suppose that $C\in(\mathcal{R}-\mathcal{A})^{n+1}$. Then $\|C\|=n+1$ and C is a non-atomic R-clump. Since C is a non-atomic R-clump, choose $A_1, A_2\in\mathcal{R}$ so that A_1 and A_2 are distinct from C and $C=A_1\cup A_2$. Then C properly includes A_1; so $\|A_1\|<\|C\|=n+1$. But A_1 is an R-clump; so by THEOREM 13, A is the union of some atomic R-clumps of cardinality less than or equal to n. Hence, $A_1\in n(\mathcal{A})$. Similarly, $A_2\in n(\mathcal{A})$; so it follows that $C=A_1\cup A_2\in n(\mathcal{A})$.

In the final portion of phase 1, TEST 5 divides the sets in the completely computed storage area $(\mathcal{R}-\mathcal{A})^{n+1}$ into two exhaustive and mutually exclusive classes; namely, $(\mathcal{R}-\mathcal{S})^{n+1}$ and $(\mathcal{S}-\mathcal{A})^{n+1}$. As soon as all the sets in $(\mathcal{R}-\mathcal{A})^{n+1}$ have been tested, it is clear that $(\mathcal{R}-\mathcal{S})^{n+1}$ and $(\mathcal{S}-\mathcal{A})^{n+1}$ are completely computed; and the instruction "GEN: $(\mathcal{R}-\mathcal{A})^{n+1}$;Z" results in a "NO". At this time, phase 1 of the n^{th} stage of the generative component terminates, and phase 2 begins.

Section 3.4: Phase 2 of the n^{th} Stage of the Generative Component.

At the beginning of Section 3.3, it was stated that the storage area $\mathcal{R}^n$ is used in phase 2 for the purpose of partially computing $\mathcal{R}^{n+1}$. A diagram of the order of operations in phase 2 appears in Figure 4. An explanation of this diagram constitutes the remainder of this section.

Figure 4: Diagram of operations in phase 2 of the n^{th} stage of the generative component.



As phase 2 begins, sets from the completely computed storage area $\mathfrak{R}^n$ are inspected one by one until the storage area is exhausted, at which time "GEN: $\mathfrak{R}^n$;X" results in a "NO" and phase 2 ends. Consider a typical set X from $\mathfrak{R}^n$ which is put in the work space. "GEN: T—X;x" directs that elements of T—X be generated one by one (in alphabetical order) until T—X is exhausted. For each $x \in T-X$, the following TEST operations are performed on the set $X \cup \{x\}$ (which is of cardinality n+1).

TEST 1: $X \cup \{x\} \in \mathfrak{R}^{n+1}$? This TEST scans the partially computed storage area $\mathfrak{R}^{n+1}$ to see whether or not $X \cup \{x\}$

is already in this storage area. If a "YES" results, a new element x of T—X (if it exists) is generated; and if a "NO" results, TEST 2 is then performed.

TEST 2: $M_X(X)=0$? If a "YES" results, then $\{x\} \not\subseteq X$ (by THEOREM 2-i); so $X \cup \{x\} \notin \mathcal{R}$ (by THEOREM 9-ii, since $X \notin \mathcal{R}$ and $\{X, \{x\}\}$ is separated). Then a copy of $X \cup \{x\}$ is stored in $\mathcal{R}^{n+1}$, and a new element in T—X (if it exists) is generated. If a "NO" results, TEST 3 is performed.

TEST 3: $M_X(T-(X \cup \{x\})) \leqslant M_X(X \cup \{x\})$? If a "YES" results, nothing is determined; and if a "NO" results, $X \cup \{x\}$ is clearly not an R-clump; so a copy of $X \cup \{x\}$ is stored in $\mathcal{R}^{n+1}$. In either case, a new element of T—X (if it exists) is then generated.

This sequence of three TEST operations continues until T—X is exhausted of elements at which time the "GEN: T—X;x" instruction results in a "NO". The process then moves back to the "GEN: $\mathcal{R}^n$;X" instruction, and a new set (if it exists) is input from $\mathcal{R}^n$. After $\mathcal{R}^n$ is emptied, phase 2 of the n^{th} generative component terminates, with $\mathcal{R}^{n+1}$ partially computed; and the $(n+1)^{th}$ stage of the analytic component begins.

Section 3.5: Concluding Remarks.

In the computation of strong clumps, it is of interest to compute the sets $\mathcal{A}^n$ and $\mathcal{S}^n$ where $n \in \mathbb{N}$, $n \leqslant \|T\|$. For each such n, $\mathcal{A}^n$ and $\mathcal{S}^n$ are completely computed by the end of the n^{th} analytic component. At the end of the

$(\|T\|-1)^{th}$ analytic component, simply add T to the storage area $\mathcal{S}^{\|T\|}$ and if, for some $t \in T$, t belongs to no previously computed set in $\mathcal{A}^n$, n=1,2,...,$\|T\|-1$, also add T to the storage area $\mathcal{A}^{\|T\|}$. At this point the actual computation procedure is terminated. Due to the structure of the GEN operation, all sets of $\mathcal{A}^n$ and $\mathcal{S}^n$ are stored in alphabetical order.

One may, if one wishes, eliminate TEST 3 of phase 2 of the n^{th} generative component. In fact, one may wish to eliminate all instructions in the n^{th} analytic component dealing with the storage area $\mathcal{R}^n$ as well as all of phase 2 of the n^{th} generative component. Such alternatives entail increased computation time in TESTS 4 through 9 in the n^{th} analytic component. Alternatively, one may wish to retain instructions dealing with $\mathcal{R}^n$ while dropping TESTS 5 through 8 (the "short check") in the n^{th} analytic component. In fact, any combination of these operations may be dropped. Thus, it is clear that one has various options in writing a computer program to realize this computation procedure. The optimum program depends entirely on the efficiency of the program relative to the particular (T,M) structures to which the computation is applied.

It should be mentioned that the strong clump computation procedure need not be carried out entirely in this manner. For example, if a negligible number of new atomic R-clumps is added in the first k analytic components after the n^{th} analytic component, one may simply use $(n+k)(\mathcal{A})$

and perform connectedness tests on those clumps of cardinality greater than n+k-1. This allows one to estimate most of the strong clumps in the (T,M) structure.

Finally, it should be remarked that it is not always necessary to completely compute or even to estimate the class of all strong clumps. It may be sufficient to compute only those with cardinality less than a certain fixed value, depending upon the application. For example, such a short-cut could be taken when using strong clumps in man-machine negotiation or in informal vocabulary control. It is probably through such partial computations in pilot programs that the practical value of strong clumps as a tool in information retrieval may be more definitely ascertained.

REFERENCES

[1] Dale, A. G., Dale, N., and Prendergraft, E. D., "A Programming System for Automatic Classification with Applications in Linguistic and Information Retrieval Research". University of Texas Linguistics Research Center, Austin, Texas, October, 1964. National Science Foundation, Grant No. GN-208.

[2] Halmos, Paul R., Measure Theory. D. Van Nostrand Company, Inc., Princeton, New Jersey, 1950.

[3] Hillman, Donald J., "Characterization and Connectivity", Report No. 1 of Document Retrieval Theory, Relevance, and the Methodology of Evaluation. Lehigh University Center for the Information Sciences, Bethlehem, Pennsylvania, May 24, 1966. National Science Foundation, Grant No. GN-451.

[4] Needham, R. M., "The Theory of Clumps,II", Report No. ML-139. Cambridge Language Research Unit, Cambridge, England, March, 1961.

[5] Parker-Rhodes, A. F. and Needham, R. M., "The Theory of Clumps", Report No. ML-126. Cambridge Language Research Unit, Cambridge, England, February, 1960.

VITA

Robert Charles Heiser, son of Leroy F. and Shirley J. (Kepler) Heiser, was born in West Reading, Pennsylvania, on January 3, 1945. He attended Reading Senior High School from which he graduated as class salutatorian in June, 1962. At Lehigh University, Mr. Heiser was a member of Phi Eta Sigma, a national scholastic honorary for freshmen, and graduated with a B. A. in mathematics in June, 1966. He then served as a research assistant at the Center for the Information Sciences (1966-1969) and as a teaching assistant for the Philosophy Department (1968-1970) while pursuing graduate study at Lehigh University.

Mr. Heiser is married to the former Ann Catharine Dougherty of Schnecksville, Pennsylvania, and has one daughter, Deborah Susan.