The Annals of Statistics
1983, Vol. 11, No. 2, 377–391

GRAPH-THEORETIC MEASURES OF MULTIVARIATE ASSOCIATION AND PREDICTION[1]

BY JEROME H. FRIEDMAN AND LAWRENCE C. RAFSKY

Stanford University and GemNet Software Corporation

Interpoint-distance-based graphs can be used to define measures of association that extend Kendall's notion of a generalized correlation coefficient. We present particular statistics that provide distribution-free tests of independence sensitive to alternatives involving non-monotonic relationships. Moreover, since ordering plays no essential role, the ideas are fully applicable in a multivariate setting. We also define an asymmetric coefficient measuring the extent to which (a vector) X can be used to make single-valued predictions of (a vector) Y. We discuss various techniques for proving that such statistics are asymptotically normal. As an example of the effectiveness of our approach, we present an application to the examination of residuals from multiple regression.

1. Introduction. The theory of generalized correlation coefficients, as advanced by Daniels, Kendall, and others (see Kendall, 1962, page 19), exposes the underlying structure of such useful tools as Spearman's ρ and Kendall's τ. The theory can be exploited to provide measures of correlation that, while still intended to uncover monotone relationships, are less specifically directed towards linearity than Pearson's product-moment formulation. Permutation tests of "no correlation" that are distribution-free can be constructed from such statistics.

It is our purpose to (1) define tests that have power against non-monotone alternatives, (2) follow the same program in a multivariate setting, and (3) define a statistic that will measure how predictable a random vector Y is from a random vector X, without regard to how well X can be predicted from Y. This last statistic forms the basis of a test of "no correlation" that has power against alternatives where many X values may be associated with (nearly) the same Y value. Our approach is motivated by interpoint-distance-based graphs and focuses on the degree to which closeness of two vectors in one space is matched by closeness of the corresponding two vectors in the other. Note that this is not the same as requiring high correlation between the corresponding interpoint distances in the two spaces; large distances are not considered. When using the statistics proposed herein, values significant under the "no correlation" null hypothesis should be used to signal the need to examine the nature of the uncovered relationship, not as a final answer to some sharply defined question.

2. Generalized correlation coefficients. Consider a sample (x_i, y_i), $i = 1, \cdots, N$, of ordered pairs. A generalized correlation coefficient, ignoring standardization, is a statistic of the form

$$\Gamma = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} b_{ij}, \tag{1}$$

where a_{ij} is a score for every pair (i, j) of X observations and b_{ij} is a score for every pair of Y observations. As is well known, the choices $a_{ij} = x_i - x_j$, $b_{ij} = y_i - y_j$; $a_{ij} = \text{rank}(x_i) - \text{rank}(x_j)$, $b_{ij} = \text{rank}(y_i) - \text{rank}(y_j)$; and $a_{ij} = \text{sign}(x_i - x_j)$, $b_{ij} = \text{sign}(y_i - y_j)$ lead to the Pearson, Spearman, and Kendall measures, respectively (Kendall, 1962). Conditioning on the sets $\{x_i\}$ and $\{y_i\}$ of observed X and Y values, a test of "no correlation" is available

Received August 1981; revised October 1982.

[1] Work supported by the Department of Energy under contract DE-AC03-76SF00515.

AMS 1980 *subject classifications.* Primary, 62G10; secondary, 62H20.

Key words and phrases. Multivariate association, interpoint distances, graph theory, linear permutation statistics, examination of residuals.

using the distribution of

$$\Gamma(\pi) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} b_{\pi(i)\pi(j)}, \tag{2}$$

where π is a permutation of $(1, 2, \cdots, N)$, with all permutations having equal probability. The tails of this distribution determine whether or not the value (1) is too positive or negative an extreme. This is an appropriate test since "no correlation" may be thought of as having all $N!$ permutations of the Y subscripts—i.e., all possible X-Y pairings—equally likely.

The notion of a generalized correlation coefficient (1) need not be restricted to two univariate variables; scores a_{ij} and b_{ij} can be defined for vectors. However, the notion of ordering which plays an essential role in the formulation of Spearman's ρ and Kendall's τ is not available for multivariate observations.

3. Interpoint-distance based graphs. In extending such measures to multivariate observations, our intuition has been guided by the notion of interpoint-distance-based graphs. These graphs have the sample observations (considered as points in Euclidean space) as nodes. Such graphs can summarize the useful properties of the $N(N-1)/2$ interpoint distances. Consider a graph for which every observation point is a node, and each node pair defines an edge. Such a (*complete*) graph has N nodes and $N(N-1)/2$ edges. Assign as a weight to each edge the Euclidean distance (or generalized dissimilarity) between the nodes defining it. Our statistics are based on spanning subgraphs of this complete graph. A *subgraph* of a given graph is a graph with all of its nodes and edges in the given graph. A *spanning subgraph* has its node set identical to the node set of the given graph.

Spanning subgraphs that we have found useful are the K nearest-neighbor graph (KNN) and the K minimal spanning tree (KMST). The KNN has an edge between each point and its K closest points. The KMST is a generalization of the minimal spanning tree (MST), well known in graph theory. Define a *path* between two prescribed nodes in a graph as an alternating sequence of nodes and edges with the prescribed nodes as first and last elements, all other nodes distinct, and each edge linking the two nodes adjacent to it in the sequence. An MST is a spanning subgraph with the property that it provides a path between every pair of nodes with minimal sum of edge weights, i.e., minimal total distance. It is immediate that MSTs have precisely $N-1$ edges and do indeed form a tree, i.e., have no cycles. Note that the MST of points in R^1 simply connects the points in sorted order.

An MST connects (provides a path between) all of the points with minimal total distance (sum of edge weights). A second-MST connects all of the points with minimal total distance subject to the constraint that it share no edges with the MST. (Two graphs that share no edges are said to be *orthogonal.*) A k-MST is a minimal spanning tree orthogonal to the $(k-1)$ through the second MST and the MST. The KMST is the graph defined by all of the edges of the first k-MST's. A KMST clearly has $K(N-1)$ edges. If $K \ll N$, the edges of both the KMST and KNN graphs will mainly be defined by node pairs of small interpoint distance.

These concepts are illustrated in Figure 1. Figure 1a displays 50 points in the plane. Figure 1b shows the (1)NN graph for these points and Figure 1c shows the (1)MST. Figures 1d and 1e show the 5NN and 5MST respectively.

Interpoint-distance-based graphs can also be defined for (multivariate) unordered categorical variables using "Hamming Distance" (the number of coordinates for which two observations realize different categories). A graph can then be constructed by connecting nodes at zero distance from each other, or by choosing a strategy for resolving ties and building a KMST or KNN.

Computational methods for constructing KNN and KMST graphs, given a set of observations and an appropriately defined distance or dissimilarity measure, can be found in the appendix of Friedman and Rafsky (1979).

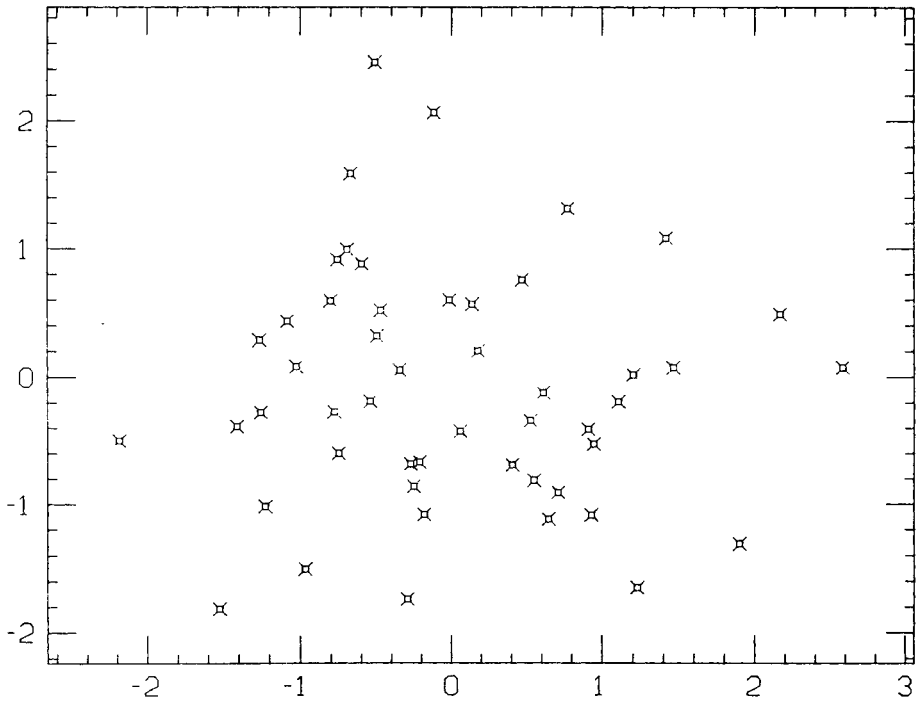

FIG. 1a: *Fifty points from a bivariate normal distribution.*

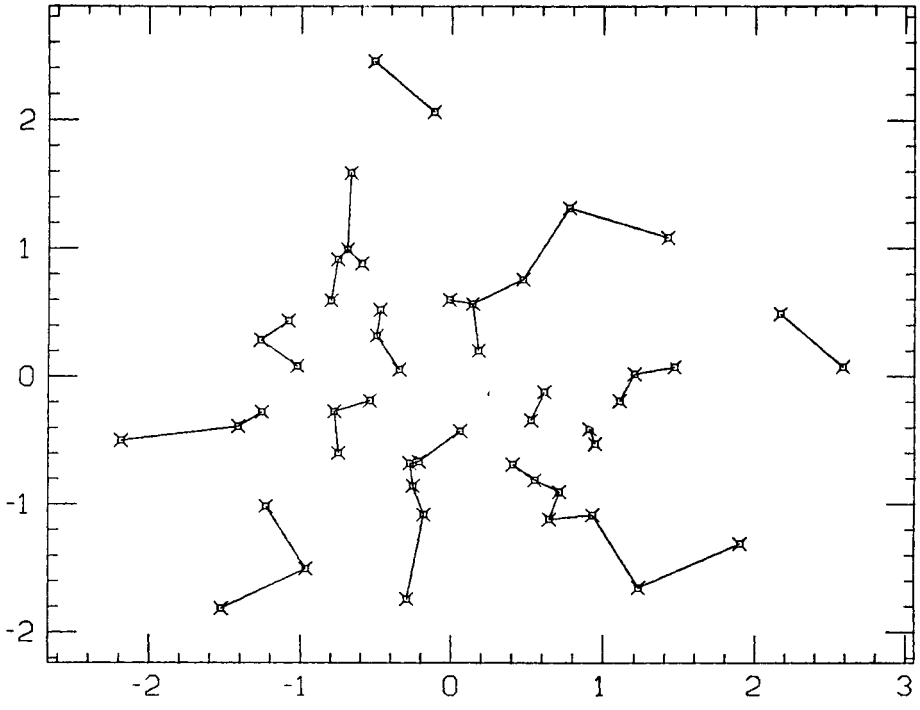

FIG. 1b: *Edges of* (1)NN (*Nearest Neighbor*) *graph shown as straight line segments.*

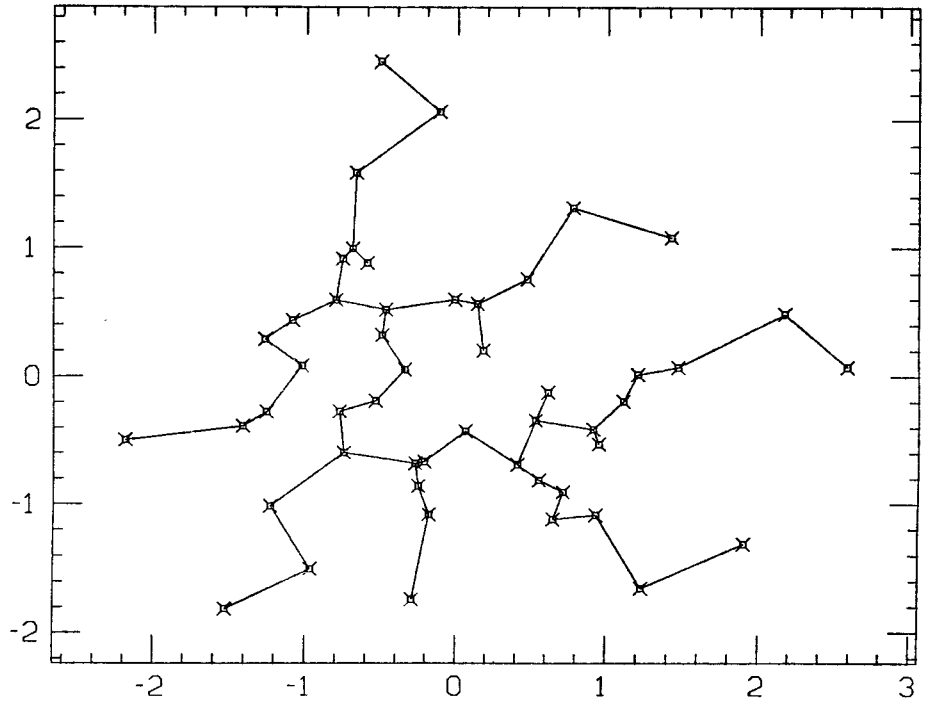

FIG. 1c. *Edges of* (1)MST (*Minimal Spanning Tree*) *graph shown as straight line segments.*

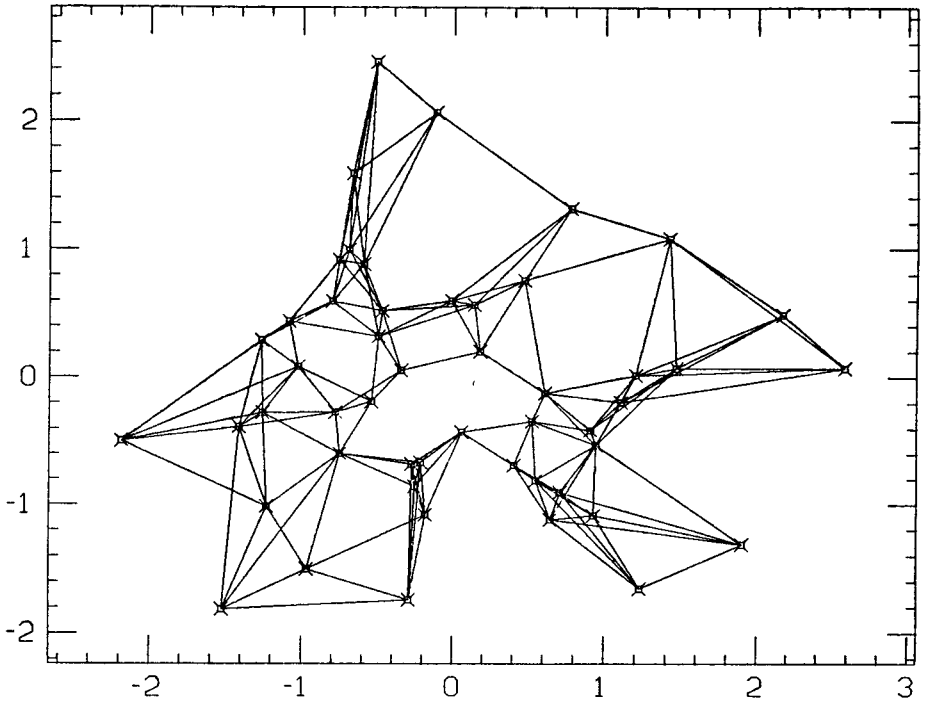

FIG. 1d: *Edges of* 5NN *graph shown as straight line segments.*

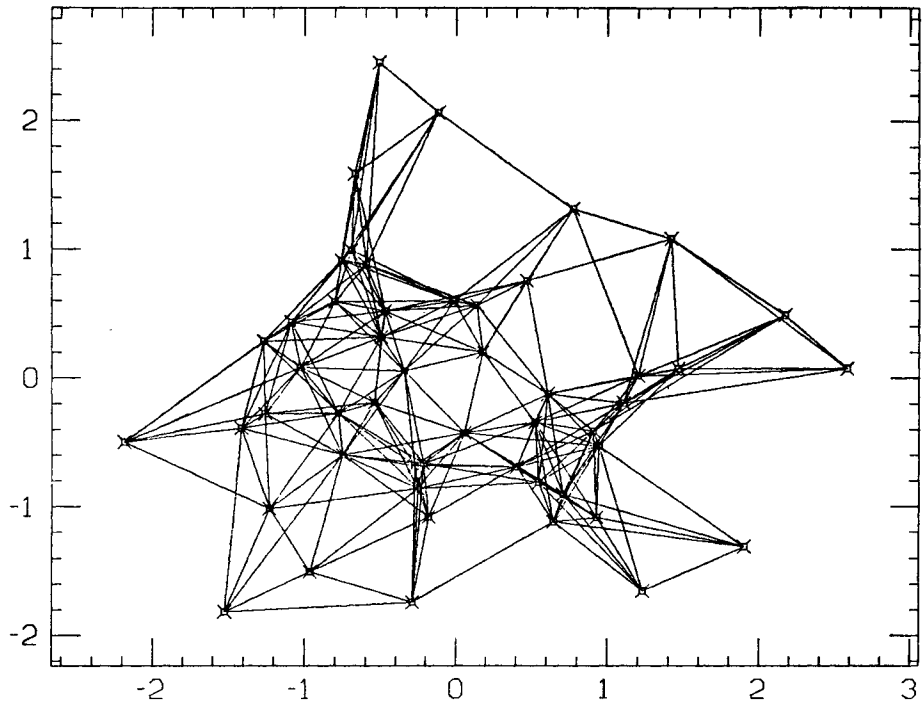

FIG. 1e: *Edges of* 5MST *graph shown as straight line segments.*

4. Association measures based on graph intersections. The Kendall measure of association τ can be defined in terms of the number of edges in the intersection of two graphs, that is, the number of edges the two graphs have in common. For univariate observations, consider the graph in which each unique pair of observations (i, j) define an edge if and only if $x_i < x_j$. Consider another such graph in which observations i and j define an edge if and only if $y_i < y_j$. Let Γ be the number of edges shared by these two graphs. This number will clearly tend to be large for a strong positive monotone relation between X and Y, and small for a negative one. In fact, Γ is related to Kendall's τ by $\tau = (\Gamma/N) - 1$.

In order to have power against more general alternatives, we use, instead of the graph defined above, the interpoint-distance-based (KMST or KNN) graphs defined in the previous section. As before, one graph G_x is defined over the X observations and a corresponding graph G_y is defined over the Y observations. The test statistic Γ is taken to be the number of edges in the intersection of the two graphs. The value of such a statistic will tend to be large if observations which are close in X also tend to be close in Y. Since KNN and KMST graphs ($K \ll N$) involve very few, if any, large distances, the test statistic will reflect only the extent to which closeness is correlated between the two spaces without regard to the correlation (or lack of it) for larger distances. This test is applicable to both univariate and multivariate observations and will be able to detect general relationships, i.e., those that are not necessarily one-one (non-monotone in the univariate case).

Measures of association based on graph intersections can be easily cast in the form of generalized correlation coefficients (1). Let

$$a_{ij} = \begin{cases} 1 \text{ if edge } (i,j) \in G_x, \\ 0 \text{ otherwise,} \end{cases} \tag{3}$$

$$b_{ij} = \begin{cases} 1 \text{ if edge } (i,j) \in G_y, \\ 0 \text{ otherwise,} \end{cases} \tag{4}$$

then clearly

$$\Gamma_1 = \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} b_{ij} \tag{5}$$

is the number of edges in the intersection.

5. Measures of prediction. The association measures defined above measure the extent to which close values in one space (X) are matched by close values in the other (Y), as reflected by the sample, without regard for the larger distances. It is this lack of regard for the larger distances that gives rise to power for general relations between the X and Y. The price paid for this generality is less power in those situations in which the relationship is one-one, many-one, or one-many. In these cases, including the larger distances in an association measure can greatly increase power.

In the many-one situation it is often important to know how well a (possibly vector valued) variable X can be used to make single valued predictions of a (possibly vector valued) variable Y without regard for how well X can be predicted from Y. The association measure defined in the previous section has sufficient generality but sacrifices some power. A more powerful test of association for such many-to-one relations would involve only small interpoint distances in the X space while making use of both small and large distances in the Y space, thus taking advantage of the (hypothesized) single valued dependence of Y and X.

As in the previous section, let G_x be an interpoint-distance-based (KMST or KNN) graph defined over the observations in the X space. For each observation i, rank the other observations $j \neq i$ in increasing order of their distance from i in the Y space. Let $R_i(j)$ be the position of the jth observation in the list of observations ordered in increasing Y distance from observation i. Defining a_{ij} as in (3) and $b_{ij} = R_i(j)$, we have as our proposed measure of prediction

$$\Gamma_2 = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} b_{ij} = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} R_i(j)$$

(6) or

$$\Gamma_2 = \sum_{(i,j) \in G_x} R_i(j).$$

Note that here rejection is for small values of Γ_2.

6. Distribution moments. The moments of $\Gamma(\pi)$ are determined by the scores a_{ij} and b_{ij}. It is straightforward, if laborious, to compute them directly from these scores. In this section, we present the first two (central) moments for the scores used in (5) and (6) in terms of easily obtained parameters of the corresponding graphs. These results are derived in the Appendix.

Consider first the association measures based on graph intersections, (3)–(5). Let e_x and e_y be the number of edges, respectively, in the two graphs. (Since G_x and G_y are spanning subgraphs of the complete graph, they both contain N nodes as does their intersection.) For the case of KMST graphs, the number of edges is always $K(N-1)$. For KNN graphs, the number of edges will depend on the actual point scatter of the observations.

Define the degree d_i of node i to be the number of edges which it (along with some other node) defines, in other words, the number of edges "incident" upon it. The average degree of the nodes in a graph is related to the number of edges by

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{N} d_i = \frac{2e}{N}. \tag{7}$$

The second graph parameter determined by the set of node degrees used in calculating the second moment is the number of edge pairs that share a common node. This parameter, C, is related to the node degrees by

$$C = \sum_{i=1}^{N} d_i(d_i - 1)/2. \tag{8}$$

The first two (and, in fact, all of the) moments are determined by the set of node degrees

of the two graphs G_x and G_y.

Let $C_x(C_y)$ be the number of edge pairs that share a common node for G_x (respectively G_y). The first two moments of Γ_1 in (5) are:

$$E(\Gamma_1 \mid e_x, e_y) = \frac{2e_xe_y}{N(N-1)}, \tag{9}$$

$$\begin{aligned}\mathrm{Var}(\Gamma_1 \mid e_x, e_y, C_x, C_y) = {} & \frac{2e_xe_y}{N(N-1)}\left\{1 - \frac{2e_xe_y}{N(N-1)}\right\} + \frac{4}{N(N-1)(N-2)} \\ & \cdot\left[C_xC_y + \frac{\{e_x(e_x-1) - 2C_x\}\{e_y(e_y-1) - 2C_y\}}{N-3}\right].\end{aligned} \tag{10}$$

For the prediction measure (6), the moments are determined by the set of node degrees of G_x (through e_x and C_x) and two parameters of the matrix R whose elements are $R_i(j)$. Let

$$A_R = \textstyle\sum_{i=1}^{N}\sum_{j=1}^{N} R_i(j)R_j(i), \tag{11}$$

and

$$B_R = \textstyle\sum_{i=1}^{N}\{\sum_{j=1}^{N} R_j(i)\}^2. \tag{12}$$

Then

$$E(\Gamma_2 \mid e_x) = e_xN \tag{13}$$

and

$$\begin{aligned}\mathrm{var}(\Gamma_2 \mid e_x, C_x, A_R, B_R) = {} & \frac{e_x^2}{N-3}\left\{\frac{N(3N+1)}{3} + \frac{4(A_R - B_R)}{N(N-1)(N-2)}\right\} \\ & + \frac{e_x}{N-3}\left\{\frac{2(N-1)(N-4)A_R + 4B_R}{N(N-1)(N-2)} - \frac{N(N-1)(N+2)}{3}\right\} \\ & + \frac{2C_x}{(N-2)(N-3)}\left\{\frac{(N+1)B_R - 2(N-1)A_R}{N(N-1)} - \frac{N(3N-4)(N^2-1)}{12}\right\}.\end{aligned} \tag{14}$$

7. Asymptotic normality. In this section, we discuss conditions under which the permutation distributions of Γ_1 and Γ_2 are asymptotically normal. Our results have as their basis the work of Daniels (1944) on the permutation distribution of generalized correlation coefficients (2).

Daniels' proof of the asymptotic normality of (2) rests on two conditions on the scores a_{ij} and b_{ij}:

$$a_{ij} = -a_{ji} \quad \text{and} \quad b_{ij} = -b_{ji}, \tag{15}$$

$$\textstyle\sum_{i,j,k} a_{ij}a_{ik} \sim N^3 \quad \text{and} \quad \sum_{i,j,k} b_{ij}b_{ik} \sim N^3. \tag{16}$$

A careful reading of Daniels's proof reveals that (15) is used only to ensure that

$$\textstyle\sum_{ij} a_{ij} = \sum_{ij} b_{ij} = 0, \tag{17}$$

but the weaker condition (17) can always be assumed without loss of generality by simply centering the a_{ij} and b_{ij}. Moreover, Daniels's arguments remain valid if (16) is replaced by the weaker condition

$$\lim_{N\to\infty} \frac{(\sum_{ijkl} a_{ij}a_{ik}a_{il})^2}{(\sum_{ijk} a_{ij}a_{ik})^3} = 0, \tag{18}$$

with a similar condition on the b_{ij}.

With the scores a_{ij} and b_{ij} defined in (3) and (4), conditions (16) and (18) become

$$\textstyle\sum_{i=1}^{N} d_i^2 \sim N^3 \quad \text{as} \quad N \to \infty \tag{19}$$

(20) and $$\lim_{N\to\infty} \frac{(\sum_{i=1}^{N} d_i^3)^2}{(\sum_{i=1}^{N} d_i^2)^3} = 0$$

for both G_x and G_y. These conditions place restrictions on the topology of the graphs G_x and G_y.

Condition (19) requires the graphs to be dense. For example, (19) can be insured if the degree of each node grows linearly with N. Choosing K proportional to N, KNN and KMST graphs meet this requirement. Since the number of edges of a graph is related to the average node degree by (7), this would require e_x and e_y to grow quadratically in N, or linearly in $N(N-1)/2$, which is the number of edges in the complete graph.

Condition (20) permits Daniels proof to be applied to sparse graphs as well. For example, consider the case in which the maximum node degree is bounded by a constant independent of N. This would obtain for KNN and KMST graphs based on Euclidean or more general (e.g., qth power) distances if K is held constant with increasing N. Sphere packing properties of p-space (Leech and Sloane, 1971) imply that the maximum degree in this case is bounded by a constant depending only on p. Under these conditions e_x and e_y grow only linearly with N so that G_x and G_y link a vanishingly small fraction of the $N(N-1)/2$ node pairs as N becomes large, resulting in very sparse graphs. Clearly the numerator of (20) grows as N^2 while the denominator grows as N^3, thereby insuring the limiting normal distribution.

There are sequences of sparse graphs for which (20) is violated and the limiting distributions are not normal. Such a sequence could be constructed of "fan" graphs which have $N-1$ nodes with degree one and one node with degree $N-1$. For these graphs, the numerator and denominator of (20) both grow as N^6 and the condition is violated. The permutation distribution of the number of edges in the intersection of these G_x with any G_y puts equal probability on each node degree of G_y. Since a sequence of G_y's can be chosen with maximal degree bounded independent of N, such distributions cannot approach normality as N increases. In order to insure an asymptotic normal distribution for Γ_1 (or Γ_2) as well as have power for the corresponding tests, it is important that as N increases, the increasing number of edges of the spanning subgraph be distributed among the node pairs in a way that prohibits a (too rapidly) decreasing fraction of the nodes from defining a (too rapidly) increasing fraction of the edges.

The asymptotic normality of the distribution of Γ_2 depends completely on G_x satisfying (19) or (20). The scores $b_{ij} = R_i(j)$ can easily be seen to satisfy (20) since the numerator grows as N^{14} while the denominator grows as N^{15}.

8. Previous work. Generalizations to a multivariate setting of nonparametric measures of monotone association have mostly been based on projection rank analogs of the sample covariance matrix. That is, the coordinate values of each observation are replaced by their corresponding ranks in the projection on each coordinate (or possibly some transformation of the ranks) and the covariance matrix computed. Tests for association can then be based on this matrix; see Puri and Sen (1971) for a survey. The earliest use (in a limited setting) of measures of association based on interpoint-distance-based-graphs appears to have been in epidemiology; see Knox and Braithwaite (1963) and references therein. For these studies, the X space (in the language of this paper) was taken to be a (two-dimensional) map of the locations of the onset of a particular disease. The Y space was the (one-dimensional) time of disease onset. A high association between positions in space and time is evidence for epidemicity (that is, real or apparent contagion).

The graphs used in these studies were all distance threshold graphs. That is, each point is connected to all points within a prespecified distance. For example, Knox (1964) used a map distance threshold of one kilometer and a time distance threshold of 59 days. Knox speculated that if the graphs were sufficiently sparse (number of edges very small compared to the complete graph) the number of observations so adjacent in both space and time would be approximately a Poisson variable.

Barton and David (1966) introduced the graph theoretic specification of the problem,

derived the first two moments (9) and (10), and set up a methodology for computing higher moments. They were also the first to study the asymptotic permutation distribution of statistics based on graph intersections. They show for the case of e_x and e_y *both* growing linearly in N and the graphs G_x and G_y having a low degree of connectiveness relative to their respective number of edges (such graphs are called "incoherent"—see Barton and David, 1966, Section 3.7, for precise definitions), that Γ_1 tends toward a Poisson distribution as N increases. Abe (1969) points out that there is no valid ground, under the Barton and David conditions, for applying the central limit theorem for a Poisson variable, and thereby deducing asymptotic normality in such situations. However, for the special case of distance threshold graphs considered by Knox, the theory of U-statistics (Hoeffding, 1948) can be applied to directly deduce asymptotic normality under very general conditions. Unfortunately, statistics based on the intersection of KNN and KMST graphs are not U-statistics, and we must appeal to our generalization of Daniels's (1944) conditions (18) and (20) for the case of sparse graphs.

Abe (1969) shows asymptotic normality for statistics based on graph intersections provided e_x and e_y satisfy $N^{5/2} < e_x e_y$ and $(e_x e_y / N^3)^r \to 0$ for all $r > 2$ as $N \to \infty$. This result is not applicable to sparse graphs (e_x and/or $e_y \sim N$), or dense graphs (e_x and/or $e_y \sim N^2$). It can, however, be applied to some cases between these two extremes. For example, $e_x e_y \sim N^\alpha$ for $2.5 < \alpha < 3$.

There has been considerable work on the asymptotic distribution of statistics similar to (2) under a variety of conditions; see, for example, Jogdeo (1968), Brown and Kildea (1978), and Shapiro and Hubert (1979). However, the asymptotics introduced by Daniels (1944) appear to provide the most general conditions for the asymptotic normality of our statistics Γ_1 and Γ_2.

Although not suggesting graphs, Mantel and Valand (1970) discussed using the direct correlation between the interpoint distances as a measure of association between two multivariate spaces. As discussed above, such measures are mainly sensitive to one-one relationships. They do mention, however, the possibility of weighting the distances so as to increase the influence of the smaller distances.

Shepard and Carrol (1965) discuss using interpoint distances in forming measures of prediction for multivariate observations in the context of parametric mapping and multidimensional scaling. Simon (1977) and Weier and Basu (1978) discuss measures of association specifically directed at the hypothesis of total independence of the variables.

9. Notes on applications.

NOTE 1. A situation for which Γ_2 is appropriate is testing goodness-of-fit. Suppose the relation between $\mathbf{Y}$ and $\mathbf{X}$ can be expressed as

$$\mathbf{Y} = f(\mathbf{X}) + \varepsilon, \tag{21}$$

where f is a single-valued function and

$$E(\varepsilon \mid \mathbf{X} = \mathbf{x}) = 0. \tag{22}$$

The "no correlation" null hypothesis is

$$H_0: f(\mathbf{X}) = \text{constant}. \tag{23}$$

If instead our hypothesis is $f(\mathbf{X}) = g(\mathbf{X})$, where $g(\mathbf{X})$ is a specified function, then we may test whether the function $g(\mathbf{X})$ exhausts the predictive relationship of $\mathbf{Y}$ on $\mathbf{X}$ or whether a more elaborate (or perhaps different) model might be in order. This is equivalent to testing that there is no association between $Y - g(\mathbf{X})$ and $\mathbf{X}$. Note that association measures directed at one-to-one relationships are not suitable for this application. The general association measure Γ_1 in (5) is suitable but is less sensitive than Γ_2 because it does not use to advantage the (presumed) single-valued relation of $\mathbf{Y}$ on $\mathbf{X}$.

NOTE 2. The two-sample runs test (Wald and Wolfowitz, 1940) and its multivariate generalization (Friedman and Rafsky, 1979) are special cases of Γ_1. The graph G_x is taken

to be the MST over the values of the (pooled) observations. The graph G_y is defined over the categorical variable that labels the sample identity of each observation: all points are connected to all others with the same label. With these definitions for G_x and G_y, the number-of-runs test statistic is $R = N - \Gamma_1 + 1$. When cast in this framework, it is easy to see how to generalize both tests to cases of more than two samples: the categorical sample identity variable simply takes on more than two values. The asymptotic null distribution for such a test is guaranteed to be normal by (20) and its first two moments are easily derived from (9) and (10).

NOTE 3. As with all methods based on interpoint distances or dissimilarities, the definition of distance (or dissimilarity) is important. KNN and KMST graphs depend on the sorted order of the edge weights of the complete graph. Different distance (or dissimilarity) measures can result in a different order. For the case of a single variable X, $d_{ij} = |X_i - X_j|$ is a natural definition of distance. For vector valued variables, there are a variety of definitions possible. Weighted Euclidean (or more general qth power norms) are often used with the weight for each variable chosen by the researcher to reflect its presumed importance within the context of the problem. Although KNN and KMST graphs are resistant to moderate changes in the coordinate weights, they are not fully robust, and the power of these tests for particular situations can depend on specific choices. Note, however, that since these statistics depend only on the order of the interpoint distances rather than the actual values, they are robust to (possibly large) changes involving only a small fraction of the observations.

NOTE 4. In addition to choice of distance (dissimilarity) measure, another important choice is the size (number of edges) of the spanning subgraphs G_x and G_y. In some cases (e.g., graphs on a single categorical variable), this is fixed by the nature of the graph definition (e.g., all points linked to all other points with the same value). In the case of KNN or KMST graphs, the size of the graph is determined by the choice of K. The best choice will depend upon the particular situation and there are, as yet, no specific guidelines. As the sample size becomes large, it is unlikely that the best choice would involve having both graphs sparse; that is, both e_x and e_y growing linearly with the sample size. Similarily, it is unlikely that both graphs should be dense, e_x and e_y growing as N^2. Choices between these two extremes are likely to be best. For example, the Wald-Wolfowitz runs test takes one graph (sample identity) to be dense and the other (MST) to be sparse.

10. An example. In order to study the effectiveness of our approach, we apply the statistic Γ_2 to the goodness-of-fit problem discussed in the previous section. For each experiment, a sample of 100 $(\mathbf{X}, Y)$ pairs $(\mathbf{X} \in R^p, Y \in R^1)$ were drawn according to the model

$$Y = f(\mathbf{X}) + \varepsilon \tag{25}$$

with

$$f(\mathbf{X}) = \sum_{i=1}^{p} (X_i - 1/2)^2, \tag{26}$$

each X_i drawn from a uniform distribution over the interval [0, 1] and each ε drawn from a standard normal distribution. For each experiment, the hypothesis (23) was tested using the reference distribution given by (13), (14) and normality. The graph G_x was taken to be the 5MST defined over the $\mathbf{X}$ points using simple equal weighted Euclidean distance. (Other, similar choices for G_x lead to nearly the same results.) A run of 100 experiments was performed for each p, $1 \leq p \leq 10$. The fraction of experiments in each run for which the value of Γ_2 was less than the five percent point of the reference distribution was used as an estimate of power at five percent significance.

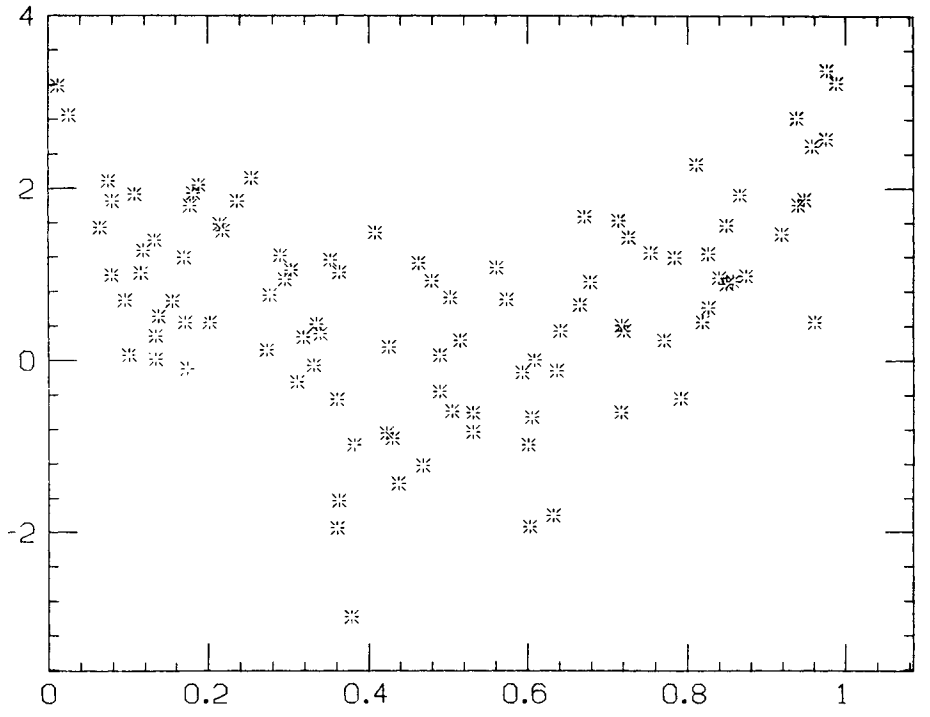

FIG. 2. *One hundred points,* $Y = 10(X - \frac{1}{2})^2 + \varepsilon$ vs. X *with* ε iid *standard normal.*

In addition, the data of each experiment were fitted to the model

$$g(\mathbf{X}) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i \tag{27}$$

and hypothesis (23) was tested using $f(\mathbf{X}) = Y - g(\mathbf{X})$. The parameters β_i, $0 \le i \le p$, were estimated by least squares.

Figure 2 shows a plot of y_i vs. x_i, $1 \le i \le 100$, for one of the experiments (the first) from the 100 experiments comprising the run for $p = 1$. Table 1 shows a summary of the results for all ten runs. Included in the table for each run is the ratio of the standard deviation of $f(\mathbf{X})$ (26) to that of ε.

Also included in Table 1 are results for a second study where we chose the null case ($f(\mathbf{X}) = 0$) to test the adequacy of the normal approximation. The number of experiments for which Γ_2 was less than the 5% point of the reference distribution is displayed.

The ability of a test based on Γ_2 to reject (23) for these data is seen to decrease with increasing dimension. For low to moderate dimension ($p \le 6$), the power is reasonably high.

11. Summary. An association measure (Γ_1) has been developed that is sensitive to general alternatives. A measure of one-way single valued association (Γ_2) has also been developed that can be used as well for testing goodness-of-fit. These measures are based on interpoint distance graphs (spanning subgraphs of the complete graph) in which the observations are nodes. Two particular spanning subgraphs—the KNN and KMST—have been proposed as being especially suitable for this purpose. Since these graphs, as well as the matrix $R_i(j)$, can be defined for vector valued observations, all results are applicable in multivariate settings. An extension to Daniels's (1944) theorem on the asymptotic normality (of the permutation distribution) of generalized correlation coefficients is used to derive the asymptotic null distributions of these statistics.

A FORTRAN program implementing the tests described is available from either author.

TABLE 1
Simulation experiment with $Y = f(\mathbf{X}) + \varepsilon$, $f(\mathbf{X}) = \sum_{i=1}^{p} (X_i - 1/2)^2$, $\mathbf{X} \sim U[0, 1]^p$, and $\varepsilon \sim N(0, 1)$. 100 observations per experiment, 100 experiments per run (value of p)

p	$\left[\frac{\text{Var}\{f(\mathbf{X})\}}{\text{Var}(\varepsilon)}\right]^{1/2}$	Power at 5% significance		
		Y vs. $\mathbf{X}$	$Y - \hat{\beta}_0 - \sum_{i=1}^{p} \hat{\beta}_i X_i$ vs $\mathbf{X}$	$f(\mathbf{X}) = 0$
1	0.72	1.0	1.0	0.07
2	1.02	1.0	1.0	0.06
3	1.26	1.0	1.0	0.07
4	1.48	0.99	0.98	0.02
5	1.67	0.97	0.90	0.08
6	1.81	0.87	0.70	0.06
7	1.98	0.76	0.49	0.10
8	2.01	0.67	0.29	0.04
9	2.25	0.59	0.22	0.03
10	2.32	0.49	0.20	0.09

APPENDIX: MOMENT CALCULATIONS

In this section, we present the detailed calculation of the first two moments of the permutation distribution (2) for the statistics Γ_1, defined in (3)–(5) and Γ_2, defined in (6). The results are presented in (9), (10), (13), and (14).

We first derive $E(\Gamma_1)$, (9). Label the *edges* of G_x arbitrarily and define the indicator variable z_i as

$$z_i = \begin{cases} 1 \text{ if } i \in G_y, & 1 \le i \le e_x, \\ 0 \text{ otherwise.} \end{cases} \tag{A1}$$

Then

$$\Gamma_1 = \sum_{i=1}^{e_x} z_i, \tag{A2}$$

and

$$E(\Gamma_1) = \sum_{i=1}^{e_x} E(z_i) = e_x \Pr(z_i = 1). \tag{A3}$$

The quantity $\Pr(z_i = 1)$ is the probability that a randomly selected edge is in G_y. This is just the ratio of the number of edges in G_y to the total possible number of edges (in the complete graph). Thus,

$$\Pr(z_i = 1) = \frac{e_y}{N(N-1)/2} \tag{A4}$$

so that from (A3) we have the result (9).

Consider next $\text{Var}(\Gamma_1)$, (10). From (A2) one has

$$\text{Var}(\Gamma_1) = \sum_{i=1}^{e_x} \text{Var}(z_i) + 2 \sum_{i<j} \text{cov}(z_i, z_j). \tag{A5}$$

Now, $\text{Var}(z_i) = p(1-p)$ so that

$$\sum_{i=1}^{e_x} \text{Var}(z_i) = e_x p(1-p) \tag{A6}$$

with $p = \Pr(z_i = 1)$ given by (A4). The second summand in (A5) can be simplified using

$$\text{Cov}(z_i, z_j) = E(z_i z_j) - p^2, \tag{A7}$$

so that it can be expressed as

$$2 \sum_{i<j} \text{Cov}(z_i, z_j) = 2 \sum_{i<j} E(z_i z_j) - e_x(e_x - 1)p^2. \tag{A8}$$

The quantity $E(z_i z_j)$ can have two distinct values which depend upon whether or not the *edge pair* (i, j) share a common defining node. Consider first the situation in which this is the case. Then $E(z_i z_j)$ is just the probability that a randomly selected edge pair shares a common node in G_y. This is just the ratio of the number of edge pairs that share a common node in G_y, C_y given by (8), to the total number of edge pairs sharing a node (in the complete graph). From (8), with $d_i = N - 1$ for the complete graph, this latter quantity is just $N(N-1)(N-2)/2$. Therefore,

$$\text{(A9)} \qquad E\{z_i z_j | (i,j) \text{ share a common node in } G_x\} = \frac{2\,C_y}{N(N-1)(N-2)}.$$

By a similar argument, $E(z_i z_j)$ for the case in which the edge pair (i, j) does not share a common node is just the ratio of the number of edge pairs that do not share a common node in G_y to the corresponding total number (in the complete graph). The numerator of this ratio is just the difference between the total number of edge pairs in G_y, $e_y(e_y - 1)/2$, and C_y. Similarily, the number of edge pairs in the complete graph that do not share a node in common is the difference between the total number of edge pairs in the complete graph, $\{N(N-1)/2\}\{N(N-1)/2 - 1\}/2$, and $N(N-1)(N-2)/2$. This latter difference reduces to $N(N-1)(N-2)(N-3)/8$, giving the result

$$\text{(A10)} \qquad E\{z_i z_j | (i,j) \text{ do not share common node in } G_x\} = \frac{4\,e_y(e_y-1) - 8\,C_y}{N(N-1)(N-2)(N-3)}.$$

The number of times the summand corresponding to (A9) appears in the double sum (A5) is simply C_x and that corresponding to (A10) is $e_x(e_x - 1)/2 - C_x$. Combining (A4)-(A10), one has the result (10).

We now derive the first two moments of Γ_2, defined in (6), as presented in (11)-(14). For notational convenience, we take $R_i(i) = 0$, $1 \le i \le N$, and consider the $N \times N$ matrix $R = [R_i(j)]$ with $R_i(j)$ as its (i, j) entry.

Consider $E(\Gamma_2)$ given in (13). From (6), we have

$$\text{(A11)} \qquad E(\Gamma_2) = \textstyle\sum_{(i,j)\in G_x} E[R_i(j)] = 2\,e_x E[R_i(j)]$$

$$\text{(A12)} \qquad E[R_i(j)] = \frac{1}{N(N-1)} \textstyle\sum_{i=1}^{N} \sum_{j=1}^{N} R_i(j) = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N-1} j = \frac{N}{2}.$$

Combining (A11) and (A12) gives the result (13). We next derive $\mathrm{Var}(\Gamma_2)$, given by (14). From (6)

$$\text{(A13)} \qquad \mathrm{Var}(\Gamma_2) = \textstyle\sum_{(i,j)\in G_x} \mathrm{Var}[R_i(j)] + \sum_{(i,j)\in G_x} \sum_{\substack{(k,l)\in G_x \\ (k,l)\neq(i,j)}} \mathrm{cov}[R_i(j), R_k(l)].$$

Now,

$$\text{(A14)} \qquad \mathrm{Var}[R_i(j)] = E[R_i^2(j)] - \{E[R_i(j)]\}^2,$$

with

$$\text{(A15)} \qquad E[R_i^2(j)] = \frac{1}{N-1} \textstyle\sum_{j=1}^{N-1} j^2 = \frac{N(2N-1)}{6},$$

so that combining (A12), (A14) and (A15) one has

$$\text{(A16)} \qquad \mathrm{Var}[R_i(j)] = \frac{N(N-2)}{12}.$$

The summand of the second sum (A13) can be expressed, using (A12), as

$$\text{(A17)} \qquad \mathrm{Cov}[R_i(j), R_k(l)] = E[R_i(j) R_k(l)] - \frac{N^2}{4}.$$

The quantity $E[R_i(j)R_k(l)]$ can take on six distinct values for *ordered* edge-pairs defined by nodes (i, j) and (k, l), $(i, j) \neq (k, l)$. These values result from the following orderings: Case 1: (i, j), (i, k) $j \neq k$; Case 2: (i, j), (k, j) $i \neq k$; Case 3: (i, j), (j, k) $i \neq k$; Case 4: (i, j), (k, i) $j \neq k$; Case 5: (i, j), (j, i); Case 6: (i, j), (k, l) all distinct. We compute the value of $E[R_i(j)R_k(l)]$ for each case in turn.

CASE 1.

$$E[R_i(j)R_i(k)] = \frac{1}{N(N-1)(N-2)} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} R_i(j)R_i(k) = \frac{N(3N-1)}{12}. \tag{A18}$$

CASE 2.

$$\begin{aligned} E[R_i(j)R_k(j)] &= \frac{1}{N(N-1)(N-2)} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} R_i(j)R_k(j) \\ &= \frac{B_R}{N(N-1)(N-2)} - \frac{N(2N-1)}{6(N-2)}, \end{aligned} \tag{A19}$$

with B_R given by (12).

CASE 3.

$$\begin{aligned} E[R_i(j)R_j(k)] &= \frac{1}{N(N-1)(N-2)} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} R_i(j)R_j(k) \\ &= \frac{N^2(N-1)}{4(N-2)} - \frac{A_R}{N(N-1)(N-2)} \end{aligned} \tag{A20}$$

with A_R given by (11).

CASE 4.

$$\begin{aligned} E[R_i(j)R_k(i)] &= \frac{1}{N(N-1)(N-2)} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} R_i(j)R_k(i) \\ &= \frac{N^2(N-1)}{4(N-2)} - \frac{A_R}{N(N-1)(N-2)}. \end{aligned} \tag{A21}$$

CASE 5.

$$E[R_i(j)R_j(i)] = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} R_i(j)R_j(i) = \frac{A_R}{N(N-1)}. \tag{A22}$$

CASE 6.

$$\begin{aligned} E[R_i(j)R_k(l)] &= \frac{1}{N(N-1)(N-2)(N-3)} \sum_{\substack{i=1 \\ k\neq i, l\neq i, k\neq j, l\neq j}}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} R_i(j)R_k(l) \\ &= \frac{1}{N(N-1)(N-2)(N-3)} \Big[\sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} R_i(j)R_k(l) \\ &\quad - \sum_{\substack{i=1 \\ i\neq j\neq l}}^{N} \sum_{j=1}^{N} \sum_{l=1}^{N} \{R_i(j)R_i(l) + R_i(j)R_j(l)\} \\ &\quad - \sum_{\substack{i=1 \\ i\neq j\neq k}}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \{R_i(j)R_k(i) + R_i(j)R_k(j)\} \Big]. \end{aligned} \tag{A23}$$

Note that the triple sums of each of the four products have been evaluated in Cases 1

through 4 above. Therefore,

$$E[R_i(j)R_k(l)] = \frac{1}{N(N-1)(N-2)(N-3)} \cdot \left\{ (N\textstyle\sum_{j=1}^{N-1} j)^2 - \frac{3N^3(N-1)^2}{4} + A_R - B_R + \frac{N^2(N-1)(2N-1)}{6} \right\},$$

which after some simplification becomes

$$E[R_i(j)R_k(l)] = \frac{N(3N^2-6N+1)}{12(N-3)} + \frac{A_R - B_R}{N(N-1)(N-2)(N-3)}. \tag{A24}$$

This completes the calculation of the six distinct values for $E[R_i(j)R_k(l)](i,j) \neq (k,l)$.

Of the $(2e_x)^2$ ordered edge pairs $\{(i,j),(k,l)\}$ in G_x, there are $2e_x$ for which $i = k$ and $j = l$, $2C_x$ for each of the Cases 1–4, $2e_x$ for Case 5, and $4e_x^2 - 8C_x - 4e_x$ for Case 6. These numbers along with (A18)–(A24) and (A17) permit the evaluation of the double sum in (A13). Then from (A16) and (A13)—with some algebraic simplification—we have the result (14).

Acknowledgment. We thank Chen-hsin-Chen for verifying the result (14).

REFERENCES

ABE, O. (1969). A central limit theorem for the number of edges in the random intersection of two graphs. *Ann. Math. Statist.* **40** 144–151.

BARTON, D. E. and DAVID, F. N. (1966). The random intersection of two graphs. In *Research Papers in Statistics*, (F. N. David, Ed.). Wiley, New York.

BROWN, B. M. and KILDEA, D. G. (1978). Reduced U-statistics and the Hodges-Lehmann estimator. *Ann. Statist.* **6** 828–835.

DANIELS, H. E. (1944). The relation between measures of correlation in the universe of sample permutations. *Biometrika* **33** 120–135.

FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

JOGDEO, K. (1968). Asymptotic normality in nonparametric methods. *Ann. Math. Statist.* **39** 905–922.

KENDALL, M. G. (1962). *Rank Correlation Methods.* Griffin, London.

KNOX, G. (1964). Epidemiology of childhood leukemia in Northumberland and Durham. *British J. Prev. Soc. Med.* **18** 17–24.

KNOX, G. and BRAITHWAITE, F. (1963). Cleft lips and palates in Northumberland and Durham. *Archs. Dis. Childh.* **38** 66–70.

LEECH, J. and SLOANE, N. J. A. (1971). Sphere packings and error correcting codes. *Canad. J. Math.* **23** 718–745.

MANTEL, N. and VALAND, R. S. (1970). A technique of nonparametric multivariate analysis. *Biometrics* **27** 547–558.

PURI, M. L. and SEN. P. K. (1971). *Nonparametric Methods in Multivariate Analysis.* Wiley, New York.

SHAPIRO, C. P. and HUBERT, L. (1979). Asymptotic normality of permutation statistics derived from weighted sums of bivariate functions. *Ann. Statist.* **7** 788–794.

SHEPARD, R. N. and CARROLL, J. D. (1966). Parametric representation of nonlinear data structures. In *Multivariate Analysis*, (P. R. Krishnaiah, Ed.). Academic, New York.

SIMON, G. (1977). A nonparametric test of total independence based on Kendall's tau. *Biometrika* **64** 277–282.

WALD, A. and WOLFOWITZ, J. (1940). On a test whether two samples are from the same population. *Ann. Math. Statist.* **11** 147–162.

WIERER, D. R. and BASU, A. P. (1978). A trivariate generalization of Spearman's rho. University of Missouri, Columbia Technical Report No. 78, Dept. of Statistics.

STANFORD LINEAR ACCELERATOR CENTER
STANFORD UNIVERSITY
BOX 4349
STANFORD, CALIFORNIA 94305

GEMNET SOFTWARE CORPORATION
2175 W. STADIUM BLVD.
ANN ARBOR, MICHIGAN 48106