ON TIME CHANGES IN A DIGRAPH*

by

T. N. Bhargava

TECHNICAL REPORT NO. 3

PREPARED UNDER RESEARCH GRANT NO. NsG-568
(PRINCIPAL INVESTIGATOR: T. N. BHARVAVA)

FOR

NATIONAL AERONAUTICS and SPACE ADMINISTRATION

DEPARTMENT OF MATHEMATICS

KENT STATE UNIVERSITY

KENT, OHIO

June, 1964

* Part of a dissertation for the Ph.D. degree, Michigan State University, 1962; with partial support of National Science Foundation grant number G-18976. Thanks are due to major adviser Professor Leo Katz.

18941

The object of this paper is to present a probabilistic model for analyzing changes through time in a binary dyadic relation on a finite set of points. The total relation on the set takes the form of an aggregate of directed binary dyadic relations between ordered pairs of points belonging to the set; equivalently, the total relation on the set can be represented by means of a digraph or an incidence matrix isomorphic with the total relation. Such a relation, changing in its structure as time proceeds, is a reasonable mathematical model, for example, for the evolution of interrelationships of the members of a social or any other group. A group of this kind is organized for a specific activity involving some sort of "communication" from one member to the other, and may be observed at successive discrete points in time generating statistics on the evolutionary process. (For a detailed treatment, see [3]). As a matter of fact, under suitable assumptions, the model presented here has potentialities for application in those situations which can be represented mathematically in terms of a finite set of points and an all-or-none relationship between ordered pairs of these points. Some of the other examples are communication networks, ecology, animal sociology, and management sciences (see [5]).

1. Introduction: Let $A = \{P_1, P_2, ---, P_N\}$ be a finite set consisting of N points, and R a binary relation defined on the set A for each ordered pair of distinct points P_i, P_j belonging to A. If $R(i,j)$ donotes the relation for the ordered pair (P_i, P_j) then either $R(i,j) = 1$ or $R(i,j) = 0$. The total relation on the set is the aggregate of the directed binary dyadic relations $R(i,j)$ and is denoted by $R(A)$. As time proceeds, some of the relations $R(i,j)$ may undergo a change while others may not. Thus $R(A)$

is a function of time, and the process may be observed at successive discrete points in time, generating statistics on the evolutionary process.

There are two common procedures for obtaining a sufficiently large number of observations for purposes of drawing statistical inferences about a process in time: (i) making a series of observations on a single set for a sufficiently long time, or (ii) making observations in parallel on a number of similar sets for a relatively shorter period of time. However, in many situations we are excluded from using either of these procedures due to ground rules laid down in certain applications. A device useful in these circumstances is to consider the total relation R(A) in terms of an aggregate of the subrelations R(a), where $a = \{P_{i_1}, P_{i_2}, \text{---}, P_{i_n}\}$ is a subset, of fixed size n, of A. For a detailed discussion of this approach, we refer the reader to [2] or [3].

There are two equivalent representations of R(A): first, in terms of a directed graph (or simply digraph) Γ(A), and second, in terms of an incidence matrix C(A). We give below some relevant definitions and notations.

The digraph Γ(A) consists of N vertices $P_1, P_2, \text{---}, P_N$ corresponding to the points of A, and directed edges $\overrightarrow{P_iP_j}$ shown by means of directed lines from P_i to P_j corresponding to R(i,j) = 1; R(i,j) = 0 is shown by absence of directed edges from P_i to P_j. Γ(A) is said to be of order N if it consists of exactly N vertices irrespective of the number of edges. A directed path from P_i to P_j is given by a chain of directed edges of the form $\overrightarrow{P_iP_{i_1}}, \overrightarrow{P_{i_1}P_{i_2}}, \text{---}, \overrightarrow{P_{i_L}P_j}$. We note that by definition, presence of loops (directed edges joining a point to itself) is not allowed, and that there cannot be more than one directed edge joining any ordered pair of points. A directed path of length L (L > 1) from P_i to itself is called a cycle of length L. Finally, a point P_j is said to be accessible from a point P_i if there is a directed path from P_i to P_j; when this happens the ordered pair $(P_i P_j)$ is called an accessible pair.

The incidence matrix C(A) is a matrix of zeros and ones such that its rows and columns correspond to the points in A, and the element C_{ij} of C(A) is one, or zero corresponding to R(i,j) = 1 or 0. A matrix C is said to be positive if all its elements are positive, and is denoted by C > 0. The element-wise (Hadamard) product of two matrices C and D is denoted by C*D.

Corresponding to the subrelation R(a) there is a subdigraph Γ(a) of Γ(A), consisting of all the directed edges in A which connect points of subset a, and a principal diagonal submatrix C(a), of C(A), such that its rows and columns correspond to the points of the subset a.

2. <u>Classification of relations</u>. It is easily seen that the number of different forms in which R(A) may be observed is $2^{N(N-1)}$, which is extremely large. However, for statistical purposes we require that the number of distinct "states" in which R(A) may be observed should not be excessive. We may, first of all, ignore the labels of the set A (in those cases where the labelling of points is of no particular importance), and obtain equivalence classes, under permutations, of the class of relations R(A). The number of these equivalence classes, although much smaller than the total number of different forms of R(A), is still quite large. For example, for N = 5 these numbers are

8,508 and 1,048,576 respectively.

Next, we observe that relations belonging to different, but nearly alike, equivalence classes are not modified very much by the addition or deletion of a few directed dyadic relations $R(i,j)$. Accordingly, we may group some of these equivalence classes together to obtain more manageable numbers of empirically observable, mutually exclusive and exhaustive classes, such that at any point in time $R(A)$ may belong to one and only one of these classes. Clearly this could be achieved in many ways. We present here three particular classification schemes: the connectivity classification, the accessibility classification, and the weighted classification. For this purpose we find it convenient to use the graphic representation, $\Gamma(A)$, of the relation $R(A)$.

The connectivity classification is based on the strength of connectedness and is a standard topological classification: $\Gamma(A)$ is said to be strongly connected if each point of A is accessible from every other point of A; unilaterally connected if for every pair of points belonging to A there is a directed path from at least one of them to the other; weakly connected if there is a chain of connections, ignoring all directions, from each point of A to every other point of A; disconnected if it is not even weakly connected. These classes may be defined more strictly, in an obvious manner, to obtain four mutually exclusive and exhaustive classes: $\Gamma(A)$ is said to be in state s_3 if it is strongly connected; in state s_2 if it is unilaterally connected but not strongly connected; in state s_1 if it is weakly connected but not unilaterally connected; in state s_0 if it is disconnected.

The accessibility classification is based on the number of accessible ordered pairs in the digraph $\Gamma(A)$: $\Gamma(A)$ is said to be in state s_i' if it consists of exactly i accessible ordered pairs. By definition the accessibility classes are mutually exclusive and exhaustive, and the number of classes depends on N, the order of $\Gamma(A)$.

The structure of the relation R(A) is determined by the symmetric relation of accessibility in both directions between pairs of elements of A. This relation is an equivalence relation on the set A, separating it into disjoint equivalence sets, in each of which every point is accessible from every other point. Further, the residual one-way relations induce a partial ordering on these equivalence sets. Let there be m equivalence sets $A_1, A_2, ---, A_m$ of sizes $N_1, N_2, ---, N_m$, respectively, where $\{A_i \mid i = 1,2,---,m\}$ constitutes a partition of A, and $\sum_{i=1}^{m} N_i = N$. If p denotes the number of accessible ordered pairs in (A), it can be shown (see [2]).

Theorem 2.1 ρ can take only values of the form

$$\sum_{i \leq j} \epsilon_{ij} N_i N_j - N$$

where $\epsilon_{ij} = 1$, if the sets A_i and A_j are ordered in the partial ordering ($\epsilon_{ii} = 1$), and $\epsilon_{ij} = 0$ otherwise.

The weighted classification is a refinement of the accessibility classification, obtained by assigning to an accessible ordered pair (P_i, P_j) a weight of N-L if the length of minimal directed path from P_i to P_j is L, and a weight of zero to all those ordered pairs which are not accessible.

Details of these classifications may be found in [2] or [3]. In this paper we discuss some statistical results for the first two classifications, namely, the connectivity and the accessibility classifications.

3. <u>Enumeration of subrelations</u>. Since our approach consists of considering $R(A)$ as an aggregate of subrelations $R(a)$, it is necessary that we find formal procedures for enumeration of subrelations $R(a)$ which are, (i) in states s_i $(i = 0,1,2,3)$ at any fixed time t, and (ii) in states s_{0j}, s_{1j}, s_{2j}, s_{3j} at times t_j, $j = 1,2,---,h$.

Let $R_{ni}(t)$ denote the number of subrelations of size n, which are in state s_i $(i = 0,1,2,3)$ at time t. To find the value of $R_{ni}(t)$, for any t, we define a function f on the class of subrelations, $\{R(a), a \subset A\}$, with the set of states $\{s_i\}$ as its range: we say that

$$f(R(a)) = \begin{cases} s_i & \text{if } R(a) \text{ is in state } s_i \\ 0 & \text{otherwise.} \end{cases}$$

Thus $R_{ni}(t)$ is equal to the number of those $R(a)$ for which $f(R(a)) = s_i$ at time t, and the main problem is that of being able to evaluate the function $f_t(R(a))$ at time t for every $R(a)$. We also note that the total number of subrelations $R(a)$, of $R(A)$, is $\binom{N}{n}$.

<u>Theorem 3.1</u>

$$f(R(a)) = \begin{cases} s_3 & \text{if, and only if, } [C^*(a)]^{n-1} > 0. \\ s_3 \text{ or } s_2 & \text{if, and only if, } \{[C^*(a)]^{n-1} + [C^{1*}(a)]^{n-1}\} > 0. \\ s_3 \text{ or } s_2 \text{ or } s_1 & \text{if, and only if, } [C^*(a) + C^{1*}(a)]^{n-1} > 0. \\ s_0 & \text{if, and only if, } [C^*(a) + C^{1*}(a)]^{n-1} \not> 0. \end{cases}$$

where C(a) is the corresponding matrix representation(defined earlier) of R(a); and $C^*(a) = C(a) + I$, I being the n x n identity matrix.

A proof of this theorem may be found in [2] or [4].

This theorem can be stated more strictly in an obvious way to obtain the values of $R_{ni}(t)$. Furthermore, it can be programmed on a computer to carry out the computations in an easier and faster way. It also follows that repeated application of this theorem gives us the number of various (h-1)-order transitions from state s_{i1} at time t_1 to state s_{ih} at time t_h through states s_{i2}, ---, $s_{i,h-1}$ at times t_2, ---, t_{h-1} respectively.

Similarly, if $R'_{ni}(t)$ denotes the number of n-point subrelations in state s'_i at time t, then the following theorem gives the values of $R'_{ni}(t)$ for all admissible values of i.

Theorem 3.2

$f'(R(a)) = s'_i$ if, and only if, the number of non-zero entries in the matrix $[C^*(a)]^{n-1}$ is exactly n + i, where f' is a function on $\{R(a), a \subset A\}$ with its range as the set $\{s'_i\}$. The proof of this theorem may be found in [2].

As before, a repeated application of this theorem counts the number of (h-1)-order transitions for the accessibility classification, and it can be programmed on a computer for computational purposes.

It may be mentioned here that an (h-1)-order transition table for a given set of data can be reduced to obtain all the transition tables of order less than h, as well as the values of $R_{ni}(t)$ (or $R'_{ni}(t)$) for $t = t_1, t_2, ---, t_h$.

4. Stochastic process. The change in R(A) over a period of time may now be described as a stochastic process and standard

tests and techniques may be used to draw statistical inferences in time for such a process. The simplest case is that of a Markov chain model with unspecified Markovian property, and such a model can be analyzed, using the standard methods as, for example, in [1].

5. Distribution theory. The derivation of the probability distributions of random variables $R_{ni}(R'_{ni})$ is based on the theory of compound probabilities, an exposition of which may be found, for instance, in Feller [6], and Fréchet [7].

Let P_i denote the probability of the occurrence of an event E_i, $i = 1,2,\text{---},M$ (E_i's are not necessarily independent); P_{ij} denote the probability of the simultaneous occurrence of the events E_i, E_j ($i,j = 1, 2, \text{---}, M$; $i \neq j$); P_{ijk} denote the probability of the simultaneous occurrence of the events E_i, E_j, E_k ($i, j, k = 1, 2, \text{---}, M$; $i \neq j \neq k$); and so on. Define sums

$$S_1 = \sum_i P_i$$

$$S_2 = \sum_{i,j} P_{ij}$$

$$S_3 = \sum_{i,j,k} P_{ijk} \text{ ---, and so on.}$$

If $P_{(m)}$ and $P_{[m]}$ be respectively the probabilities of the simultaneous occurrence of exactly, and at least m among M events, then the principle of inclusion and exclusion (Feller, [6], Chapter 4) gives immediately

$$P_{(m)} = S_m - \binom{m+1}{m} S_{m+1} + \binom{m+2}{m} S_{m+2} - \text{---} \pm \binom{M}{m} S_M$$

and

$$P_{[m]} = P_{(m)} + P_{(m+1)} + \text{---} + P_{(M)}$$

$$P_{[m]} = S_m - \binom{m}{m-1} S_{m+1} + \binom{m+1}{m-1} S_{m+2} - \text{---} \pm \binom{M-1}{m-1} S_M$$

Let $\alpha_{(k)}$ denote the K-th factorial moment of the distribution. Then

$$\alpha_{(k)} = k!\, S_k \quad \text{(Frechet [7])}.$$

In particular,

Mean = S_1

Variance = $2S_2 + S_1 - S_1^2$

We make use of the above relationships and formulae to derive results for the probability distributions of the random variables $R_{ni}(R'_{ni})$.

Let the point P_i hold relation R with d_i ($0 \leq d_i \leq N-1$) other points in the set A; in digraph $\Gamma(A)$, d_i is the number of directed edges starting from P_i. We say that d_i is the "valency" of the point P_i, and assume that for any i, d_i is fixed, and the d_i directed edges emanating from P_i follow a hypergeometric law. We also assume that the distribution of the terminals of the directed edges originating at P_i is completely random. If all the d_i's are equal ($d_1 = d_2 = \text{---} = d_N = d$, say) we have the simple restricted case; otherwise the general unrestricted case. The results for the general case get quite complicated and as such most of the results presented here are for the restricted case.

Under the assumptions stated above, it is easy to see that

$$P[R(i,j) = 1] = d_i/(N - 1),$$
$$P[R(i,j) = 0] = 1 - d_i/(N - 1),$$
$$P[R(i,j) = 1 \mid R(i,k) = 1] = (d_i-1)/(N-2),$$
$$P[R(i,j) = 1 \mid R(i,k) = 0] = d_i/(N-2),$$
$$P[R(i,j) = 0 \mid R(i,k) = 1] = 1 - (d_i- 1)/(N-2),$$
$$P[R(i,j) = 0 \mid R(i,k) = 0] = 1 - d_i/(N-2),$$

and so on.

We can now derive, at least theoretically, the probabilities of the type $P[f(R(a))= s_i]$, $P[f(R(a_1))= s_i,\ f(R(a_2))= s_j]$, ---, where a_1, a_2 are two distinct subsets, of size n, of the set A. By computing the sums of the form

$$\sum_a P[f(R(a))= s_i],$$

$$\sum_{a_1, a_2} P[f(R(a_1))= s_i,\ f(R(a_2))= s_j], \text{ ---}$$

we may obtain the probabilities $P_{(m)}$, $P_{[m]}$, the moments $\alpha_{(k)}$, and the exact probability distributions of the random variables $R_{ni}(R'_{ni})$.

However, these computations become extremely tedious even for small values of n, and the derivation of exact distributions becomes almost impossible. As a consequence we propose, later on, simple approximations which are comparatively much easier to obtain, and are, therefore, of practical importance. Furthermore, although the results for the restricted case may be obtained as a special case of the general case (by putting $d_1 = d_2 = \text{---} = d_N = d$, say), we present the two cases separately due to the marked simplicity of the former.

In what follows, we need the following notations. Factorial notation: $\chi^{(\theta)} = \chi(\chi - 1) \text{ --- } (\chi - \theta + 1)$, functions $u_m = \sum_{i_1<i_2<\text{---}<i_m} d_{i_1} d_{i_2} \text{ --- } d_{i_m}$; and power sums $v_m = \sum_i d_i^{\,m}$. The functions u_m may be expressed in terms of the sums v_m (Perron, [9]); for $m \leq 4$, these relations are

$$u_1 = v_1$$
$$u_2 = (v_1^2 - v_2)/2!$$

$$u_3 = (v_1^3 - 3v_2v_1 + 2v_3)/3!$$
$$u_4 = (v_1^4 - 6v_2v_1^2 + 3v_2^2 + 8v_3v_1 - 6v_4)/4!$$

6. Dyadic case: $n = 2$. For digraphs (or subdigraphs) of order two, the connectivity and the accessibility classifications are identical; there are only three different states ($s_3 \equiv s_2'$; $s_2 \equiv s_1'$; $s_0 \equiv s_0'$) in which a two-point subrelation may be found, and $R_{23} \equiv R_{22}'$, $R_{22} \equiv R_{21}'$, $R_{20} \equiv R_{20}'$. (See [4]). Let $a = \{P_j, P_k\}$ be a two-point subset of the set A, and let R(a) be the corresponding subrelation. It is seen easily that R(a) is in state s_3 if there are two directed edges in $\Gamma(a)$: one from P_j to P_k, and the other from P_k to P_j; in state s_2 if there is only one directed edge in $\Gamma(a)$: either from P_j to P_k, or from P_k to P_j; in state s_0 if there are no directed edges at all in $\Gamma(a)$. As such,

$$P[f(R(a)) = s_3] = d_j d_k/(N-1)^2,$$
$$P[f(R(a)) = s_2] = \{d_j(N-1-d_k) + d_k(N-1-d_j)\}/(N-1)^2,$$
and $$P[f(R(a)) = s_0] = (N-1-d_j)(N-1-d_k)/(N-1)^2.$$

Furthermore,

$$E(R_{2i}) = S_{1i} = \sum_a P[f(R(a)) = s_i]$$

where the summation extends over all the $\binom{N}{2}$ unordered pairs in the set A. We may write

$$E(R_{2i}) = \sum_{j<k=1}^{N} P[f(R(a)) = s_i]$$

and using the values of $P[f(R(a)) = s_i]$, given above, obtain the expressions for $E(R_{2i})$, $i = 0,2,3$. In the restricted case (all d_i's equal to d, say) the probabilities $P[f(R(a)) = s_i]$ remain the same for any of the subsets a, and therefore,

$$E(R_{2i}) = \binom{N}{2} P[f(R(a)) = s_i].$$

Simple reductions give us the following theorem:

Theorem 6.1 For the general case,

$$E(R_{23}) = (v_1^2 - v_2)/\ 2(N-1)^2$$

$$E(R_{22}) = v_1 - (v_1^2 - v_2)/(N-1)^2$$

$$E(R_{20}) = N(N-1)/2 - v_1 + (v_1^2 - v_2)/\ 2(N-1)^2.$$

For the restricted case,

$$E(R_{23}) = Nd^2/\ 2(N-1)$$

$$E(R_{22}) = Nd(N-1-d)/(N-1)$$

$$E(R_{20}) = N(N-1-d)^2/\ 2(N-1)$$

Let $V(R_{ni})$ denote the variances of the random variables R_{ni}.

Theorem 6.2 For the general case,

$$V(R_{23}) = (6u_4 - u_2{}^2)/(N-1)^4 + 2(-4u_4 + u_3u_1 - 3u_3)/ (N-1)^3(N-2) + u_2/(N-1)^2$$

$$V(R_{22}) = 4(6u_4 - u_2{}^2)/(N-1)^4 + 8(-4u_4 + u_3u_1 - 3u_3)/ (N-1)^3(N-2) + 4u_2/(N-1)^2$$

$$V(R_{20}) = (6u_4 - u_2{}^2)/(N-1)^4 + 2(-4u_4 + u_3u_1 - 3u_3)/ (N-1)^3(N-2) + u_2/(N-1)^2$$

For the restricted case,

$$V(R_{23}) = Nd^2(N-1-d)^2/\ 2(N-1)^3$$

$$V(R_{22}) = 2Nd^2(N-1-d)^2/(N-1)^3$$

$$V(R_{20}) = Nd^2(N-1-d)^2/\ 2(N-1)^3$$

Proof. Let S_{mi} denote the sum S_m, defined in section 5, for the random variable R_{2i} $(m = 1,2,---)$, so that $V(R_{2i}) = 2S_{2i} + S_{1i} - S_{1i}{}^2$. The sums S_{1i}, equal to $E(R_{2i})$, already have been obtained in the previous theorem, and as such, we only need to compute the sums S_{2i} $(i = 3,2,0)$.

$$S_{2i} = \sum_{a_1,a_2} P[f(R(a_1)) = s_i,\ f(R(a_2)) = s_i]$$

where $a_1 = \{P_j, P_k\}$, $a_2 = \{P_h, P_\ell\}$ are any two different subsets of the set A, and the summation extends over all such subsets of A. If the subsets a_1, a_2 are composed of four different points, then the two events $[f(R(a_1)) = s_i]$, $[f(R(a_2)) = s_i]$ are independent, and

$$P[f(R(a_1)) = s_i,\ f(R(a_2)) = s_i] = P[f(R(a_1)) = s_i] \cdot P[f(R(a_2)) = s_i].$$

However, if the subsets have one point in common (if both the points are common, we do not have <u>different</u> subsets a_1, a_2), say $P_h = P_k$, then the events $[f(R(a_1)) = s_i]$, $[f(R(a_2)) = s_i]$ are not independent, and we have to compute the compound probability $P[f(R(a_1)) = s_i$ and $f(R(a_2)) = s_i]$.

Thus,

$$S_{2i} = \Sigma P[f(R(a_1)) = s_i] \cdot P[f(R(a_2)) = s_i] + P[f(R(a_1)) = s_i \text{ and } f(R(a_2)) = s_i]$$

where the first summation is over all pairs of disjoint subsets a_1, a_2, and the second summation is over the remaining pairs of (different) subsets a_1, a_2 (with one point in common).

Each set of four different points may form a pair of dyads in three ways, and each set of three different points may have any of the three points as the common point in the pair of dyads. With these considerations, we may write

$$S_{2i} = 3 \sum_{\substack{h<j<k<\ell \\ =1}}^{N} P[f(R(a_1))= s_i]\cdot P[f(R(a_2))= s_i]$$

$$+ 3 \sum_{\substack{j,k,\ell=1 \\ j<k<\ell}}^{N} P[f(R(a_1))= s_i \text{ and } f(R(a_2))= s_i]$$

These probabilities are easily obtainable, and these sums now can be computed in terms of the power sums v_p's ($p= 1,2,---$). The expressions $V(R_{2i}) = S_{2i} + S_{1i} - S_{1i}^2$ then are computed to give the variances for the general case. The variances for the restricted case can be obtained by putting $d_1= d_2=---= d_N= d$ in the expressions for the general case; however, there is rather a simple way of obtaining them directly.

For fixed i, the probability $P[f(R(a_1))= s_i]$ is the same for any subset a_1 of A; the compound probability $P[f(R(a_1)) = s_i$ and $f(R(a_2)) = s_i]$, where a_1 and a_2 have one point in common, is also the same for every such pair of subsets. Further, the number of ways of choosing four different points from N points and forming two dyads from them is $3\binom{N}{4}$; the number of ways of choosing three different points from N points and forming two dyads from them is $3\binom{N}{3}$. Therefore,

$$S_{2i} = 3\binom{N}{4} P[f(R(a))= s_i]^2 + 3\binom{N}{3} P[f(R(a_1))= s_i, \text{ and } f(R(a_2))= s_i]$$

where the subsets a_1 and a_2 have one point in common. The sums S_{1i}, being known, the variances $V(R_{2i})$ can now be computed rather easily.

We note that the computation of the compound probability $P[f(R(a_1))= s_i$, and $f(R(a_2))= s_i]$ has to be carried out rather carefully. For example the compound event $[f(R(a_1))= s_2$ and $f(R(a_2))= s_2]$ may take place in any one of the following four ways:

In the subdigraph $\Gamma(a_1)$ the directed edge may be from either P_j to P_k or from P_k to P_j; with either of these two cases the subdigraph $\Gamma(a_2)$ may also have a directed edge either from P_k to P_ℓ or from P_ℓ to P_k. The probabilities for each of these four cases will be different in general and need to be evaluated separately and carefully.

To obtain higher moments of the random variables R_{2i}, we need to compute the sums S_{mi}, $m \geq 3$. The number of different types of terms in S_m rapidly increases and the resulting expressions for S_m become rather complicated algebraically, but the principles of computation remain the same. We give here the derivation of $S_{3,3}$, for the restricted case, to illustrate the method. The sum

$$S_{3,3} = \Sigma \Pr \{[f(R(a_1)) = s_3], [f(R(a_2)) = s_3], [f(R(a_3)) = s_3]\}$$

where the summation extends over all the triples of two-point subsets $a_1 = \{P_i, P_j\}$, $a_2 = \{P_k, P_\ell\}$, $a_3 = \{P_m, P_n\}$. For any particular triple of subsets a_1, a_2, and a_3 the number of distinct points in the three subsets may be six, five, four or three; there is only one way of forming a triple of two-point subsets on three points. If the points are P_i, P_j, P_ℓ, then the only possible triple is $\{P_i, P_j\}$, $\{P_i, P_\ell\}$, and $\{P_j, P_\ell\}$. Thus S_{33} consists of five different kinds of terms, and there are $\binom{N}{6}\frac{6!}{3!2!2!2!}$ different ways of obtaining the first kind of term (six distinct points); $\binom{N}{5}\frac{5!}{2!2!}$ ways of obtaining the second kind of term (five distinct points); $\binom{N}{4}\frac{4!}{2!}$ ways of obtaining the third kind of term (four distinct points); $\binom{N}{4}\frac{4!}{2!}$ ways of obtaining the fourth kind of term (four distinct points); $\binom{N}{3}$ ways of obtaining the fifth kind of term (three distinct points). The probability element of each type is determined easily, and

$$S_{33} = \binom{N}{6}\frac{6!}{3!\,(2!)^3}\left[\frac{d^2}{(N-1)^2}\right]^3 + \binom{N}{5}\frac{5!}{(2!)^2}\left[\frac{d^2}{(N-1)^2}\right]^2\frac{d}{(N-1)}\,\frac{(d-1)}{(N-2)}$$

$$+ \binom{N}{4}\frac{4!}{2!}\left[\frac{d^2}{(N-1)^2}\right]^2\frac{(d-1)^2}{(N-2)^2} + \binom{N}{4}\frac{4!}{3!}\left[\frac{d^2}{(N-1)^2}\right]^2\frac{(d-1)}{(N-2)}\,\frac{(d-2)}{(N-3)}$$

$$+ \binom{N}{3}\frac{d^2}{(N-1)^2}\,\frac{d(d-1)}{(N-1)(N-2)}\,\frac{(d-1)^2}{(N-2)^2}$$

7. <u>Triadic Case</u>: $n = 3$. The complexity of problems concerned with the probability distributions of the random variables R_{ni} (R'_{ni}) increases many times for $n > 2$. In general, a subrelation may belong to a given state s_i (s'_i) through several different kinds of structures (non-labelled digraphs). Consequently, the computations of various probabilities of the type $P[f(R(a)) = s_i]$, $P[f(R(a_1)) = s_i$, $f(R(a_2)) = s_i]$, --- become increasingly more involved, and it becomes almost impossible to obtain the various sums S_m even for $m = 2$, and the restricted case. As a matter of fact, for the general case, even for $n = 2$, the expressions for S_2 are quite involved (see last section). However, the principle inovlved is quite simple, and theoretically there seems to be no difficulty in computing these.

Let a subrelation $R(a)$ be found in state s_i through one, and only one, of I different structures s_{i1}, s_{i2}, --- s_{iI}

(each of type s_i). Then

$$E(R_{3i}) = S_{1i} = \sum_a P[f(R(a)) = s_i]$$

$$= \sum_a \sum_{p=1}^{I} P[f(R(a)) = s_{ip}]$$

For the restricted case,

$$E(R_{3i}) = S_{1i} = \binom{N}{3} \sum_{p=1}^{I} P[f(R(a)) = s_{ip}]$$

$$S_{2i} = \sum_{a_1,a_2} P[f(R(a_i)) = s_i \text{ and } f(R(a_2)) = s_i]$$

where a_1 and a_2 are any two distinct three-point subsets of the set A. We may rewrite

$$S_{2i} = \sum_{a_1,a_2} \sum_{p,q=1}^{I} P[f(R(a_1)) = s_{ip} \text{ and } f(R(a_2)) = s_{iq}]$$

$$= \sum_{a_1,a_2} \sum_{p=1}^{I} P[f(R(a_1)) = s_{ip} \text{ and } f(R(a_2)) = s_{ip}]$$

$$+ \sum_{a_1,a_2} \sum_{p\neq q=1}^{I} P[f(R(a_1)) = s_{ip} \text{ and } f(R(a_2)) = s_{iq}]$$

$$= \sum_{p=1}^{I} S_{2ip} + \sum_{p<q=1}^{I} E(R_{3ip}R_{3iq})$$

where $S_{2ip} = \sum_{a\ ,a} P[f(R(a_1)) = s_{ip} \text{ and } f(R(a_2)) = s_{ip}]$.

To calculate S_{2ip} and $E(R_{3ip}R_{3iq})$ for the restricted case, we notice that the various probabilities do not depend on which subsets a_1 and a_2 are taken, and there are $\binom{N}{6}\frac{6!}{3!3!2!}$ different pairs of triads (three-point subsets) with no points in common; $\binom{N}{5}\frac{5!}{2!2!2!}$ different pairs of triads with one point in common; $\binom{N}{4}\frac{4!}{2!2!}$ different pairs of traids with two points in common. Therefore,

$$S_{2ip} = \binom{N}{6}\frac{6!}{3!3!2!} \{P[f(R(a)) = s_{ip}]\}^2$$
$$+ \binom{N}{5}\frac{5!}{2!2!2!} P[f(R(a_i)) = s_{ip} \text{ and } f(R(a_2)) = s_{ip} \mid a_1 a_2 \text{ have one point in common }]$$

and,

$$E(R_{3ip}R_{3iq}) = \binom{N}{6}\frac{6!}{3!3!2!} P[f(R(a)) = s_{ip}]\cdot P[f(R(a))=s_{iq}]$$
$$+ \binom{N}{5}\frac{5!}{2!2!2!} P[f(R(a_1))= s_{ip} \text{ and } f(R(a_2))= s_{iq} \mid a_1a_2 \text{ have one point in common}]$$
$$+ \binom{N}{4}\frac{4!}{2!2!} P[f(R(a_1))= s_{ip} \text{ and } f(R(a_2))= s_{iq} \mid a_1a_2 \text{ have two points in common}]$$

We give below, without proof, the expected values of the random variables R_{3i} and R'_{3i} for the restricted case. The expressions for the variances are quite lengthy and as such are not given here; these may be found, however, in [2].

Theorem 7.1 For the restricted case,

$$E(R_{33}) = Nd^3[2N^3 - 15N^2 + 30N - 18 + 3d(N^2-3) -6d^2(N-1) + 2d^3] / 6(N-1)^2(N-2)^2$$

$$E(R_{32}) = N(N-1-d)d^2\ [2N^3 - 12N^2 + 23N - 14 + d(5N-8) -4d^2(N-1) + 2d^3] \ / \ 2(N-1)^2(N-2)^2$$

$$E(R_{31}) = Nd(N-1-d)^2(N-2-d)\ [-(N-2) + 2d(N-1) - 2d^2] \ / \ 2(N-1)^2(N-2)^2$$

$$E(R_{30}) = N(N-1-d)^3(N-2-d)\ [(N-2)^2 + 4d(N-2) - 2d^2] \ / \ 6(N-1)^2(N-2)^2$$

<u>Theorem 7.2</u> For the restricted case,

$$E(R'_{36}) = E(R_{33}), \text{ given in theorem 7.1}$$

$$E(R'_{34}) = Nd^2(N-1-d)\ (2N-3-d)\ [-(N-2) + 2d(N-1) - 2d^2] \ / \ 2(N-1)^2(N-2)^2$$

$$= Nd^2(N-1-d)\ [-(2N-3)(N-2) + d(4N^2 - 9N + 4) \ / \ 2(N-1)^2(N-2)^2$$

$$E(R'_{33}) = Nd^2(N-1-d)^2\ (N-2-d) \ / \ (N-1)^2(N-2)$$

$$E(R'_{32}) = Nd(N-1-d)^2\ (N-2-d)\ [-(N-2) + 3d(N-1) - 3d^2] \ / \ 2(N-1)^2\ (N-2)^2$$

$$E(R'_{31}) = Nd(N-1-d)^3\ (N-2-d)^2 \ / \ (N-1)^2\ (N-2)^2$$

$$E(R'_{30}) = N(N-1-d)^3\ (N-2-d)^3 \ / \ 6(N-1)^2\ (N-2)^2$$

8. Approximate Probability Distributions. We present now simple approximations for the probability distributions of the random variables R_{ni} $(n \geq 2)$ under the assumption that N is quite large compared to the d_i's. When this happens, that is $N >> d_i$ $(i = 1, 2, ---, N)$ the probability of having a directed edge between any ordered pair of points (which is $d_i/(N-1)$) is quite small compared to the probability of not

having a directed edge between an ordered pair of points (which is $1 - d_i/(N-1)$). Therefore, the probability that R(a) has a certain structure is proportional to the number of directed edges in its corresponding digraph $\Gamma(a)$. In other words, the probability that R(a) is in state s_i through those structures s_{ip} which have large number of directed edges, is quite small and may be ignored; we may approximate the probability $P[f(R(a))= s_i]$ by the sum of the probabilities $P[f(R(a))= s_{ip} \mid s_{ip}$ has smallest number of directed edges among $s_{i1}, s_{i2}, ---, s_{iI}.]$

Let $\overline{P}[f(R(a))= s_i]$ denote the approximations as described above, to the probabilities $P[f(R(a))= s_i]$. It is easy to see that

(i) $\overline{P}[f(R(a))= s_3] = P[\Gamma(a)$ consists of n directed edges in the form of an n-cycle]

(ii) $\overline{P}[f(R(a))= s_2] = P[\Gamma(a)$ consists of n-1 directed edges in the form of an (n-1)- step chain]

(iii) $$\overline{P}[f(R(a))= s_0] = \begin{cases} P[\Gamma(a) \text{ has no directed edges}], & n=2, \\ P[\Gamma(a) \text{ has no or one directed edge}], & n>2. \end{cases}$$

(iv) $\overline{P}[f(R(a))= s_1] = 1 - \Sigma\overline{P}[f(R(a))= s_i] \quad i = 0,2,3$

These probabilities can be computed easily for any n, and the approximate probability distributions obtained.

Using the factorial notation $\chi^{(\theta)}$ we find after some reductions and approximations:

(i) $\bar{P}[f(R(a))= s_3] = \dfrac{(n-1)!}{[(N-1)^{(n-1)}]^n} \prod_{k=1}^{n} d_{i_k}(N-1-d_{i_k})^{(n-2)}$,

general case

$= \dfrac{(n-1)!}{[(N-1)^{(n-1)}]^n} [d(N-1-d)^{(n-2)}]^n$,

restricted case

(ii) $\bar{P}[f(R(a))= s_2] =$

$= \dfrac{(n-1)!}{[(N-1)^{(n-1)}]^n} \sum_{j=1}^{n} [(N-1-d_{i_j})^{(n-1)} \prod_{\substack{k=1\\k\neq j}}^{n-1} d_{i_k}(N-1-d_{i_k})^{(n-2)}]$,

general case

$= \dfrac{n!}{[(N-1)^{(n-1)}]^n} (N-1-d)^{(n-1)} [d(N-1-d)^{(n-2)}]^{n-1}$,

restricted case

(iii) $\bar{P}[f(R(a))= s_0] =$

$= \dfrac{1}{[(N-1)^{(n-1)}]^n} \prod_{j=1}^{n} (N-1-d_{i_j})^{(n-1)}$

$+ \dfrac{n-1}{[(N-1)^{(n-1)}]^n} \sum_{j=1}^{n} [d_{i_j}(N-1-d_{i_j})^{(n-2)} \prod_{\substack{k=1\\k\neq j}}^{n} (N-1-d_{i_k})^{(n-1)}]$,

general case (n>2)

$= \dfrac{[(N-1-d)^{(n-1)}]^n}{[(N-1)^{(n-1)}]^n} + \dfrac{n(n-1)}{[(N-1)^{(n-1)}]^n} d(N-1-d)^{(n-2)}[(N-1-d)^{(n-1)}]^{n-1}$

restricted case (n>2)

(iv) $\bar{P}[f(R(a))= s_1]$ is obtained by difference, i.e.

$1 - \sum_{i=0,2,3} \bar{P}[f(R(a))= s_i]$

By investigating the forms of the approximate probabilities $P_{(m)i} = P[R_{ni} = m]$, we may propose approximations for the probability distributions of the random variables R_{ni}. The probabilities $P_{(m)i}$ are expressed in terms of the sums $S_{k,i}$ and the first term in $S_{k,i}$ (approximated values as on page 21) can be approximately written as

$$S_{ki}^{(1)} = \binom{N}{nk} \frac{(nk)!}{(n!)^k k!} \{\bar{P}[f(R(a)) = s_i]\}^k$$

This may be further approximated using the fact that N is quite large and d_i's are quite small ($N >> d_i$).

The other terms in $S_{k,i}$ ($i = 3,2,1,0$) are seen to be, at most, of order $0(\frac{1}{N})$; the number of terms in $S_{k,i}$ depends only on k, not on N. Thus for fixed k and d, the sum of these remaining terms tends to zero for increasing N, and we may approximate to $S_{k,i}$ by the approximate value of its first term.

With this approximate value, say $\tilde{S}_{k,i}$ of $S_{k,i}$ we obtain

$P_{(m),i} = P[R_{ni} = m] \sim \sum_{j=0}^{\binom{N}{n}-m} (-1)^j \binom{m+j}{m} \tilde{S}_{m+j,i}$ and then check to see if it is the general term of a well-known distribution.

For each of the random variables R_{ni} ($i=3,2,1,0$) we find that $P_{(m),i}$ can be approximated by a general term of a Poisson distribution and hence the following theorem:

Theorem 8.1 For large N, and $d<<N$

$$P[R_{ni} = m] \sim \frac{1}{m!} \theta_i^m \exp(-\theta_i), \quad i=3,2,1,0$$

where

$$\theta_3 = \frac{d^n}{n}\left(1 - \frac{d-1}{N - \frac{n+1}{2}}\right)^{n^2-2n}$$

$$\theta_2 = (N-1)d^{n-1}\left(1 - \frac{d}{N-\frac{n}{2}}\right)^{n-1}\left(1 - \frac{d-1}{N-\frac{n+1}{2}}\right)^{n^2-3n+2}$$

$$\theta_0 = \frac{N^n}{n!}\left(1 - \frac{d}{N-\frac{n}{2}}\right)^{n^2-n} + \frac{N^{n-1}}{(n-2)!}\, d\left(1 - \frac{d-1}{N-\frac{n+1}{2}}\right)^{n-2}\left(1-\frac{d}{N-\frac{n}{2}}\right)^{n^2-2n+1}$$

$$\theta_1 = \frac{N^n}{n!} - (\theta_3 + \theta_2 + \theta_0)$$

That is, the random variables R_{ni} are distributed approximately as Poisson distributions with parameters θ_i, $i=3,2,1,0$.

We may remark that for the random variable R_{n3} our result agrees with that of Katz [8] who obtained a Poisson approximation with parameter $\frac{d^2}{2}$ for the simple case $n=2$ for the distribution of the number of mutuals in a social group (R_{23} in our terminology).

9. <u>A simple approximation for expected values</u>. We have seen that the expected values of R_{ni} and R'_{ni} have rather involved expressions even for moderate values of n, and are computationally difficult to obtain. Consequently, simple approximations for $E(R_{ni})$ and $E(R'_{ni})$ would be convenient. The natural way seems to be to investigate their limiting values as N tends to infinity.

Let us suppose that the valencies d_i are simple linear functions of N, that is $d_i = \theta_i(N-1)$, $0 < \theta_i \leq 1$ $(i=1,2,---,N)$; $\theta_i = d_i | (N-1)$ is the ratio of the number of points with which P_i holds the relation R to the total number of points with which P_i may hold the relation R. If N be large enough so that all θ_i's are almost the same, or if all d_i's be the same (restricted case) so that all θ_i's are equal then this common value, say θ, is much more meaningful and denotes the `rate of valency´ for the relation R(A). Let $\theta = d/(N-1)$ for the restricted case, and $\theta = \sum_{i=1}^{N} d_i/N(N-1) = \bar{d}/(N-1)$ for the general case, where $\bar{d}$ is the `average valency´ ($= \sum_{i=1}^{N} d_i/N$) for a point in the set A.

Let $\overset{*}{E}(R_{ni}) = \lim_{N\to\infty} E(R_{ni}/\binom{N}{n})$, and $\overset{*}{E}(R'_{ni}) = \lim_{N\to\infty} E(R'_{ni})/\binom{N}{n}$

We obtain easily:

Theorem 9.1

$$\overset{*}{E}(R_{23} \equiv R'_{22}) = \theta^2$$

$$\overset{*}{E}(R_{22} \equiv R'_{21}) = 2\theta(1-\theta),$$

$$\overset{*}{E}(R_{20} \equiv R'_{20}) = 1 - 2\theta + \theta^2$$

Theorem 9.2

$$\overset{*}{E}(R_{33} \equiv R'_{36}) = \theta^3(\theta-2)(2\theta^2-2\theta-1),$$

$$\overset{*}{E}(R_{32}) = 6\theta^2(1-\theta)^2(1+\theta-\theta^2),$$

$$\overset{*}{E}(R_{31}) = 6\theta^2(1-\theta)^4$$

$$\overset{*}{E}(R_{30}) = (1-\theta)^4(1+4\theta-2\theta^2);$$

$$\overset{*}{E}(R'_{34}) = 6\theta^3(1-\theta)^2(2-\theta),$$

$$\overset{*}{E}(R'_{33}) = 6\theta^2(1-\theta)^3,$$

$$\overset{*}{E}(R'_{32}) = 9\theta^2(1-\theta)^4,$$

$$\overset{*}{E}(R'_{31}) = 6\theta(1-\theta)^5,$$

$$\overset{*}{E}(R'_{30}) = (1-\theta)^6$$

Theorem 9.3

$$\overset{*}{E}(R_{43}) = \theta^4(6 + 36\theta - 104\theta^2 + 24\theta^3 + 177\theta^4 - 252\theta^5 + 156\theta^6 - 48\theta^7 + 6\theta^8),$$

$$\overset{*}{E}(R_{42}) = 2\theta^3(1-\theta)^3(12 + 36\theta - 72\theta^2 - 16\theta^3 + 90\theta^4 - 57\theta^5 + 11\theta^6),$$

$$\overset{*}{E}(R_{41}) = 2\theta^3(1-\theta)^6(52 - 51\theta + 5\theta^3),$$

$$\overset{*}{E}(R_{40}) = (1-\theta)^6(1 + 6\theta + 21\theta^2 - 72\theta^3 + 78\theta^4 - 36\theta^5 + 6\theta^6),$$

$$\overset{*}{E}(R'_{4,12}) = \theta^4(6 + 36\theta - 104\theta^2 + 24\theta^3 + 177\theta^4 - 252\theta^5 + 156\theta^6 - 48\theta^7 + 6\theta^8) = \overset{*}{E}(R_{43}),$$

$$\overset{*}{E}(R'_{49}) = 8\theta^4(1-\theta)^3(6 - 13\theta^2 + 9\theta^3 - \theta^5),$$

$$\overset{*}{E}(R'_{48}) = 6\theta^5(1-\theta)^4(10 - 20\theta + 14\theta^2 - 3\theta^3),$$

$$\overset{*}{E}(R'_{47}) = 12\theta^4(1-\theta)^5(6 - \theta - 5\theta^2 + 3\theta^3),$$

$$\overset{*}{E}(R'_{46}) = 4\theta^3(1-\theta)^6(8 + 21\theta - 30\theta^2 + 11\theta^3),$$

$$\overset{*}{E}(R'_{45}) = 12\theta^3(1-\theta)^7(2 + 3\theta),$$

$$\overset{*}{E}(R'_{44}) = 3\theta^3(1-\theta)^8(32 + 3\theta),$$

$$\overset{*}{E}(R'_{43}) = 4\theta^2(1-\theta)^9(6 + 11\theta),$$

$$\overset{*}{E}(R'_{42}) = 42\theta^2(1-\theta)^{10},$$

$$\overset{*}{E}(R'_{41}) = 12\theta(1-\theta)^{11},$$

$$\overset{*}{E}(R'_{40}) = (1-\theta)^{12}$$

We give below some computational checks of these limiting (approximate) values of $E(R_{ni})$ and $E(R'_{ni})$ with their exact values, for the particular cases $N = 25$, $d = 3$, $n = 2,3$; and $\theta = 1/8 = 0.125$.

Table 9.1 Showing the approximate and exact values of $E(R_{ni})$ for n = 2, 3, and N = 25, d = 3.

	n = 2		n = 3	
	Exact value	Approximate value	Exact value	Approximate value
$E(R_{n3})$	4.7	4.7	10.0	10.3
$E(R_{n2})$	65.6	65.6	184.9	183.1
$E(R_{n1})$	-	-	111.8	126.4
$E(R_{n0})$	229.7	229.7	1993.3	1980.2
Total	300	300	2300	2300

Table 9.2 Showing the approximate and exact values of $E(R'_{ni})$ for n = 3, and N = 25, d = 3.

	Exact Value	Approximate value
$E(R'_{36})$	10.0	10.3
$E(R'_{34})$	35.1	38.7
$E(R'_{33})$	149.9	144.5
$E(R'_{32})$	180.2	189.6
$E(R'_{31})$	911.8	884.8
$E(R'_{30})$	1013.1	1032.2
Total	2300	2300

For θ = .1,.2,...,.9, the values of the ratios $\overset{*}{E}(R_{ni})$ and $\overset{*}{E}(R'_{ni})$ have been computed for n = 2,3, and 4; these values arranged in the form of tables are given in [2]. In applications these tables showing ratios $\overset{*}{E}(R_{ni})$ and $\overset{*}{E}(R'_{ni})$, and their corresponding graphs with θ as the independent variable may be used:

(i) to estimate the ratios $\overset{*}{E}(R_{ni})$ or $\overset{*}{E}(R'_{ni})$ for predetermined values of θ and n,

(ii) to find that value of θ (and hence, d) which for a predetermined n, will give a certain preassigned ratio $\overset{*}{E}(R_{ni})$ or $\overset{*}{E}(R'_{ni})$,

(iii) to find that value of n which for a fixed θ , will give a certain preassigned ratio $\overset{*}{E}(R_{ni})$ or $\overset{*}{E}(R'_{ni})$.

REFERENCES

[1] Anderson, T. W., and Goodman, L. A.: "Statistical Inference about Markov Chains," The Annals of Mathematical Statistics, Vol. 28, No. 1, March, 1957, pp. 89-110.

[2] Bhargava, T. N.: "A Stochastic Model for a Binary Dyadic Relation with Application to Group Dynamics," Ph.D. Thesis, 1962, Michigan State University.

[3] __________: "On Treatment of Group Dynamics as a Stochastic Process," submitted for publication.

[4] __________, and Katz, Leo: "A Stochastic Model for a Binary Dyadic Relation with Applications to Social and Biological Sciences," paper presented at the 34th Session of the International Statistical Institute, Ottawa, 1963.

[5] Feller, W.: An Introduction to Probability Theory and Its Applications, John Wiley & Sons, 1950.

[6] Fréchet, M.: "Les Probabilités Associeés à un Système d'Evènements Compatibles et Dependants," Actualités Scientifiques et Industrielles, Nos. 859 and 942, Hermann et Cie, Paris, 1940, 1943.

[7] Katz, L., and Wilson, T. R.: "The Variance of the Number of Mutual Choices in Sociometry," Psychometrika, Vol. 21, No. 3, September, 1956.

[8] Perron, D.: "Den Theorie der Matrizen"; Mathematische Annalen, Vol. 64, 1907, pp. 248-263.