NBER WORKING PAPER SERIES

# THE STATISTICAL ANALYSIS OF LOCAL STRUCTURE IN SOCIAL NETWORKS

Paul W. Holland*

and

Samuel Leinhardt**

Working Paper No. 44

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

June 1974

Preliminary: not for quotation



*NBER Computer Research Center. **Carnegie-Mellon University, School of Urban and Public Affairs. Research supported in part by National Science Foundation Grants GS-39778 to Carnegie-Mellon University and GJ-1154X2 to the National Bureau of Economic Research, Inc.

## Abstract

We introduce the concept of a triad census of a digraph and show how it can be used to enumerate various types of subgraph configurations. We give the basic probabilities needed for computing means and variances for a triad census under the U|MAN distribution for digraphs. These concepts are combined to provide a way of testing propositions about social structure using sociometric data. An application to 408 sociograms is given.

## Acknowledgements

We are grateful to James A. Davis, J. Richard Dietrich and Christopher Winship for their aid in the progress of this research.

## Contents



## Figures



## Tables

## 1. Introduction

Graph theory and network concepts are commonly used by social scientists to operationalize theoretical statements about structural regularities in social systems. These concepts are especially appropriate for models of the fine structure of interpersonal relations. Here the identification of individuals as "nodes" and interpersonal relations as "edges" in a graph is immediate. Graph-theoretic concepts are used to represent both models of social structure and data on social structure. A basic reference detailing the application of graph theory to models of social structure is Harary, Norman and Cartwright [1965]. The usual data representation, the sociogram, was introduced by Moreno [1934].

Theoretical statements incorporating network analogies have been formalized in mathematical models of social and perceptual behavior. For example, the cognitive balance theory of Heider [1944] was formalized by Cartwright and Harary [1956], Davis [1967] and Flament [1963]. Homans' [1950] propositions about behavior in groups were formalized by Davis and Leinhardt [1972] and Holland and Leinhardt [1971]. Radcliff-Brown's [1940] and Nadel's [1957] theoretical statements on kinship and role systems were formalized by White [1963] and White and Lorrain [1972]. We view these models as "global" in the sense that they imply that the entire organization of the system will fit into some relatively simple patterns.

Sociometric data have commonly been collected by investigators interested in small-scale social systems. In the years since the introduction of the technique, data have been compiled on a large variety of groups by numerous investigators (see, for example, Davis and Leinhardt [1972]). The complexity of this data has led to the development of various techniques and algorithms for organizing and simplifying them (see, for example, Moreno, et. al. [1960] and Boyle [1969]). Most analyses of sociometric data to social structure have focused on global properties such as cliques, status hierarchies and communication paths.

Although there is an abundance of theory, models, data and data analytic procedures, it seems fair to say that our understanding of small-scale social structure has not advanced much beyond the fundamental insights of Moreno, Heider and Homans. We believe that a major reason for this is the discrepancy between the local level at which the data are collected and the global level at which the models are conceptualized. We propose to bridge this gap by examining local structural properties which are expected to hold, on the average, across entire social systems. This approach permits the formalization and empirical study of propositions concerning the effect of social organization on individual perception and behavior. Such propositions are quite common in the sociological and social psychological literature. Davis [1963] presents a review of propositions about average local properties of social networks. He takes 56 major sociological and social psychological propositions from the writings of a variety of authors and restates them in graph-theoretic terms. He then shows that the propositions are statements about the consequences of local structure in interpersonal relations.

In the absence of a statistical methodology for testing empirical tendencies of local structure, even Davis' theoretical reformulation remains little more than an interesting exercise in formalism. Our purpose here is to propose such a methodology, (for an earlier statement of our approach see: (Holland and Leinhardt [1970].) Using this methodology a variety of propositions concerning local structure in networks of interpersonal relations can be modeled in graph-theoretic terms and tested by determining the discrepancy between empirically observed structure and that which would obtain by chance. While we emphasize the sociometric interpretation of graphs, this interpretation is not essential to the development or use of our methods.

In sections 2 through 5 we develop the technical material needed for our method. In section 6 the use of the method is described and some examples are discussed.

## 2. The Triad Census of a Digraph

### A. Some notation and preliminaries

For a detailed discusion of many concepts and results from graph theory, we refer the reader to Harary, Norman and Cartwright [1965]. However, we briefly discuss a few graph theoretic concepts which we use repeatedly.

Digraphs and Sociomatrices: The basic mathematical entity that concerns us is the binary directed graph, or digraph. A digraph is a set of "nodes" and a set of directed "lines" or "edges" connecting pairs of nodes. In sociometric choice data, the nodes correspond to the individual group members, and there is a directed line connecting node i to node j if and only if person i chooses person j according to the sociometric choice criterion employed. The adjective "binary" in the definition of a digraph refers to the added restriction that we only consider whether or not a line connects i to j so that we do not allow for the possibility that choices may have "strengths" attached to them or that they may be of different types. Thus the digraph is the mathematical representation of the simplest form of sociometric choice data -- unranked choices on a single criterion.

We denote the number of nodes in a digraph by g -- the group size. If the nodes are numbered in some arbitrary way from 1 to g, then we may create a useful matrix representation of the digraph as follows. Let $\underset{\sim}{X}$ be a g by g matrix whose (i,j) entry is:

$$X_{ij} = \begin{cases} 1 \text{ if } i \to j \\ 0 \text{ otherwise} \end{cases} \qquad (2\text{-}1)$$

Note that we use $i \to j$ to mean that there is a directed line from node i to node j-- there may also be a directed line from node j to node i, but this possibility is neither implied nor denied by the notation $i \to j$. The matrix $\underset{\sim}{X}$ is called the adjacency matrix in graph theory and the sociomatrix in sociometry. We use the latter term. The main diagonal of $\underset{\sim}{X}$ corresponds to self-choice in the sociometric context, and for many reasons it is convenient for us to assume that $X_{ii}=0$. Because self-choice is often disallowed in sociometric data, this restriction is generally not significant. There are some applications of digraphs which allow self-choice (especially if the nodes correspond to groups of people rather than to individuals), but we ignore this possibility here. The sociomatrix, $\underset{\sim}{X}$, is not exactly the same thing as the digraph because it implies that the nodes have been labeled from 1 to g in some arbitrary way. Strictly speaking, each sociomatrix corresponds to a labeled digraph, whereas two different sociomatrices can represent the same unlabeled digraph. In this latter situation the two sociomatrices will only differ by a simultaneous row-column permutation. There correspond to any unlabeled digraph a certain number of labeled digraphs. These are called the labeled versions of the digraph.

Subgraphs, Dyads and Triads: Fundamental to the methods we discuss is the notion of a subgraph of a digraph. If we delete some of the nodes and all the lines in a digraph that either go to or come from the deleted nodes, the resulting entity is a subgraph of the digraph. If we delete all but $k$ of the nodes, then we shall call the result a $k$ - subgraph. Since a $k$-subgraph is also a digraph with $k$ nodes, it may be represented by a $k$ by $k$ sociomatrix, and it has a certain number of labeled versions. There are $\binom{g}{k}$ k-subgraphs in a digraph with g nodes.

Two digraphs with g nodes are said to be isomorphic if they are the same digraph in the sense that they can both be represented by the same sociomatrix. This means that if digraph 1 is represented by sociomatrix $\underset{\sim}{X}$ and digraph 2 by $\underset{\sim}{Y}$ then there is a row-column permutation of $\underset{\sim}{Y}$ call it $\underset{\sim}{Z}$, such that, as matrices, $\underset{\sim}{X}=\underset{\sim}{Z}$. The notion of isomorphism partitions digraphs into isomorphism classes. For example, when g=2, there are 3 isomorphism classes or dyads as illustrated in Figure 2-1. We adopt the names null, asymmetric and mutual to describe the three dyad types.

Figure 2-1 goes about here

When g=3, there are 16 isomorphism classes or triads. These are illustrated and named in Figure 2-2. We have adopted the triad naming convention of Holland and Leinhardt [1970], which is described in the legend of Figure 2-2.

Figure 2-2 goes about here

When $g \geq 4$, the number of isomorphism classes of digraphs grows very rapidly -- for g=4 there are 218, and for g=5 there are 9608 (see Harary [1955]). We refer to the isomorphism classes of digraphs with g nodes as the digraph "types".

B. The Triad Census

In a digraph there are $\binom{g}{3}$ distinct 3-subgraphs formed by selecting each of the possible subsets of 3 nodes and their corresponding lines. These subgraphs can be classified by their isomorphism type. Let $T_u$ denote the number of these 3-subgraphs of isomorphism type u (where u ranges over the 16 triad names given in Figure 2-2). The triad census $\underset{\sim}{T}$ is the vector of 16 components given by:

$$\underset{\sim}{T} = (T_{003}, T_{012}, \ldots, T_{300}). \tag{2-2}$$

We have adopted the following ordering of the components of $\underset{\sim}{T}$ to simplify communication: 003, 012, 102, 021D, 021U, 021C, 111D, 111U, 030T, 030C, 201, 120D, 120U, 120C, 210, 300.

A triad census may be regarded as a way of reducing the entire sociomatrix $\underset{\sim}{X}$ to a smaller set of 16 summary statistics. When $g \geq 5$

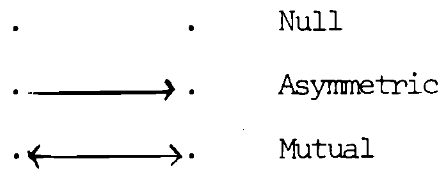
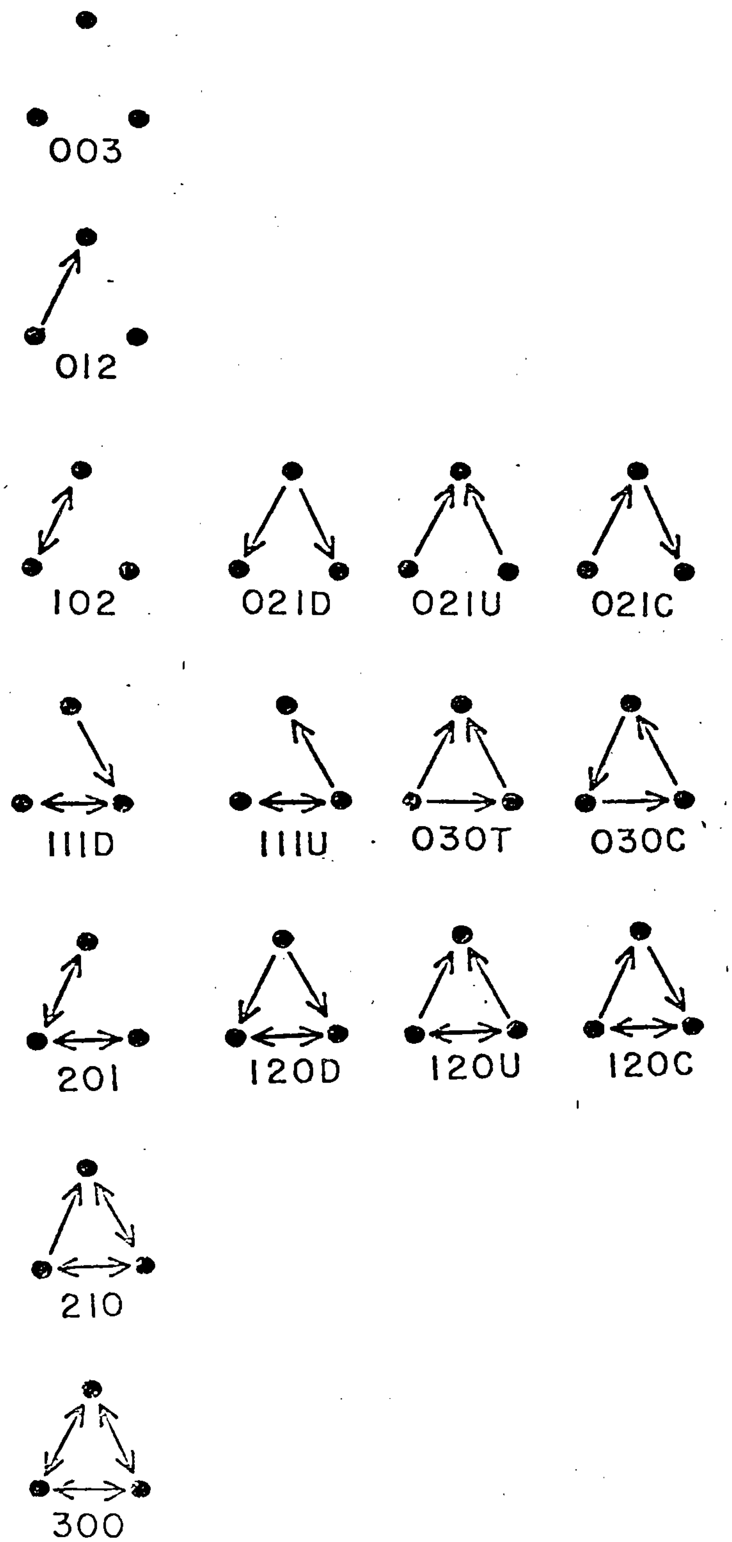



Figure 2-1: The 3 Isomorphism Classes for Digraphs with g=2 (i.e., the Dyad Types).

Figure 2-2: The 16 Isomorphism Classes for digraphs with g=3 (i.e., the Triad Types). Triad naming convention: first digit=number of mutual dyads; second digit=number of asymmetric dyads; third digit=number of null dyads; trailing letters further differentiate among triad types.



this is a real reduction in information since $\underset{\sim}{X}$ contains more than 16 elements of data. In general, knowing the triad census of a digraph does not uniquely determine the digraph. However, as we show in section 4, $\underset{\sim}{T}$ contains a surprising amount of useful information about $\underset{\sim}{X}$.

A triad census is a special case of the more general concept of a $\underline{k}$-subgraph census. Thus we might also consider a dyad census, a tetrad census or even a pentad census. We show in section 4 that a dyad census can be computed from a triad census. Furthermore there are so many tetrad types (218) and pentad types (9608) that a $\underline{k}$-subgraph census for $\underline{k} \geq 4$ is often more cumbersome than the original sociomatrix. Thus the triad census occupies an important position among $\underline{k}$-subgraph censuses in that it is both manageable and yet contains a substantial amount of useful information.

For various random digraphs, the first and second moments of a triad census are readily computed. These computations are illustrated in section 5 for a particular random digraph (Corollary 2). This permits us to test propositions about average local structure in specific digraphs using the triad census (see section 6 ).

Because a triad census $\underset{\sim}{T}$ is a summation over all 3-subgraphs, the information in $\underset{\sim}{T}$ is relatively stable and is not significantly affected by a few changes in the lines of the digraph. For this reason, reduction to the triad census is not as affected by sociometric measurement error or masking (see Holland and Leinhardt [1973]) as are methods which focus on specific linkages and individual nodes. This property of the triad census has both advantages and disadvantages. On the one hand, conclusions drawn from $\underset{\sim}{T}$ are likely to be relatively robust, but on the other, if interest focuses on a few specific nodes or lines, the triad census may not be useful.

3. Configurations in Digraphs

A. Reasons for considering configurations

Davis and Leinhardt [1972] present a procedure in which a version of the triad census is used to test a model of small-scale social structure. Distilling a set of structural propositions from Homans [1950], they examined the triad types to determine which ones were logically inconsistent with the propositions. They felt that, formalized in this fashion, a fair interpretation of Homans was that the inconsistent triads would be empirically rare. They then showed that a social structure combining clusters, Davis [1967], with a transitive hierarchy, French [1956], was implied by the propositions. Their empirical analysis of sociometric data lent some support to the model, as did a latter analysis by Davis [1970]. However, a re-examination of this model revealed that it and most other structural models of interpersonal affect assumed that the affective choice relation is transitive, Holland and Leinhardt [1971]. The models themselves could be expressed as transitive graphs (t-graphs) upon which some additional constraints had been applied. Transitivity, however, is not a characteristic of a triad. Rather, it is a property of a lower-order configuration which is contained in varying degree by some triad types. For example, assume the triad type 300 represents the sociogram of a three-person group. (Refer to Figure 2-2 for triad names.) It is transitive from the point of view of each member. That is, the transitive condition, $i \to j$, $j \to k$, $i \to k$, holds for all 6 permutations of i, j, k. On the other hand, if one examines the 120C triad type, the transitivity condition holds only once while it is contradicted, i.e., $i \to j$, $j \to k$ <u>and</u> <u>not</u> $i \to k$, twice. Thus, if Homans' propositions are in fact statements about the propensity for affective choice to tend toward transitivity and avoid intransitivity, then a failure to recognize the complex nature of triads may yield erroneous conclusions. Indeed, our reanalysis of Davis' [1970] results yielded a high level of confirmation for the t-graph model (see Holland and Leinhardt [1971]).

It is useful to pursue the notion of transitivity in interpersonal affect because it demonstrates the utility of thinking of structural propositions in terms of configurations of social relations. Consider, for example, the statement by Heider [1957] in regard to positive interpersonal sentiment in triadic situations: "In the p-o-x triad, the case of three positive relations may be considered psychologically...transitive" (p.206). For Heider, transitivity defines cognitive balance in affective situations. His examples of balanced psychological structures are configurational in that they emphasize transitivity from the point of view of p, the perceiver, and not necessarily from the points of view of the other entities.

In an alternative to the t-graph model, Mazur [1971] put forth another structural proposition. He argued that interpersonal affect data could be explained by the proposition: "Friends are likely to agree, and unlikely to disagree; close friends are very likely to agree, and very unlikely to disagree" (p. 308). If one assumes that an asymmetric pair represents "friends" while a mutual pair represents "close friends" then Mazur's proposition amounts to a set of hypotheses concerning the relative empirical frequency of simultaneous choice of individuals by "friends" and "close friends". Since the choices of the third individual are not at issue, Mazur's proposition is not about triads but refers instead to configurations contained in triads.

An examination of Davis [1963] reveals numerous other structural propositions that can be formalized in terms of configurations. The concepts of cross-pressures (Berelson, et al., [1954]), homophily-heterophily (Lazarsfeld and Merton [1954]), structural balance (Cartwright and Harary [1956]), distributive justice (Homans [1961]), cliques (Homans [1950]; Lazarsfeld and Merton [1954]; Lipset, et al. [1956]), innovation in ideas (Katz and Lazarsfeld [1955]), attitude change (Homans [1950]) and conflict (Coleman [1957]), as reformulated by Davis, are all propositions about local structure and may be described by configurations within a digraph.

B. Structure of Configurations

Configurations and subgraphs are similar except that in a configuration only some of the lines between a subset of nodes are of interest. We shall try to make this vague idea more precise. In section 2 we introduced the concepts of labeled and unlabeled subgraphs of a digraph, and the sociomatrix, $\underset{\sim}{X}$. For any 3-subgraph we may construct a 3 by 3 sociomatrix such as the one given in (3-1):

$$\begin{bmatrix} - & X_{ij} & X_{ik} \\ X_{ji} & - & X_{jk} \\ X_{ki} & X_{kj} & - \end{bmatrix} \tag{3-1}$$

where i, j and k are 3 distrinct nodes. There are $2^6=64$ possible zero-one sociomatrices of the form (3-1), and they correspond to the set of all possible labeled 3-node digraphs. These in turn are the labeled versions of the 16 non-isomorphic unlabeled 3-node digraphs. It is easier to illustrate how to construct a configuration than to give a precise definition of this concept. We begin with the pairs of subscripts that appear on the entries of (3-1). The first step is to select a subset of these ordered pairs of subscripts. For example, to define the configuration that corresponds to "intransitivity", we select these three ordered pairs: ij, jk, ik. The order of the pairs indicates the direction of the relation. Next, associate a zero or a one with each ordered pair of subscripts that has been selected. For example, for intransitivity we associate a 1 with ij, another 1 with jk and a 0 with ik. These correspond to: $i \to j$, $j \to k$ <u>and</u> <u>not</u> $i \to k$. For convenience these items can be arranged in a matrix as follows:

$$\begin{pmatrix} ij & jk & ik \\ 1 & 1 & 0 \end{pmatrix} . \tag{3-2}$$

The first row of the matrix that describes a configuration is the "reading rule" for the configuration. The second row defines the configuration's type. Two different configurations can have the same reading rule but be of different types. For example, the configuration that corresponds to "transitivity" has the same reading rule as (3-2) but a different type. It is given by:

$$\begin{pmatrix} ij & jk & ik \\ 1 & 1 & 1 \end{pmatrix} . \qquad (3\text{-}3)$$

It is convenient to have a picture for a configuration. We adopt the following conventions for drawing them. For an ordered pair of nodes, ab: (i) if ab is not in the reading rule, then no arrow is drawn from a to b; (ii) if ab is in the reading rule and has a 1 associated with it, then a solid arrow is drawn from a to b; (iii) if ab is in the reading rule and has a 0 associated with it, then a dashed arrow is drawn from a to b. For example, the matrix given in (3-2) can be represented by Figure 3-1.

Figure 3-1 goes about here

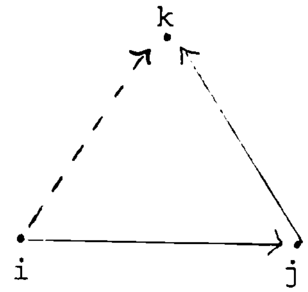
Another example of a configuration comes from Mazur's proposition about agreement among friends. The reading rule for configurations which portray situations in which friends agree or disagree in their choices of a third individual is: ij, ji, ik, jk. If close friends are pairs of individuals who choose one another then disagreement among close friends can be represented by the following configuration matrix:

$$\begin{pmatrix} ij & ji & ik & jk \\ 1 & 1 & 1 & 0 \end{pmatrix} \qquad (3\text{-}4)$$

It is obvious that this matrix is equivalent to the matrix:

$$\begin{pmatrix} ij & ji & ik & jk \\ 1 & 1 & 0 & 1 \end{pmatrix} . \qquad (3\text{-}5)$$

Figure 3-1: Pictorial Representation of the Configuration Defined by (3-2).



The pictorial representation of (3-4) and (3-5) is given in Figure 3-2.

Figure 3-2 goes about here

While we have only discussed configurations that involve 3 nodes, i.e., 3-configurations, it is clear that this is not an essential part of the concept. In general, one may have reading rules that involve k distinct nodes, and these result in various types of k-configurations. If a 3-configuration has a reading rule that involves all 6 of the ordered pairs ij, ji, ik, ki, jk, kj, then it is equivalent to a triad.

The main reason for considering configurations is that they are a more refined set of concepts than the triads. A single triad may contain many different configurations. Furthermore, many sociological and social psychological propositions about networks make predictions about configurations rather than about triads. However, in the next section we show that the triad census can be used to enumerate all 3-configurations. This, along with other known properties of the triad census, is what makes it a basic tool.

4. Linear Combinations of Triad Frequencies

Once we consider a vector such as the 16 triad frequencies that make up a triad census, $\underset{\sim}{T}$, it is natural to consider any linear combination of the elements of the vector. We use the vector notation:

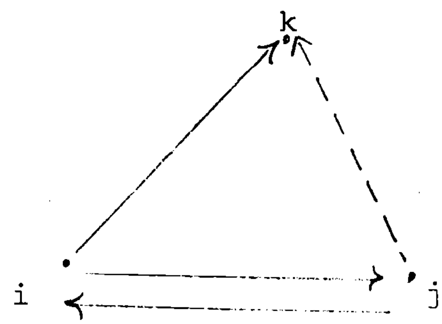
$$\underset{\sim}{\ell}'\underset{\sim}{T} = \sum_u \ell_u T_u \qquad (4\text{-}1)$$

where u is always assumed to run over the triad types given in Figure 2-2.

A. Dyads from Triads

A simple property of a triad census $\underset{\sim}{T}$ is that it determines how many nodes there are in the digraph. This is because the total number of triads in a digraph with g nodes is $\binom{g}{3}$ and furthermore if $\underset{\sim}{e}' = (1, 1, \ldots, 1)$

Figure 3-2: Pictorial Representation of Configurations Given by (3-4).



$$\underset{\sim}{e}' \underset{\sim}{T} = \sum_u T_u = \binom{g}{3} . \tag{4-2}$$

Thus, to find g we merely compute $\binom{g}{3}$ from (4-2) and solve for g. Equation (4-2) is a simple and yet important example of a linear combination of triad frequencies. Let M, A and N denote the number of mutual, asymmetric and null dyads in the digraph. Furthermore, let $m_u$, $a_u$, $n_u$ denote the number of mutual, asymmetric and null dyads that are contained in a triad of type u. Since every dyad in a digraph is contained in exactly g-2 triads, it is easy to see that

$$\underset{\sim}{m}' \underset{\sim}{T} = \sum_u m_u T_u = (g-2) M, \tag{4-3}$$

$$\underset{\sim}{a}' \underset{\sim}{T} = \sum_u a_u T_u = (g-2) A, \tag{4-4}$$

and

$$\underset{\sim}{n}' \underset{\sim}{T} = \sum_u n_u T_u = (g-2) N. \tag{4-5}$$

If we regard g as given, then M, A and N can be written as linear combinations of the triad frequencies where the weights are $m_u/(g-2)$, $a_u/(g-2)$, and $n_u/(g-2)$. The weights used for enumerating M, A and N are given in Table 4-1.

Since the total number of directed lines (or "choices"), C, is given by

$$C = 2M+A, \tag{4-6}$$

it is clear that C, too, can be expressed as a linear combination of the triad frequencies. The weights for C are given in Table 4-1.

## B. In- and Out-degrees from Triads

The row and column sums of a sociomatrix $\underset{\sim}{X}$ will be denoted by $\{X_{i+}\}$ and $\{X_{+j}\}$, respectively. In the sociometric context, $X_{i+}$ is the number of choices made by individual i, while $X_{+j}$ is the number of choices received by individual j. $X_{i+}$ and $X_{+i}$ are also called the out-degree and the in-degree, respectively, of node i. In this notation we have C given in (4-6) also given by

$$C = X_{++} = \sum_i X_{i+} = \sum_j X_{+j} \quad . \tag{4-7}$$

It is sometimes convenient to summarize a set of numbers by their mean and variance. We now do this to the two sets of "degrees", $\{X_{i+}\}$ and $\{X_{+j}\}$. From (4-7) we see that the mean in-degree and the mean out-degree are the same and are given by

$$\overline{X} = C/g \quad . \tag{4-8}$$

Since C can be expressed as a linear combination of triad frequencies it follows that $\overline{X}$ can also be so expressed. What is more interesting is the fact that the variances of the in-degrees and of the out-degrees can both be determined by C and linear combinations of the triad frequencies. Because this fact is useful to us and not widely known we include a proof of it now. We denote the variance of the $\{X_{i+}\}$ by $S^2_{out}$ and of the $\{X_{+j}\}$ by $S^2_{in}$ where

$$S^2_{out} = (1/g) \sum_i (X_{i+} - \overline{X})^2 \tag{4- 9}$$

and

$$S^2_{in} = (1/g) \sum_j (X_{+j} - \overline{X})^2 \quad . \tag{4-10}$$

The proof is divided into two parts given in Lemmas 1 and 2 below.

Lemma 1: Let $B_{in} = \sum_i \sum_{j<k} X_{ji} X_{ki}$ and $B_{out} = \sum_i \sum_{j<k} X_{ij} X_{ik}$ then

(a)
$$S^2_{in} = (2/g)\, B_{in} - \overline{X}\, (\overline{X} - 1) \tag{4-11}$$

(b)
$$S^2_{out} = (2/g)\, B_{out} - \overline{X}\, (\overline{X} - 1) \tag{4-12}$$

where $\overline{X}$ is defined in (4-8) and (4-7).

Proof: Since the proofs of (a) and (b) are nearly identical, we only prove (a).

$$S^2_{in} = (1/g) \sum_i (X_{+i} - \overline{X})^2 = (1/g) \sum_i (X_{+i})^2 - (\overline{X})^2 \quad . \tag{4-13}$$

But

$$\sum_i (X_{+i})^2 = \sum_{i,j,k} X_{ji} X_{ki}$$

$$= \sum_i \sum_{j<k} X_{ji} X_{ki} + \sum_i \sum_{k<j} X_{ji} X_{ki} + \sum_i \sum_j X^2_{ji}$$

$$= 2 \sum_i \sum_{j<k} X_{ji} X_{ki} + \sum_i \sum_j X_{ji} \quad . \tag{4-14}$$

In (4-14) we made use of the fact that

$$X^2_{ij} = X_{ij} \tag{4-15}$$

Hence we have

$$(1/g) \sum_i (X_{+i})^2 = (2/g)\, B_{in} + \overline{X} \quad . \tag{4-16}$$

Putting (4-16) into (4-13) completes the proof.||

The next lemma shows that $B_{in}$ and $B_{out}$ may both be expressed as linear combinations of the triad frequencies so that together with Lemma 1 it proves that $S^2_{in}$ and $S^2_{out}$ are determined by the triad frequencies. Before we state Lemma 2 we need to define two special configurations which we shall call $B_{in}$ and $B_{out}$ configurations.

A $B_{in}$ configuration is defined by the matrix

$$\begin{pmatrix} ji & ki \\ 1 & 1 \end{pmatrix}, \tag{4-17}$$

while a $B_{out}$ configuration is defined by the matrix

$$\begin{pmatrix} ij & ik \\ 1 & 1 \end{pmatrix}. \tag{4-18}$$

In Figure 4-1 we illustrate $B_{in}$ and $B_{out}$ configurations using the conventions of Figure 3-1 and 3-2.

<u>Figure 4-1 goes here</u>

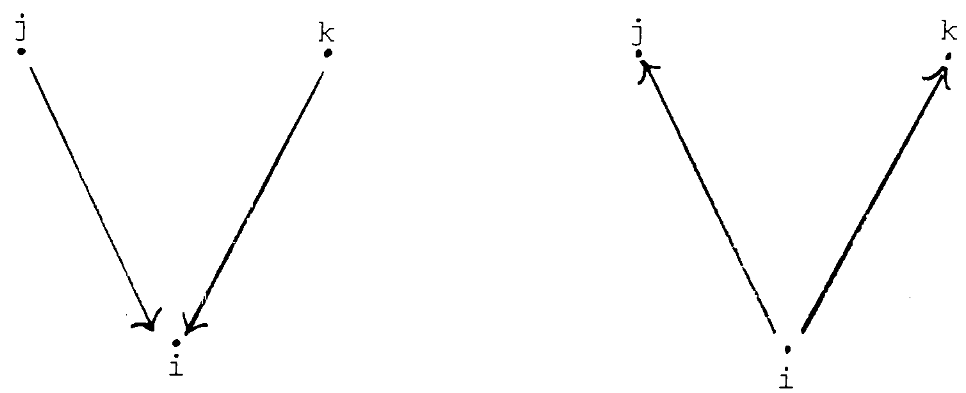
It is easy to see that $B_{in}$ gives the number of $B_{in}$ configurations and that $B_{out}$ gives the number of $B_{out}$ configurations. Lemma 2 shows that $B_{in}$ and $B_{out}$ can also be computed from the triad census.

<u>Lemma</u> 2: $B_{in}$ <u>and</u> $B_{out}$ <u>are given by</u>

(a) $$B_{in} = \sum_u b_{in,u} T_u \tag{4-19}$$

(b) $$B_{out} = \sum_u b_{out,u} T_u \tag{4-20}$$

<u>where</u> $b_{in,u}$ <u>and</u> $b_{out,u}$ <u>are given in Table</u> 4-1.




$B_{in}$ configuration　　$B_{out}$ configuration

Figure 4-1: Pictorial Representation of $B_{in}$ and $B_{out}$ Configurations.

Proof: Again we only prove (a). First we note that every $B_{in}$ configuration in the graph is contained in exactly one 3-subgraph. Hence to count $B_{in}$ configurations, we only need to weight each triad frequency by the number of $B_{in}$ configurations contained in that triad type. To finish the proof, we observe that these weights are exactly the values of $b_{in,u}$ given in Table 4-1. ||

C. Counting Configurations from Triads

The proof of Lemma 2 makes use of the fact that the number of configurations of certain types can be counted using only the triad census, $\underset{\sim}{T}$. This leads us to consider counting all of the possible types of 3-configurations using only the triad census. Evidently, this requires no new ideas since any configuration that involves 3 nodes is contained in exactly one 3-subgraph. Thus, to count the number of 3-configurations of a given type that arise in a digraph, we need only find out how many of these are in each of the 16 triad types and then sum the correspondingly weighted triad frequencies. If a configuration only involves two nodes, the same rule applies except that in order to take care of the fact that each 2-configuration is contained in g-2 triads, we must divide the weighted sum of triad frequencies by g-2. This is why the factor g-2 arises when we count dyad types from the triad frequencies.

As an example, suppose we wish to find the number of intransitive configurations in a given digraph. An intransitive configuration is a 3-configuration and can therefore be enumerated by the above rule. The weights used for counting intransitive as well as transitive configurations from $\underset{\sim}{T}$ are given in Table 4-1.

As a final example of counting a 3-configuration using $\underset{\sim}{T}$ we consider the one given in (3-3), (3-4) and in Figure 3-2. Each weight given in Table 4-1 for this configuration (the column marked "close friends disagreeing") is found by counting the number of mutual dyads in the given triad type for which only one member of the mutual pair chooses the other triad member.

<u>Table 4-1 goes about here</u>

The discussion of sections 4A and 4B shows why the linear combinations of triad frequencies are important. They give additional information about the graph, information which may not be obvious from the triad census itself, but which is implied by it. Thus the set of <u>all</u> linear combinations of the triad frequencies is a natural extension of the triad census. While not all linear combinations of triad frequencies have interpretations in terms of counting configurations, many do and the concept of configurations is a key to understanding the set of all linear combinations of the triad frequencies.

## 5. <u>The Distribution of a Random Subgraph Census</u>

Various notions of "random" digraphs have statistical utility in the analysis of sociometric data (see Holland and Leinhardt [1970] and the appendix to this paper for further discussion). When we obtain a triad census from an actual sociomatrix, it is also useful to know what we would expect the triad census to be from a random digraph. While the exact probability distribution of a triad census from a random digraph is very complicated, it is also true that when there is a sufficient number of nodes

Table 4-1: Selected Weighting Vectors.

| TRIAD TYPES | $e$ | $m_u$ | $a_u$ | $n_u$ | $c_u$ | $b_{in,u}$ | $b_{out,u}$ | Transitivity | Intransitivity | Close Friends Disagreeing |
|---|---|---|---|---|---|---|---|---|---|---|
| 003 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 012 | 1 | 0 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| 102 | 1 | 1 | 0 | 2 | 2 | 0 | 0 | 0 | 0 | 0 |
| 021D | 1 | 0 | 2 | 1 | 2 | 0 | 1 | 0 | 0 | 0 |
| 021U | 1 | 0 | 2 | 1 | 2 | 1 | 0 | 0 | 0 | 0 |
| 021C | 1 | 0 | 2 | 1 | 2 | 0 | 0 | 0 | 1 | 0 |
| 111D | 1 | 1 | 1 | 1 | 3 | 1 | 0 | 0 | 1 | 0 |
| 111U | 1 | 1 | 1 | 1 | 3 | 0 | 1 | 0 | 1 | 1 |
| 030T | 1 | 0 | 3 | 0 | 3 | 1 | 1 | 1 | 0 | 0 |
| 030C | 1 | 0 | 3 | 0 | 3 | 0 | 0 | 0 | 3 | 0 |
| 201 | 1 | 2 | 0 | 1 | 4 | 1 | 1 | 0 | 2 | 2 |
| 120D | 1 | 1 | 2 | 0 | 4 | 2 | 1 | 2 | 0 | 0 |
| 120U | 1 | 1 | 2 | 0 | 4 | 1 | 2 | 2 | 0 | 0 |
| 120C | 1 | 1 | 2 | 0 | 4 | 1 | 1 | 1 | 2 | 1 |
| 210 | 1 | 2 | 1 | 0 | 5 | 2 | 2 | 3 | 1 | 1 |
| 300 | 1 | 3 | 0 | 0 | 6 | 3 | 3 | 6 | 0 | 0 |
| $\sum_u \ell_u T_u$ | $\binom{g}{3}$ | $(g-2)M$ | $(g-2)A$ | $(g-2)N$ | $(g-2)C$ | $B_{in}$ | $B_{out}$ | | | |

(say, at least 10) and when certain other nondegeneracies obtain (such as a sufficient number of dyads of each type), a random triad census has an approximate multivariate normal distribution. ( See Holland and Leinhardt [1970] for some simulation results.) In this section, we give formulas for the means, variances and covariances of the components of $\underset{\sim}{T}$ for a very general class of random digraphs. We also give special results for a particular class of random digraphs which we have found useful.

A. Moments of a Subgraph Census

We begin by considering the number of k-subgraphs of a given digraph that are of a particular isomorphism class. We justify this extra generality on the grounds that it entails no additional difficulties and that the general results for k-subgraphs may be more useful for related problems. We specialize to triads in section 5B.

Let u and v denote two isomorphism classes for k-subgraphs. (In the triad case, the range of values for u and v are given in Figure 2-2.) For k-subgraphs where $k \geq 4$, the possibilities are very numerous -- some are given in Moon [1968].

In Theorem 1, below, we give formulas for the means, variances and covariances for the number of k-subgraphs of types u and v for a random digraph generated by a completely general stochastic mechanism. In Corollary 1 we specialize this result to the case of a triad census, and in Corollary 2 we specialize even further to the moments of a triad census where the random digraph is of the special variety used in Holland and Leinhardt [1970].

We need some notation. Let K and L be subscripts that refer to the $\binom{g}{k}$ distinct k-subgraphs of a given digraph. Thus, we shall speak of K being of a particular isomorphism class, etc.

We next define the indicator variables $H_K(u)$ by:

$$H_K(u) = \begin{cases} 1 \text{ if } K \text{ is of isomorphism class } u \\ 0 \text{ otherwise .} \end{cases} \tag{5-1}$$

The number of k-subgraphs of the given digraph that are of type u, $H_u$, is then given by the sum over K of $H_K(u)$, i.e. ,

$$H_u = \sum_K H_K(u) \quad . \tag{5-2}$$

Since we provide formulas that hold for a wide class of random digraphs, we need a notation for various probabilities that arise in the calculations and which must be computed explicitly for any particular application of the general results. Thus we define

$$p_K(u) = P \left\{K \text{ is of type } u\right\} = P \left\{H_K(u) = 1\right\} \tag{5-3}$$

$$\begin{aligned} p_{K,L}(u,v) &= P \left\{K \text{ is of type } u \text{ and } L \text{ is of type } v\right\} \\ &= P \left\{H_K(u) = 1 \text{ and } H_L(v) = 1\right\} \quad . \end{aligned} \tag{5-4}$$

We also need a notation for certain average probabilities which can be computed from $p_K(u)$ and $p_{K,L}(u,v)$. The first of these is easy since it is just the average value of $p_K(u)$ over all values of K,i.e.,

$$\bar{p}(u) = \frac{1}{\binom{g}{k}} \sum_K p_K(u) \quad . \tag{5-5}$$

In order to define the other average probabilities that arise, we need to examine the relationship between two k-subgraphs more carefully.

Let

$|K \cap L|$ = the number of nodes that K and L have in common. (5-6)

The possible values for $|K \cap L|$ are $0,1,\ldots,k$. When $|K \cap L| = 0$, K and L are disjoint, while when $|K \cap L| = k$, K and L are identical. In general, there are

$$\binom{g}{k}\binom{g-k}{k-j}\binom{k}{j} \tag{5-7}$$

pairs of k-subgraphs of a digraph on g nodes for which $|K \cap L| = j$. Now we define the average value of $p_{K,L}(u,v)$ over all choices of K and L such that $|K \cap L|=j$ as

$$\bar{p}_j(u,v) = \frac{1}{\binom{g}{k}\binom{g-k}{k-j}\binom{k}{j}} \sum_{|K \cap L|=j} p_{K,L}(u,v). \tag{5-8}$$

It should be emphasized that the various probabilities defined in (5-3), (5-4), (5-5), and (5-8) must be calculated explicitly in order to apply the general results to particular cases. Thus, we use Theorem 1 to specify what probabilities must be computed and combined to get the moments of a subgraph census.

Theorem 1: Using the notation given above and assuming that a random digraph is generated by some specific but completely general stochastic mechanism the first and second moments of $H_u$ defined in (5-2) are given by:

(a) $E(H_u) = \binom{g}{k}\bar{p}(u)$ (5-9)

(b) $\mathrm{Var}(H_u) = \binom{g}{k}\bar{p}\ (u)\ (1-\bar{p}(u)) + \binom{g}{k}\sum_{j=0}^{k-1}\binom{g-k}{k-j}\binom{k}{j}[\bar{p}_j(u,u)-(\bar{p}(u))^2]$ (5-10)

(c) $\mathrm{Cov}(H_u,H_v) = -\binom{g}{k}\bar{p}(u)\bar{p}(v) + \binom{g}{k}\sum_{j=0}^{k-1}\binom{g-k}{k-j}\binom{k}{j}[\bar{p}_j(u,v) - \bar{p}(u)\bar{p}(v)]$ (5-11)

<u>Proof</u>: First we prove (a).

$$E(H_u) = E(\sum_K H_K(u)) = \sum_K E(H_K(u)) ,$$

but $H_K(u)$ is an indicator variable so that its expected value is merely its probability of being one, thus

$$E(H_u) = \sum_K P\left\{H_K(u) = 1\right\} = \binom{g}{k}\bar{p}\ (u) ,$$

proving part (a).

Since formula (b) is the special case of formula (c) for u=v, we only prove (c).

$$\mathrm{Cov}\ (H_u, H_v) = \mathrm{Cov}\ (\sum_K H_K(u), \sum_L H_L\ (v))$$

$$= \sum_{K,L} \mathrm{Cov}\ (H_K(u), H_L(v)).$$

But $\mathrm{Cov}\ (H_K(u), H_L(v)) =$

$$P\left\{H_K(u) = 1 \text{ and } H_L(v) = 1\right\} - P\left\{H_K(u) = 1\right\} P\left\{H_L(v) = 1\right\}$$

$$= p_{K,L}(u,v) - p_K\ (u)\ p_L(v),$$

thus

$$\mathrm{Cov}\ (H_u,H_v) = \sum_{K,L} p_{K,L}\ (u,v) - \sum_{K,L} p_K(u)\ p_L(v)$$

$$= \sum_{j=0}^{k} \sum_{|K \cap L|=j} p_{K,L}(u,v) - \left[\binom{g}{k} \bar{p}(u)\right]\left[\binom{g}{k} \bar{p}(v)\right]$$

$$= \sum_{j=0}^{k} \binom{g}{k}\binom{g-k}{k-j}\binom{k}{j} \bar{p}_j(u,v) - \binom{g}{k}^2 \bar{p}(u)\bar{p}(v) \quad . \tag{5-12}$$

We now use the following fact about binomial coefficients:

$$\sum_{j=0}^{k} \binom{g-k}{k-j}\binom{k}{j} = \binom{g}{k} \tag{5-13}$$

Using (5-13), we may rewrite (5-12) as follows:

$$\mathrm{Cov}\ (H_u, H_v) =$$

$$\binom{g}{k} \sum_{j=0}^{k} \binom{g-k}{k-j}\binom{k}{j} \bar{p}_j(u,v) - \binom{g}{k} \sum_{j=0}^{k} \binom{g-k}{k-j}\binom{k}{j} \bar{p}(u)\bar{p}(v)$$

$$= \binom{g}{k} \sum_{j=0}^{k} \binom{g-k}{k-j}\binom{k}{j} [\bar{p}_j(u,v) - \bar{p}(u)\bar{p}(v)] \ . \tag{5-14}$$

We get (c) from (5-14) and the observation that if $u \neq v$ then $\bar{p}_k(u,v)=0$ since a k-subgraph can only be of one isomorphism type. We get (b) from (5-14) and the fact that if u=v then $\bar{p}_k\ (u,u) = \bar{p}(u)$. ||

B. Moments of a Triad Census

The following corollary follows immediately from Theorem 1 in that it specializes it to k=3 -- the triad census. It is almost exactly the same as Theorem 1 of Holland and Leinhardt [1970] except that it allows a more general stochastic mechanism to generate the random digraph.

Corollary 1: Under the assumptions of Theorem 1, the first and second moments of a triad census $\underset{\sim}{T}$ are given by:

(a) $E(T_u) = \binom{g}{3} \bar{p}(u)$ (5-15)

(b) $\mathrm{Var}(T_u) = \binom{g}{3} \bar{p}(u)(1-\bar{p}(u)) + \binom{g}{3} \sum_{j=0}^{2} \binom{g-3}{3-j}\binom{3}{j} [\bar{p}_j(u,u)-(\bar{p}(u))^2]$ (5-16)

(c) $\mathrm{Cov}(T_u, T_v) = -\binom{g}{3} \bar{p}(u)\bar{p}(v) + \binom{g}{3} \sum_{j=0}^{2} \binom{g-3}{3-j}\binom{3}{j} [\bar{p}_j(u,v)-\bar{p}(u)\bar{p}(v)]$ (5-17)

In Holland and Leinhardt [1970] we derived some of the relevant probabilities for a particular random digraph that was introduced in Davis and Leinhardt [1972]. Since we have discussed these derivations extensively elsewhere, we shall only give a description of the random digraph and the complete tables of the relevant probabilities.

The U|MAN Random Digraph Distribution: This is the probability distribution on the set of all digraphs with g nodes which makes all digraphs with given values of M, A, and N (defined in (4-3), (4-4) and (4-5)) equally likely. In other words, this is the uniform distribution on the set of all labeled digraphs having given values of M, A, and N -- hence the notation U|MAN. One way to generate random digraphs from this distribution is as follows. First randomly allocate M mutual pairs to the $\binom{g}{2}$ possible pairs. Next randomly allocate A asymmetric pairs to the remaining $\left(\binom{g}{2} - M\right)$ pairs and then randomly and independently orient the asymmetric pairs. The result is a random labeled digraph with the given values of M, A, and N and all such digraphs are equally likely to be generated by this mechanism.

A random digraph with the U|MAN distribution possesses certain properties which simplify the calculation of $p_K(u)$, $p_{K,L}(u,v)$, $\bar{p}_j(u,v)$ and $\bar{p}(u)$. The U|MAN distribution on the set of digraphs is "homogeneous" in the sense that it is invariant under permutations of the labels given to the nodes.

Because of this property, $p_K(u)$ does not depend on K so that we have

$$p_K(u) = \bar{p}\ (u) = p(u) \qquad (5\text{-}18)$$

$= P$ {triad involving nodes 1,2,3 is of type u} .

Furthermore, the probabilities $p_{K,L}(u,v)$ only depend on u,v and $|K \cap L|$ so that if $|K \cap L| = j$

then

$$p_{K,L}(u,v) = \bar{p}_j(u,v) = p_j(u,v) \quad . \qquad (5\text{-}19)$$

Finally it is also easy to see that for the U|MAN distribution

$$p_0(u,v) = p_1(u,v) \quad . \qquad (5\text{-}20)$$

Corollary 2 summarizes the above discussion.

Corollary 2: For a triad census $\underset{\sim}{T}$ from a random digraph with the U|MAN distribution the means, variances and covariances are given by:

(a) $E(T_u) = \binom{g}{3}\ p(u)$ (5-21)

(b) $\mathrm{Var}\ (T_u) = \binom{g}{3}\ p(u)(1-p(u)) + \binom{g}{3}^2 \left(\frac{g+7}{g-2}\right) [p_0(u,u) - (p(u))^2]$

$+ 3 \binom{g}{3} (g-3)\ [p_2(u,u) - (p(u))^2]$ (5-22)

(c) $\mathrm{Cov}\ (T_u, T_v) = -\binom{g}{3}\ p(u)\ p(v) + \binom{g}{3}^2 \left(\frac{g+7}{g-2}\right) [p_0(u,v) - (p(u))^2]$

$+ 3 \binom{g}{3} (g-3)\ [p_2(u,v) - p(u)\ p(v)],$ (5-23)

where

$p(u)$, $p_0(u,v)$ and $p_2(u,v)$ are given in Tables 5-1, 5-2, and 5-3.

In Holland and Leinhardt [1970] we illustrated how some of the values of p(u), $P_0(u,v)$ and $p_2(u,v)$ are calculated so we do not repeat these derivations here. Tables 5-1,2,3 give the formulas for p(u), $p_0(u,v)$ and $p_2(u,v)$ for the U|MAN distribution, extending the corresponding tables in Holland and Leinhardt [1970] which only gave values for the intransitive triad types, i.e., those which contain at least one intransitive configuration.

In Tables 5-1,2,3, we have used the decending factorial notation in which $X^{(k)} = X(X-1)\cdots(X-k+1)$. Furthermore, Tables 5-1,2,3 contain only the numerator for the probabilities. The denomenators are respectivly $D_1$, $D_2$ and $D_3$ where

$$D_1 = \binom{g}{2}^{(3)} \tag{5-24}$$

$$D_2 = \binom{g}{2}^{(6)} \tag{5-25}$$

$$D_3 = \binom{g}{2}^{(5)} \tag{5-26}$$

Tables 5-1,2,3 go about here

C. Moments of Linear Combinations of a Triad Census

Once all the work has been done to generate the means, variances and covariances for a random triad census $\underset{\sim}{T}$, it is very simple to calculate the corresponding moments of any linear combination, $\underset{\sim}{\ell}'\underset{\sim}{T}$. It is an elementary result from probability theory that if $\underset{\sim}{\ell}'\underset{\sim}{T}$ and $\underset{\sim}{s}'\underset{\sim}{T}$ denote two linear combinations of the components of a triad census then

$$E(\underset{\sim}{\ell}'\underset{\sim}{T}) = \underset{\sim}{\ell}'\underset{\sim}{\mu}_T \tag{5-27}$$

$$\mathrm{Var}\,(\underset{\sim}{\ell}'\underset{\sim}{T}) = \underset{\sim}{\ell}'\underset{\sim}{\Sigma}_T\,\underset{\sim}{s} \tag{5-28}$$

and

$$\mathrm{Cov}\,(\underset{\sim}{\ell}'\underset{\sim}{T}, \underset{\sim}{s}'\underset{\sim}{T}) = \underset{\sim}{\ell}'\underset{\sim}{\Sigma}_T\,\underset{\sim}{s} \tag{5-29}$$

where $\underset{\sim}{\mu}_T = E(\underset{\sim}{T})$ is the vector of expected values of the $T_u$ and $\underset{\sim}{\Sigma}_T$ is the covariance matrix of $\underset{\sim}{T}$.

In section 6 we show how to use these results to develop statistical tests of propositions about local structure in empirical sociomatrices.

Table 5-1: Numerators for $p(u)$ Under U|MAN Distribution.

| u | p(u) |
|---|---|
| 003 | $n^{(3)}$ |
| 012 | $3\,a\,n^{(2)}$ |
| 102 | $3\,m\,n^{(2)}$ |
| 021D | $\frac{3}{4}\,n\,a^{(2)}$ |
| 021U | $\frac{3}{4}\,n\,a^{(2)}$ |
| 021C | $\frac{3}{2}\,n\,a^{(2)}$ |
| 111D | $3\,m\,a\,n$ |
| 111U | $3\,m\,a\,n$ |
| 030T | $\frac{3}{4}\,a^{(3)}$ |
| 030C | $\frac{1}{4}\,a^{(3)}$ |
| 201 | $3\,n\,m^{(2)}$ |
| 120D | $\frac{3}{4}\,m\,a^{(2)}$ |
| 120U | $\frac{3}{4}\,m\,a^{(2)}$ |
| 120C | $\frac{3}{2}\,m\,a^{(2)}$ |
| 210 | $3\,a\,m^{(2)}$ |
| 300 | $m^{(3)}$ |

Table 5-2: Numerator for $p_0(u,v)$ Under U|MAN Distribution.

(part 1)

| | 003 | 012 | 102 | 021D |
|---|---|---|---|---|
| 003 | $n^{(6)}$ | --- | --- | --- |
| 012 | $3\ a\ n^{(5)}$ | $9\ a^{(2)}n^{(4)}$ | --- | --- |
| 102 | $3\ m\ n^{(5)}$ | $9\ m\ a\ n^{(4)}$ | $9\ m^{(2)}n^{(4)}$ | --- |
| 021D | $\frac{3}{4}\ a^{(2)}n^{(4)}$ | $\frac{9}{4}\ a^{(3)}n^{(3)}$ | $\frac{9}{4}\ m\ a^{(2)}n^{(3)}$ | $\frac{9}{16}\ a^{(4)}n^{(2)}$ |
| 021U | $\frac{3}{4}\ a^{(2)}n^{(4)}$ | $\frac{9}{4}\ a^{(3)}n^{(3)}$ | $\frac{9}{4}\ m\ a^{(2)}n^{(3)}$ | $\frac{9}{16}\ a^{(4)}n^{(2)}$ |
| 021C | $\frac{3}{2}\ a^{(2)}n^{(4)}$ | $\frac{9}{2}\ a^{(3)}n^{(3)}$ | $\frac{9}{2}\ m\ a^{(2)}n^{(3)}$ | $\frac{9}{8}\ a^{(4)}n^{(2)}$ |
| 111D | $3\ m\ a\ n^{(4)}$ | $9\ m\ a^{(2)}n^{(3)}$ | $9\ m^{(2)}a\ n^{(3)}$ | $\frac{9}{4}\ m\ a^{(3)}n^{(2)}$ |
| 111U | $3\ m\ a\ n^{(4)}$ | $9\ m\ a^{(2)}n^{(3)}$ | $9\ m^{(2)}a\ n^{(3)}$ | $\frac{9}{4}\ m\ a^{(3)}n^{(2)}$ |
| 030T | $\frac{3}{4}\ a^{(3)}n^{(3)}$ | $\frac{9}{4}\ a^{(4)}n^{(2)}$ | $\frac{9}{4}\ m\ a^{(3)}n^{(2)}$ | $\frac{9}{16}\ a^{(5)}\ n$ |
| 030C | $\frac{1}{4}\ a^{(3)}n^{(3)}$ | $\frac{3}{4}\ a^{(4)}n^{(2)}$ | $\frac{3}{4}\ m\ a^{(3)}n^{(2)}$ | $\frac{3}{16}\ a^{(5)}\ n$ |
| 201 | $3\ m^{(2)}n^{(4)}$ | $9\ m^{(2)}a\ n^{(3)}$ | $9\ m^{(3)}n^{(3)}$ | $\frac{9}{4}\ m^{(2)}a^{(2)}n^{(2)}$ |
| 120D | $\frac{3}{4}\ m\ a^{(2)}n^{(3)}$ | $\frac{9}{4}\ m\ a^{(3)}n^{(2)}$ | $\frac{9}{4}\ m^{(2)}a^{(2)}n^{(2)}$ | $\frac{9}{16}\ m\ a^{(4)}n$ |
| 120U | $\frac{3}{4}\ m\ a^{(2)}n^{(3)}$ | $\frac{9}{4}\ m\ a^{(3)}n^{(2)}$ | $\frac{9}{4}\ m^{(2)}a^{(2)}n^{(2)}$ | $\frac{9}{16}\ m\ a^{(4)}n$ |
| 120C | $\frac{3}{2}\ m\ a^{(2)}n^{(3)}$ | $\frac{9}{2}\ m\ a^{(3)}n^{(2)}$ | $\frac{9}{2}\ m^{(2)}a^{(2)}n^{(2)}$ | $\frac{9}{8}\ m\ a^{(4)}n$ |
| 210 | $3\ m^{(2)}a\ n^{(3)}$ | $9\ m^{(2)}a^{(2)}n^{(2)}$ | $9\ m^{(3)}\ a\ n^{(2)}$ | $\frac{9}{4}\ m^{(2)}a^{(3)}n$ |
| 300 | $m^{(3)}\ n^{(3)}$ | $3\ m^{(3)}\ a\ n^{(2)}$ | $3\ m^{(4)}\ n^{(2)}$ | $\frac{3}{4}\ m^{(3)}a^{(2)}n$ |

(part 2)

| | 021U | 021C | 111D | 111U |
|---|---|---|---|---|
| 003 | --- | --- | --- | --- |
| 012 | --- | --- | --- | --- |
| 102 | --- | --- | --- | --- |
| 021D | --- | --- | --- | --- |
| 021U | $\frac{9}{16}\ a^{(4)}n^{(2)}$ | --- | --- | --- |
| 021C | $\frac{9}{8}\ a^{(4)}n^{(2)}$ | $\frac{9}{4}\ a^{(4)}n^{(2)}$ | --- | --- |
| 111D | $\frac{9}{4}\ m\ a^{(3)}n^{(2)}$ | $\frac{9}{2}\ m\ a^{(3)}n^{(2)}$ | $9\ m^{(2)}a^{(2)}n^{(2)}$ | --- |
| 111U | $\frac{9}{4}\ m\ a^{(3)}n^{(2)}$ | $\frac{9}{2}\ m\ a^{(3)}n^{(2)}$ | $9\ m^{(2)}a^{(2)}n^{(2)}$ | $9\ m^{(2)}a^{(2)}n^{(2)}$ |
| 030T | $\frac{9}{16}\ a^{(5)}\ n$ | $\frac{9}{8}\ a^{(5)}\ n$ | $\frac{9}{4}\ m\ a^{(4)}\ n$ | $\frac{9}{4}\ m\ a^{(4)}n$ |
| 030C | $\frac{3}{16}\ a^{(5)}\ n$ | $\frac{3}{8}\ a^{(5)}\ n$ | $\frac{3}{4}\ m\ a^{(4)}\ n$ | $\frac{3}{4}\ m\ a^{(4)}n$ |
| 201 | $\frac{9}{4}m^{(2)}a^{(2)}n^{(2)}$ | $\frac{9}{2}\ m^{(2)}a^{(2)}n^{(2)}$ | $9\ m^{(3)}a\ n^{(2)}$ | $9\ m^{(3)}a\ n^{(2)}$ |
| 120D | $\frac{9}{16}\ m\ a^{(4)}n$ | $\frac{9}{8}\ m\ a^{(4)}\ n$ | $\frac{9}{4}\ m^{(2)}a^{(3)}n$ | $\frac{9}{4}\ m^{(2)}a^{(3)}n$ |
| 120U | $\frac{9}{16}\ m\ a^{(4)}n$ | $\frac{9}{8}\ m\ a^{(4)}\ n$ | $\frac{9}{4}\ m^{(2)}a^{(3)}n$ | $\frac{9}{4}\ m^{(2)}a^{(3)}n$ |
| 120C | $\frac{9}{8}\ m\ a^{(4)}n$ | $\frac{9}{4}\ m\ a^{(4)}\ n$ | $\frac{9}{2}\ m^{(2)}a^{(3)}n$ | $\frac{9}{2}\ m^{(2)}a^{(3)}n$ |
| 210 | $\frac{9}{4}\ m^{(2)}a^{(3)}n$ | $\frac{9}{2}\ m^{(2)}a^{(3)}n$ | $9\ m^{(3)}a^{(2)}n$ | $9\ m^{(3)}a^{(2)}n$ |
| 300 | $\frac{3}{4}\ m^{(3)}a^{(2)}n$ | $\frac{3}{2}\ m^{(3)}a^{(2)}n$ | $3\ m^{(4)}a\ n$ | $3\ m^{(4)}a\ n$ |

(part 3)

| | 030T | 030C | 201 | 120D |
|---|---|---|---|---|
| 003 | --- | --- | --- | --- |
| 012 | --- | --- | --- | --- |
| 102 | --- | --- | --- | --- |
| 021D | --- | --- | --- | --- |
| 021U | --- | --- | --- | --- |
| 021C | --- | --- | --- | --- |
| 111D | --- | --- | --- | --- |
| 111U | --- | --- | --- | --- |
| 030T | $\frac{9}{16}\ a^{(6)}$ | --- | --- | --- |
| 030C | $\frac{3}{16}\ a^{(6)}$ | $\frac{1}{16}\ a^{(6)}$ | --- | --- |
| 201 | $\frac{9}{4}\ m^{(2)}a^{(3)}n$ | $\frac{3}{4}\ m^{(2)}a^{(3)}n$ | $9\ m^{(4)}n^{(2)}$ | --- |
| 120D | $\frac{9}{16}\ m\ a^{(5)}$ | $\frac{3}{16}\ m\ a^{(5)}$ | $\frac{9}{4}\ m^{(3)}a^{(2)}n$ | $\frac{9}{16}\ m^{(2)}a^{(4)}$ |
| 120U | $\frac{9}{16}\ m\ a^{(5)}$ | $\frac{3}{16}\ m\ a^{(5)}$ | $\frac{9}{4}\ m^{(3)}a^{(2)}n$ | $\frac{9}{16}\ m^{(2)}a^{(4)}$ |
| 120C | $\frac{9}{8}\ m\ a^{(5)}$ | $\frac{3}{8}\ m\ a^{(5)}$ | $\frac{9}{2}\ m^{(3)}a^{(2)}n$ | $\frac{9}{8}\ m^{(2)}a^{(4)}$ |
| 210 | $\frac{9}{4}\ m^{(2)}a^{(4)}$ | $\frac{3}{4}\ m^{(2)}a^{(4)}$ | $9\ m^{(4)}\ a\ n$ | $\frac{9}{4}\ m^{(3)}a^{(3)}$ |
| 300 | $\frac{3}{4}\ m^{(3)}a^{(3)}$ | $\frac{1}{4}\ m^{(3)}a^{(3)}$ | $3\ m^{(5)}\ n$ | $\frac{3}{4}\ m^{(4)}a^{(2)}$ |

(part 4)

| | 120U | 120C | 210 | 300 |
|---|---|---|---|---|
| 003 | --- | --- | --- | --- |
| 012 | --- | --- | --- | --- |
| 102 | --- | --- | --- | --- |
| 021D | --- | --- | --- | --- |
| 021U | --- | --- | --- | --- |
| 021C | --- | --- | --- | --- |
| 111D | --- | --- | --- | --- |
| 111U | --- | --- | --- | --- |
| 030T | --- | --- | --- | --- |
| 030C | --- | --- | --- | --- |
| 201 | --- | --- | --- | --- |
| 120D | --- | --- | --- | --- |
| 120U | $\frac{9}{16}\ m^{(2)}a^{(4)}$ | --- | --- | --- |
| 120C | $\frac{9}{8}\ m^{(2)}a^{(4)}$ | $\frac{9}{4}\ m^{(2)}a^{(4)}$ | --- | --- |
| 210 | $\frac{9}{4}\ m^{(3)}a^{(3)}$ | $\frac{9}{2}\ m^{(3)}a^{(3)}$ | $9\ m^{(4)}\ a^{(2)}$ | --- |
| 300 | $\frac{3}{4}\ m^{(4)}a^{(2)}$ | $\frac{3}{2}\ m^{(4)}a^{(2)}$ | $3\ m^{(5)}\ a$ | $m^{(6)}$ |

Table 5-3: Numerator for $p_2$ (u,v) Under U|MAN Distribution. (part 1)

| | 003 | 012 | 102 | 021D |
|---|---|---|---|---|
| 003 | $n^{(5)}$ | --- | --- | --- |
| 012 | $2an^{(4)}$ | $a\ n^{(3)}(4a+n-7)$ | --- | --- |
| 102 | $2mn^{(4)}$ | $4\ m\ a\ n^{(3)}$ | $mn^{(3)}(4m+n-7)$ | --- |
| 021D | $\frac{1}{4}a^{(2)}n^{(3)}$ | $\frac{1}{2}a^{(2)}n^{(2)}(a+n-4)$ | $\frac{1}{2}\ m\ a^{(2)}n^{(2)}$ | $\frac{1}{16}a^{(3)}n(a+4n-7)$ |
| 021U | $\frac{1}{4}a^{(2)}n^{(3)}$ | $\frac{1}{2}a^{(2)}n^{(2)}(a+n-4)$ | $\frac{1}{2}\ m\ a^{(2)}n^{(2)}$ | $\frac{1}{16}a^{(3)}n(a+4n-7)$ |
| 021C | $\frac{1}{2}a^{(2)}n^{(3)}$ | $a^{(2)}n^{(2)}(a+n-4)$ | $m\ a^{(2)}n^{(2)}$ | $\frac{1}{8}a^{(3)}n(a+4n-7)$ |
| 111D | $m\ a\ n^{(3)}$ | $man^{(2)}(2a+n-4)$ | $man^{(2)}(2m+n-4)$ | $\frac{1}{4}ma^{(2)}n(a+2n-4)$ |
| 111U | $m\ a\ n^{(3)}$ | $man^{(2)}(2a+n-4)$ | $man^{(2)}(2m+n-4)$ | $\frac{1}{4}ma^{(2)}n(a+2n-4)$ |
| 030T | 0 | $\frac{3}{4}\ a^{(3)}n^{(2)}$ | 0 | $\frac{3}{8}\ a^{(4)}\ n$ |
| 030C | 0 | $\frac{1}{4}\ a^{(3)}n^{(2)}$ | 0 | $\frac{1}{8}\ a^{(4)}\ n$ |
| 201 | $m^{(2)}n^{(3)}$ | $2m^{(2)}an^{(2)}$ | $2m^{(2)}n^{(2)}(m+n-4)$ | $\frac{1}{4}\ m^{(2)}a^{(2)}n$ |
| 120D | 0 | $\frac{1}{2}\ m\ a^{(2)}n^{(2)}$ | $\frac{1}{4}\ m\ a^{(2)}n^{(2)}$ | $\frac{1}{4}\ m\ a^{(3)}\ n$ |
| 120U | 0 | $\frac{1}{2}\ m\ a^{(2)}n^{(2)}$ | $\frac{1}{4}\ m\ a^{(2)}n^{(2)}$ | $\frac{1}{4}\ m\ a^{(3)}\ n$ |
| 120C | 0 | $m\ a^{(2)}n^{(2)}$ | $\frac{1}{2}\ m\ a^{(2)}n^{(2)}$ | $\frac{1}{2}\ m\ a^{(3)}\ n$ |
| 210 | 0 | $m^{(2)}a\ n^{(2)}$ | $2\ m^{(2)}a\ n^{(2)}$ | $\frac{1}{2}\ m^{(2)}a^{(2)}\ n$ |
| 300 | 0 | 0 | $m^{(3)}n^{(2)}$ | 0 |

Table 5-3: (cont'd) (part 2)

| | 021U | 021C | 111D | 111U |
|---|---|---|---|---|
| 003 | --- | --- | --- | --- |
| 012 | --- | --- | --- | --- |
| 102 | --- | --- | --- | --- |
| 021D | --- | --- | --- | --- |
| 021U | $\frac{1}{16}a^{(3)}n(a+4n-7)$ | --- | --- | --- |
| 021C | $\frac{1}{8}a^{(3)}n(a+4n-7)$ | $\frac{1}{4}a^{(3)}n(a+4n-7)$ | --- | --- |
| 111D | $\frac{1}{4}ma^{(2)}n(a+2n-4)$ | $\frac{1}{2}ma^{(2)}n(a+2n-4)$ | $ma^{(2)}n^{(2)}+m^{(2)}an^{(2)}+m^{(2)}a^{(2)}n$ | --- |
| 111U | $\frac{1}{4}ma^{(2)}n(a+2n-4)$ | $\frac{1}{2}ma^{(2)}n(a+2n-4)$ | $ma^{(2)}n^{(2)}+m^{(2)}an^{(2)}+m^{(2)}a^{(2)}n$ | $ma^{(2)}n^{(2)}+m^{(2)}an^{(2)}+m^{(2)}a^{(2)}n$ |
| 030T | $\frac{3}{8}a^{(4)}n$ | $\frac{3}{4}a^{(4)}n$ | $\frac{3}{4}ma^{(3)}n$ | $\frac{3}{4}m\ a^{(3)}n$ |
| 030C | $\frac{1}{8}a^{(4)}n$ | $\frac{1}{4}a^{(4)}n$ | $\frac{1}{4}ma^{(3)}n$ | $\frac{1}{4}m\ a^{(3)}n$ |
| 201 | $\frac{1}{4}m^{(2)}a^{(2)}n$ | $\frac{1}{2}m^{(2)}a^{(2)}n$ | $m^{(2)}an(m+2n-4)$ | $m^{(2)}an(m+2n-4)$ |
| 120D | $\frac{1}{4}ma^{(3)}n$ | $\frac{1}{2}m\ a^{(3)}n$ | $\frac{1}{4}ma^{(2)}n(a+2m-4)$ | $\frac{1}{4}ma^{(2)}n(a+2m-4)$ |
| 120U | $\frac{1}{4}ma^{(3)}n$ | $\frac{1}{2}m\ a^{(3)}n$ | $\frac{1}{4}ma^{(2)}n(a+2m-4)$ | $\frac{1}{4}ma^{(2)}n(a+2m-4)$ |
| 120C | $\frac{1}{2}ma^{(3)}n$ | $m\ a^{(3)}n$ | $\frac{1}{2}ma^{(2)}n(a+2m-4)$ | $\frac{1}{2}ma^{(2)}n(a+2m-4)$ |
| 210 | $\frac{1}{2}m^{(2)}a^{(2)}n$ | $m^{(2)}a^{(2)}n$ | $m^{(2)}an(m+2a-4)$ | $m^{(2)}an(m+2a-4)$ |
| 300 | 0 | 0 | $m^{(3)}an$ | $m^{(3)}an$ |

Table 5-3: (cont'd) (part 3)

| | 030T | 030C | 201 | 120D |
|---|---|---|---|---|
| 003 | --- | --- | --- | --- |
| 012 | --- | --- | --- | --- |
| 102 | --- | --- | --- | --- |
| 021D | --- | --- | --- | --- |
| 021U | --- | --- | --- | --- |
| 021C | --- | --- | --- | --- |
| 111D | --- | --- | --- | --- |
| 111U | --- | --- | --- | --- |
| 030T | $\frac{9}{16} a^{(5)}$ | --- | --- | --- |
| 030C | $\frac{3}{16} a^{(5)}$ | $\frac{1}{16} a^{(5)}$ | --- | --- |
| 201 | 0 | 0 | $m^{(3)}n[m+4n-7]$ | --- |
| 120D | $\frac{3}{8} m\ a^{(4)}$ | $\frac{1}{8} m\ a^{(4)}$ | $\frac{1}{2} m^{(2)}a^{(2)}n$ | $\frac{1}{16} ma^{(3)}[a+4m-7]$ |
| 120U | $\frac{3}{8} m\ a^{(4)}$ | $\frac{1}{8} m\ a^{(4)}$ | $\frac{1}{2} m^{(2)}a^{(2)}n$ | $\frac{1}{16} ma^{(3)}[a+4m-7]$ |
| 120C | $\frac{3}{4} m\ a^{(4)}$ | $\frac{1}{4} m\ a^{(4)}$ | $m^{(2)}a^{(2)}n$ | $\frac{1}{8} m\ a^{(3)}[a+4m-7]$ |
| 210 | $\frac{3}{4} \cdot m^{(2)}a^{(3)}$ | $\frac{1}{4} m^{(2)}a^{(3)}$ | $4\ m^{(3)}a\ n$ | $\frac{1}{2} m^{(2)}a^{(2)}[m+a-4]$ |
| 300 | 0 | 0 | $2\ m^{(4)}n$ | $\frac{1}{4} m^{(3)}a^{(2)}$ |

Table 5-3: (cont'd) (part 4)

| | 120U | 120C | 210 | 300 |
|---|---|---|---|---|
| 003 | --- | --- | --- | --- |
| 012 | --- | --- | --- | --- |
| 102 | --- | --- | --- | --- |
| 021D | --- | --- | --- | --- |
| 021U | --- | --- | --- | --- |
| 021C | --- | --- | --- | --- |
| 111D | --- | --- | --- | --- |
| 111U | --- | --- | --- | --- |
| 030T | --- | --- | --- | --- |
| 030C | --- | --- | --- | --- |
| 201 | --- | --- | --- | --- |
| 120D | --- | --- | --- | --- |
| 120U | $\frac{1}{16}\, m\, a^{(3)}[a+4m-7]$ | --- | --- | --- |
| 120C | $\frac{1}{8}\, m\, a^{(3)}[a+4m-7]$ | $\frac{1}{4}\, m\, a^{(3)}[a+4m-7]$ | --- | --- |
| 210 | $\frac{1}{2}\, m^{(2)} a^{(2)}[m+a-4]$ | $m^{(2)} a^{(2)}[m+a-4]$ | $m^{(3)} a[m+4a-7]$ | --- |
| 300 | $\frac{1}{4}\, m^{(3)} a^{(2)}$ | $\frac{1}{2}\, m^{(3)} a^{(2)}$ | $2\, m^{(4)} a$ | $m^{(5)}$ |

## 6. Testing Structural Hypotheses

In this section we make use of the tools developed in the previous sections in a procedure for testing propositions about local structure in a sociomatrix. In outline, the procedure has these steps:

(1) Operationalize the proposition into a hypothesis that a particular 3-configuration will tend to occur or fail to occur in the sociomatrix (this hypothesis will usually be directional).

(2) Find the weighting vector which when applied to the triad census will enumerate the critical 3-configuration.

(3) Use the weighting vector to enumerate the critical 3-configuration as well as to compute its mean and variance for a random triad census.

(4) Set up a test statistic that compares the observed and the expected number of critical configurations and use this discrepancy as a basis for testing the structural proposition.

We discuss each of these steps in turn.

### A. Operationalizing a Structural Proposition

In this initial step we take a proposition about the structure of a network and translate it, if possible, into a prediction about the number of 3-configurations of a particular type in observed sociomatrices. As an example we consider Mazur's [1971] proposition mentioned earlier in section 3A. We first note that a simple sociomatrix indicating choice or non-choice does not represent the strength of the relation so that Mazur's distinction between "friends" and "close friends" has to be made on some other basis if this proposition is to be operationalized and tested on binary choice data. We follow Mazur's suggestion and assume that mutual dyads indicate "close friends" while asymmetric dyads indicate "friends". Mazur's proposition leads us to examine not one but seven different 3-configurations. They all have the reading rule given by:

$$ij \quad ji \quad ik \quad jk \ . \tag{6-1}$$

The first two pairs in the reading rule refer to the pair of individuals who are designated "close friends", "friends" or neither. The second two pairs in the reading rule refer to the choice or non-choice of other group members by the pair in question. For example, the configuration that corresponds to "close friends agreeing on their choices" is given by:

$$\begin{pmatrix} ij & ji & ik & jk \\ 1 & 1 & 1 & 1 \end{pmatrix} . \tag{6-2}$$

On the other hand "close friends agreeing on their non-choices" is the configuration:

$$\begin{pmatrix} ij & ji & ik & jk \\ 1 & 1 & 0 & 0 \end{pmatrix} . \tag{6-3}$$

"Close friends disagreeing on their choices" is given by either:

$$\begin{pmatrix} ij & ji & ik & jk \\ 1 & 1 & 0 & 1 \end{pmatrix} \tag{6-4}$$

or

$$\begin{pmatrix} ij & ji & ik & jk \\ 1 & 1 & 1 & 0 \end{pmatrix} , \tag{6-5}$$

as mentioned earlier in section 3B. Since the reading rule is understood here to be the one given in (6-1) we shall use the following shorthand notation for these configurations: (6-2) is denoted by 1111, (6-3) by 1100, and (6-4) or (6-5) by 1101/1110. Thus, we only use the lower part of the configuration matrix and if there are two equivalent forms they are separated by a slash. In this notation Mazur's proposition makes predictions about the following seven configurations: 1111, 1110, 1101/1110, 1011/0111, 1000/0100, 1010/0101, 1001/0110. The first three deal with agreement and disagreement among "close friends" while the second four deal with agreement and disagreement among "friends".

Now for the sign of the prediction. We propose to compare the observed number of the seven configurations mentioned above with the corresponding number expected in a random sociomatrix having the same number of mutual, asymmetric and null dyads. This is the U|MAN distribution for a random digraph discussed in section 5, and it appears to be especially appropriate for Mazur's proposition because they are based on mutual and asymmetric pairs. In view of the foregoing remarks, we would formulate Mazur's predictions as given by the second column of Table 6-1.

Table 6-1 goes about here

Note that under the U|MAN distribution the seven predictions in Table 6-1 are not independent. For any observed sociomatrix the number of configurations of the first four types in Table 6-1, those grouped under "friends", sum to (g-2)A. The last three types, that are grouped under "close friends", sum to (g-2)M. These summations also apply to the expected values under the U|MAN distribution. There are therefore only 5 independent predictions. Such considerations must often be taken into consideration when examining more than one configuration.

Table 6-1: Results of Testing Mazur's Proposition.

| | Configuration | Model | Predicted sign | Median $\tau(\ell)$ for 408 sociomatrices |
|---|---|---|---|---|
| Friends | 1011/0111 | Agreement | + | 2.36 |
| | 1000/0100 | | + | 1.82 |
| | 1010/0101 | Disagreement | - | -1.73 |
| | 1001/0110 | | | -1.49 |
| Close Friends | 1111 | Agreement | + | 3.48 |
| | 1100 | | + | 1.73 |
| | 1101/1110 | Disagreement | - | -3.81 |

B. The Weighting Vector

In this step, we compute the number of ways each critical configuration occurs within the 16 triad types. Each resulting set of 16 numbers forms the elements of the weighting vector that is applied to the triad census to enumerate the critical configuration. For example, consider the configuration denoted 1111 above from (6-2). As we go through the list of triad types in Figure 2-2 we see that this configuration does not occur in any triads until we come to the triad 120U. This triad contains exactly one configuration of type 1111. Triads 210 and 300 contain, respectively, 1 and 3 configurations of this type. Hence, the weighting vector for the 1111 configuration is (0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 3). Table 6-2 gives the weighting vectors for all of the configurations critical to Mazur's proposition.

Table 6-2 goes about here

The formulas for the mean and variance of the number of these configurations in a random triad census from the U|MAN distribution are given by formulas (5-27) and (5-28). A computer program that carries out these calculations has been written and is implimented on the TROLL interactive computer system of the National Bureau of Economic Research, Inc.

C. Setting up the Test Statistic

We let $\underset{\sim}{\ell}$ denote the weighting vector for a critical configuration, and $\underset{\sim}{T}$ denote the triad census vector as usual. Then $\underset{\sim}{\ell}'\underset{\sim}{T}$ is the number of times the critical configuration occurs in the observed sociomatrix. Under the U|MAN distribution the expected number of these configurations is $\underset{\sim}{\ell}'\underset{\sim}{\mu}_T$ where $\underset{\sim}{\mu}_T$ is given by Corollary 2 (a). The variance is $\underset{\sim}{\ell}'\underset{\sim}{\Sigma}_T\underset{\sim}{\ell}$ where $\underset{\sim}{\Sigma}_T$ is a

Table 6-2: Weighting Vector for Configurations Critical to Mazur's Proposition.

| Triad | 1011 0111 | 1000 0100 | 1010 0101 | 1001 0110 | 1111 | 1100 | 1101 1110 |
|---|---|---|---|---|---|---|---|
| 003 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 012 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 102 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 021D | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 021U | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| 021C | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 111D | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 111U | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 030T | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 030C | 0 | 0 | 0 | 3 | 0 | 0 | 0 |
| 201 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| 120D | 2 | 0 | 0 | 0 | 0 | 1 | 0 |
| 120U | 0 | 0 | 2 | 0 | 1 | 0 | 0 |
| 120C | 1 | 0 | 0 | 1 | 0 | 0 | 1 |
| 210 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 300 | 0 | 0 | 0 | 0 | 3 | 0 | 0 |

16x16 matrix whose main diagonal elements are given by Corollary 2(b) and off-diagonal elements are given by Corollary 2(c).

The difference

$$\underset{\sim}{\ell}'\underset{\sim}{T} - \underset{\sim}{\ell}'\underset{\sim}{\mu}_T \tag{6-6}$$

is the discrepancy between the observed and expected number of critical configurations. Under the assumption that the triad census is random (U|MAN) the test statistic, $\tau(\underset{\sim}{\ell})$, defined by:

$$\tau(\underset{\sim}{\ell}) = \frac{\underset{\sim}{\ell}'\underset{\sim}{T} - \underset{\sim}{\ell}\ \underset{\sim}{\mu}_T}{\sqrt{\underset{\sim}{\ell}\ \underset{\sim}{\Sigma}_T\underset{\sim}{\ell}}} \tag{6-7}$$

has an approximate normal distribution with mean zero and variance one. Approximate significance levels for values of $\tau(\underset{\sim}{\ell})$ may be obtained from tables of the normal distribution. See Holland and Leinhardt [1970] for some simulation results bearing on the adequacy of this approximation.

Using a computer program, we have calculated the value of $\tau(\underset{\sim}{\ell})$ in 408 sociomatrices (randomly selected from those collected by Davis and Leinhardt [1972]) for the seven critical configurations for Mazur's proposition. In the last column of Table 6-1 we give the median value of $\tau(\underset{\sim}{\ell})$ over these 408 sociomatrices for each of the seven critical configurations. The results indicate that by and large the predictions for $\tau(\underset{\sim}{\ell})$ that derive from Mazur's proposition are supported. All of the median $\tau(\underset{\sim}{\ell})$ values have the predicted sign. Furthermore, Mazur's prediction that close friends will agree more than friends seems to be substantiated by direct comparison of the median $\tau(\underset{\sim}{\ell})$ values.

D. Another Example, Transitivity

Consider the following proposition stated by Davis, Holland and Leinhardt (1971): "Interpersonal choices tend to be transitive -- if P chooses O and O chooses X, then P is likely to choose X" (p. 309). Elsewhere (Holland and Leinhardt [1971]), we have described the social structural consequences of the transitivity of interpersonal affect when social status is associated with asymmetry and clustering is associated with mutuality. Here we use our method to explore the proposition's empirical validity.

There are two critical configurations, given in (3-2) (intransitivity) and (3-5) (transitivity). The corresponding weighting vectors are given in Table 4-1. Using the 408 sociomatrices mentioned above we obtained median $\tau(\ell)$ values of 5.18 for transitivity and -3.89 for intransitivity. These results support the proposition that interpersonal choices tend to be transitive.

However, these results and those found for Mazur's proposition are not independent of each other. Indeed, the weighting vector for transitively is the componentwise sum of the weighting vector for the configuration 1011/0111 and two times the weighting vector 1111. The weighting vector for intransitivity is the sum of the weighting vectors for configurations 1010/0101 and 1101/1110.

The transitivity prediction is therefore weaker than Mazur's in that it involves fewer configurations.

We may view the very strong observed effect of transitivity as the sum of two more modest effects corresponding to the configurations 1011/0111 and 1111 or conversely the effects of these two configurations may be viewed as the result of being "pulled along" by the very strong transitivity effect.

## 7. Summary and Discussion

Our purpose has been to describe a method for formalizing and testing theoretical propositions about regularities in social structure. The method involves the use of graph theoretic concepts to restate verbal propositions about local structure in terms of configurations. Although configurations are relatively simple graph theoretic concepts we showed that they could be used to represent some important social structural propositions. Arguing that most theories of structure in interpersonal relations concern average local properties we proceeded to develop statistical procedures for testing the tendency of a particular property to hold across a social network. With these procedures an investigator can determine whether the discrepancy between the empirical occurrence, and the chance occurrence of local structure is statistically significant. Some examples were presented to illustrate the use of the method and its interpretation.

Sociometric data are quite common. Because of their prevalence and variety they represent an important and fundamental resource in the study of structure in interpersonal relations. However, these data are complex and, like all empirical measurements, they contain an unknown amount of measurement error. Their inherent complexity and the likely presence of measurement error significantly reduce their applicability to the study of global organization in small social systems. Moreover, such application can also be questioned from the point of view of the level at which the data are collected. Interpersonal affect or choice data, the most frequently collected sociometric data, represent "surveys" of individual attitudes.

In effect, individuals are asked to provide information on the nature of their local ties. The task each group member is presented with is to judge whether a single link exists from that member to other group members. Whatever overall consistency exists in these sets of responses derives from local regularities in interpersonal relations. Although some simple global models may fit these data, the data are more properly used in the study of local conditions.

The method we have described possesses several advantages when used to study local structure in small-scale social systems. First, it leads investigators to develop operational constructs which are amenable to empirical testing. Many of the theoretical propositions advanced by sociologists and social psychologists concern the behavioral and experimental consequences of various arrangements of interpersonal relations. However, these propositions are rarely stated with precision and an absence of ambiguity that would permit them to be tested. Second, the method focuses attention on the analysis of average local structure in sociometric data. It thereby exploits the essential feature of these data. Finally the method facilitates the analysis of structural tendencies in large collections of groups. Only through analyses of large collections of network data will investigators be able to detect general tendencies in the structure of social relational systems.

## Appendix: Conditional Uniform Distributions for Random Graphs

Various considerations lead us to study random digraphs whose distributions are different from and often more complicated than the U|MAN distribution discussed in Section 5. In this appendix we describe some of these distributions and propose an approximation that may help with some of the more complicated ones.

### A. Examples of Random Digraphs

The Uniform Distribution: This is the basic distribution from which all the others we discuss here may be obtained by conditioning on particular statistics of the graph. For the uniform distribution, all labeled digraphs with g nodes are equally-likely. It is easy to generate the sociomatrix $(X_{ij})$ for a uniformly distributed random digraph, because the $X_{ij}$ are independent zero-one random variables with

$$P\{X_{ij}=1\} = \frac{1}{2} \qquad i \neq j. \tag{A-1}$$

The U|C Distribution: This is a simple conditional distribution of the uniform distribution just described. C defined in (4-6) denotes the number of directed edges in the graph. Thus U|C is the uniform distribution conditioned on C. It makes all labeled digraphs with a specified value of C equally-likely. It is also easy to generate the sociomatrix for such a random graph by selecting at random and without replacement C of the g(g-1) possible ordered pairs of nodes and allocating the C directed edges to them. In the uniform distribution C is a random variable with a binomial distribution whereas in the U|C distribution C is not random and is fixed at a specified value.

The $U|\{X_{i+}\}$ Distribution: This is the uniform distribution conditioned by the out-degrees. For the $U|\{X_{i+}\}$ distribution, all labeled digraphs with the specified out-degrees are equally-likely. To generate $(X_{ij})$ from $U|\{X_{i+}\}$ observe that all the rows of $(X_{ij})$ are statistically independent and that in the $i^{\underline{th}}$ row we merely need to choose at random and without replacement $X_{i+}$ columns (excluding the $i^{\underline{th}}$) for the ones and set the rest equal to zero. This distribution is an important baseline for the allocation of choices in a sociogram. By transposing $(X_{ij})$ the $U|\{X_{i+}\}$ becomes the $U|\{X_{+j}\}$ distribution. Note that in the $U|\{X_{i+}\}$ or the $U|\{X_{+j}\}$ distribution the value of C is fixed since

$$C = \sum_i X_{i+} = \sum_j X_{+j} \tag{A-2}$$

The $U|\{X_{i+}\}, \{X_{+j}\}$ Distribution: In this distribution all labeled digraphs with the specified values of both $\{X_{i+}\}$ and $\{X_{+j}\}$ are equally-likely. It is a highly non-trivial distribution and no simple way seems to be known for generating random graphs with this exact distribution. Nevertheless, it is of potential importance in sociometric data analysis since it conditions out both choices-made and choices-received.

The $U|M,\{X_{i+}\}$ Distribution: In this distribution we combine both U|MAN and $U|\{X_{i+}\}$. We need only specify M and $\{X_{i+}\}$ since $\{X_{i+}\}$ fixes C and M and C fix A. N is determined by M,A and g. This is also a highly non-trivial distribution and no simple way is known for generating graphs from it. Again, it is of potential importance in sociometry because it conditions out both choices-made and mutuality. Indeed, we would have prefered to use this distribution rather than the U|MAN but currently are not able to.

<u>The</u> $U|M$, $\{X_{i+}\}$, $\{X_{+j}\}$ <u>Distribution</u>: In this distribution, all digraphs with the specified values of M, $\{X_{i+}\}$, and $\{X_{+j}\}$ are equally-likely. Again this is a very difficult distribution to work with and no simple way is known for generating graphs from it. Its value to sociomatric data analysis stems from the interpretation that it controls for: (1) choices-made (possibly constrained by the experimental technique), (2) choices-received (a measure of status and isolation) and (3) mutuality (a measure of friendship).

In summary, there are a variety of possible types of random graphs besides the U|MAN distribution. We have merely catalogued a few of the important ones that stem from the notion of the uniform distribution. There are a host of possibilities for non-uniform distributions but we will not discuss them here. In a sense, the main virtue of U|MAN is that it is the most highly conditioned uniform distribution that fixes M and A for which we are currently able to compute the probabilities defined in Section 5.

## B. Approximate Distributions for the Triad Census

While the U|MAN distribution is useful, it does not control for the out-degrees (which may reflect experimental constraints like the fixed-choice procedure) or the in-degrees (which reflect status and isolation). We are left with the gnawing possibility that an observed triad census $\underset{\sim}{T}$ from real data departs substantially from its U|MAN expected value $\mu_{\underset{\sim}{T}}$ not because there is structure in the sociogram, but because $\mu_{\underset{\sim}{T}}$ does not control for all the simple effects we would like to condition out in our analysis.

Since a direct attack on the exact $U|M$, $\{X_{i+}\}$ or $U|M$, $\{X_{i+}\}$, $\{X_{+j}\}$ distribution appears to be too difficult, at least at present, we propose the following indirect and approximate approach which makes use of the simple nature of the conditional distributions for the multivariate normal.

We observe that $\underset{\sim}{T}$ is a sum of loosely correlated indicator variables (i.e. see (5-2) ), and thus it is plausible that under various random distributions $\underset{\sim}{T}$ has an approximate multivariate normal distribution for large values of g. Thus we have approximately

$$\underset{\sim}{T} \sim \mathcal{N}(\underset{\sim}{\mu}, \underset{\sim}{\Sigma}) \tag{A-3}$$

for $\underset{\sim}{\mu}$ and $\underset{\sim}{\Sigma}$ computed say from U|MAN. Next we make use of the following fact. If $\underset{\sim}{T}$ has the multivariate normal distribution in (A-3) and $\underset{\sim}{L}\,\underset{\sim}{T}$ is a vector of linear combinations of the elements of $\underset{\sim}{T}$ then:

$$E(\underset{\sim}{T} \mid \underset{\sim}{L}\,\underset{\sim}{T} = \underset{\sim}{L}\,\underset{\sim}{t}) = \underset{\sim}{\mu} + \underset{\sim}{\Sigma}\,\underset{\sim}{L}' \, (\underset{\sim}{L}\,\underset{\sim}{\Sigma}\,\underset{\sim}{L}')^{-1} \, \underset{\sim}{L}\,(\underset{\sim}{t}-\underset{\sim}{\mu}) \tag{A-4}$$

and

$$\mathrm{Cov}\,(\underset{\sim}{T} \mid \underset{\sim}{L}\,\underset{\sim}{T} = \underset{\sim}{L}\,\underset{\sim}{t}) = \underset{\sim}{\Sigma} - \underset{\sim}{\Sigma}\,\underset{\sim}{L}' \, (\underset{\sim}{L}\,\underset{\sim}{\Sigma}\,\underset{\sim}{L}')^{-1} \, \underset{\sim}{L}\,\underset{\sim}{\Sigma} \tag{A-5}$$

Thus it is fairly simple to approximately condition $\underset{\sim}{T}$ on linear combinations of its elements. Now we make one further approximation, instead of conditioning on $\{X_{i+}\}$ we condition on the mean, $\bar{C}$, and variance, $S^2_{out}$, of the $\{X_{i+}\}$. While this is clearly a reduction in the level of conditioning used, it promises to be useful. For example, if $S^2_{out} = 0$ this means that all the $\{X_{i+}\}$ are equal (i.e. the fixed - choice procedure). Now we use the result of section 4 that $\bar{C}$ and $S^2_{out}$ are essentially linear combinations of the triad frequencies to transform the problem to one of conditioning $\underset{\sim}{T}$ on some of its linear combinations. There are various ways to impliment this approach and only direct empirical will show us which ones are the most useful.

References

Berelson, B.R., Lazarsfeld, P.F. and McPhee, W.N. [1954] Voting: A Study of Opinion Formation in a Presidential Campaign Chicago, University of Chicago Press.

Boyle, R.P. [1969] "Algebraic systems for normal and hierarchical sociograms" Sociometry 32, 99-119.

Cartwright, D. and Harary, F. [1956] "Structural Balance: A generalization of Heider's theory" Psych. Review 63, 277-293.

Coleman, J.S. [1957] Community Conflict Glencoe, Illinois, Free Press.

Davis, J.A., [1963] "Structural balance, mechanical solidarity and interpersonal relations" Am. J. of Soc. 68, 444-463.

Davis, J.A. [1967] "Clustering and structural balance in graphs" Human Relations 20, 181-187.

Davis, J.A. [1970] "Clustering and hierarchy in interpersonal relations: testing two graph theoretical models on 742 sociograms" Am. Soc. Review 35, 843-852.

Davis, J.A., Holland, P.W., and Leinhardt, S. [1971] "Comments on Professor Mazur's hypothesis about interpersonal sentiments" Am. Soc. Review 36, 309-311.

Davis, J.A., and Leinhardt, S. [1972] "The structure of positive interpersonal relations in small groups" in Sociological Theories in Progress, Vol. 2 (edited by J. Berger) Boston, Houghton Mifflin.

Flament, C. [1963] "Applications of graph theory to group structure" Englewood Cliffs, N.J., Prentice-Hall.

French, J.R.P. [1956] "A formal theory of social power" Psych. Review 63, 181-195.

Harary, F., Norman, R.Z. and Cartwright, D. [1965] Structural Models, New York, Wiley.

Heider, F. [1944] "Social Perception and Phenomenal Causality" Psych. Review 51, 358-374.

Heider, F. [1958] The Psychology of Interpersonal Relations New York, Wiley.

Holland, P.W. and Leinhardt, S. [1970] "A method for detecting structure in sociometric data" Am. J. of Soc. 70 492-513.

Holland, P.W. and Leinhardt, S. [1971] "Transitivity in structural models of small groups." Comparative Group Studies 2, 107-124.

Holland, P.W. and Leinhardt, S.[1972] "Some evidence on the transitivity of positive interpersonal sentiment" Am. J. of Soc. 72, 1205-1209.

Holland, P.W. and Leinhardt, S. [1973] "The structural implications of measurement error in sociometry" J. Math. Soc. 3, 85-111.

Homans, G.C. [1950] The Human Group New York, Harcourt, Brace and World.

Homans, G.C. [1961] Social Behavior: Its Elementary Forms. New York, Harcourt, Brace and World.

Katz, E. and Lazarsfeld, P.F. [1955] Personal Influence: The Part Played by People in the flow of Mass Communications. Glencoe, Illinois, Free Press.

Lazarsfeld, P.F. and Merton, R.K. [1954] "Friendship as a social process: a substantive and methodological analysis" in Freedom and Control in Modern Society pg. 18-66 (edited by M. Berger, T. Abel, and C.H. Page) Princeton, N.J., Van Nostrand.

Lipset, S.M., Trow, M.A., and Coleman, J.S. [1956] Union Democracy: The Internal Politics of the International Typographical Union, Glencoe, Illinois, Free Press.

Lorrain, F. and White, H.C. [1971] "Structural equivalence of individuals in social networks" J. Math. Soc. 1, 49-80.

Mazur, A. [1971] "Comments on Davis Graph Model" Am. Socio. Review 36, 308-311.

Moon, J.W. [1968] Topics on Tournements. New York, Holt, Rinehart and Winston.

Moreno, J.L. [1934] Who Shall Survive? Washington, D.C., Nervous and Mental Disease Pub. Co.

Moreno, J.L. et al. [1960] The Sociometry Reader. Glencoe, Illinois, Free Press.

Nadel, S.F. [1957] The Theory of Social Structure- London, Cohen and West.

Radcliff-Brown, A.R. [1940] "On social structure" _J. Royal Anthro. Inst._ _70_, 1-12.

White, H.C. [1963] _An Anatomy of Kinship: Mathematical Models for Structures of Cumulated Roles_. New Jersey, Prentice-Hall.