TECHNICAL REPORT NO. 968

OCTOBER 1996

Asymptotic Properties of Random Interval Graphs and Their Use in Cluster Analysis

by

Erhard Godehardt and Bernard Harris
ASYMPTOTIC PROPERTIES OF
RANDOM INTERVAL GRAPHS

AND THEIR USE IN CLUSTER ANALYSIS

Erhard Godehardt
Heinrich Heine Universität, Düsseldorf

Bernard Harris
University of Wisconsin, Madison

This research was partially supported by the Deutsche Forschungsgemeinschaft - DFG
Grants Go 490/4-2 and Go 490/4-3.
1. INTRODUCTION AND SUMMARY

Let $x_1, x_2, ... x_n$ be vectors in $S$, a subset of $E_k$. The objective is to divide these $n$ vectors into classes $C_1, C_2, ..., C_p$, where $C_i \cap C_j = \emptyset$, $i \neq j$, and $\cup C_i \subseteq S$. These classes are known as clusters. The experimenter seeks to place "similar" vectors in the same class and "dissimilar" vectors into different classes. To accomplish this, a "dissimilarity" measure $\delta$ and a threshold $d$ are defined. In many applications, $\delta$ satisfies the mathematical postulates of a metric. In other instances, $\delta$ is so defined that it may violate the triangle inequality. Then, if $\delta(x, y) < d$, $x$ and $y$ are considered to be similar and are placed in the same cluster. We will assume that $x_1, x_2, ..., x_n$ are realizations of the random variables $X_1, X_2, ..., X_n$ respectively.

To progress further, we need to introduce the notion of a graph. A graph, $G_n$, with $n$ vertices is defined as follows. $G_n = (V, E)$, where $V$ is a set with $|V| = n$, and $E$ is a set of unordered pairs of elements of $V$. Hence $0 \leq |E| \leq \binom{n}{2}$. With no loss of generality, we can assume $V = \{1, 2, ..., n\}$. The elements of $V$ will be referred to as the vertices of the graph $G_n$ and the elements of $E$ will be called the edges of $G_n$. There is a natural correspondence between graph theory and cluster analysis. Graphs provide a means of representing a binary relation and clusters are a collection of subsets of the data such that the elements in a given cluster are related to each other. Hence, the use of graph theoretical methods to study clustering processes is inherently reasonable. For an extensive discussion of this relationship, the reader is referred to the book by H. H. Bock (1974). The article by P. Arabie and L. J. Hubert in the book edited by P. Arabie, L. J. Hubert and G. De Soete (1996) is also relevant in this context. If the graphs are obtained from realizations of random variables, they are realizations of random graphs. There is an extensive literature on random graphs and the following two models for random graphs are frequently employed.
**Bernoulli (or binomial) model.** A sequence of \( \binom{n}{2} \) independent and identically distributed Bernoulli trials is conducted. If the \( i \)th Bernoulli trial is a "success", then the \( i \)th pair of elements of \( V \) is placed in \( E \). In this way, a random graph is constructed. Note that \( |E| \) is a random variable. This model was introduced by E. N. Gilbert (1959).

**Hypergeometric model.** \( M \) of the \( \binom{2}{n} \) possible elements of \( E \) are chosen at random without replacement (see P. Erdős and A. Rényi (1960)).

Both of these models are not specifically intended for use with real valued data and in fact would violate the triangle inequality if relationships between real valued observations were given in terms of a metric. Therefore probability models for random interval graphs (or random coincidence graphs) are more suited for studying the homogeneity of real valued data (or vector valued data). In the sequel, we will restrict our attention to random interval graphs.

A fundamental concern in the use of clustering algorithms, is that the computational technique finds similarities, when only random effects are present. In the present paper, we will use graph theoretic methods to develop tests of the hypothesis that the clusters are chance occurrences against the hypothesis that the clusters are real (causal).

Accordingly, the null hypothesis is formulated as follows. We assume that the random variables \( X_1, X_2, \ldots, X_n \) are independently, identically distributed by the distribution function \( F_X(x) \), where \( F_X(x) \) is assumed to be known. Then, with this assumption, the probability distributions of various characteristics of an induced interval graph can
be calculated and utilized to see if the clustering obtained is compatible with these probability distributions. In particular, we obtain asymptotic probability distributions for the number of edges, the number of triangles, and in general the number of complete subgraphs of order m; the number of isolated vertices, the number of maximal complete subgraphs of order m, and the degrees of the vertices.

With only rare exceptions, the exact distributions are complicated and cumbersome to utilize. Consequently, much of the paper is devoted to determining the asymptotic behavior of the probability distributions of these characteristics.

There is an extensive literature on clustering methods. The reader is referred to the books of H. H. Bock (1974), J. van Ryzin (1977), J. A. Hartigan (1975), E. Godehardt (1990) for extensive discussions of these various methods. There is also a substantial number of related results in the combinatorial literature and in the statistical literature. W. Eberl and R. Hafner (1971) investigate the distribution of the number of edges (which they refer to as coincidences) and in R. Hafner (1972), these results are extended to multiple coincidences. In these papers, they require a symmetry condition, which is not needed for the present methodology (however, in the present work their symmetry condition does hold for edges). Also, for the case of multiple coincidences, Hafner assumes that the graph is connected and can be assumed to be a tree, whereas here complete subgraphs are considered. H. Maehara (1990) studied similar questions, but restricted his investigation to the case of the uniform distribution on the circle. In S. Jammalamadaka and X. Zhou (1988) and S. Jammalamadaka and S. Janson (1986), goodness-of-fit statistics based on the number of edges are studied.

2. RANDOM INTERVAL GRAPHS

3
Theorem 1. The probability that $X_1, X_2, ..., X_m$, or equivalently, the vertices 
$\{1, 2, ..., m\}$, form a $K_{m,d}$, $m = 1, 2, ..., n$ is

$$P\left(\max_{1 \leq i \leq m} X_i - \min_{1 \leq i \leq m} X_i \leq d\right)$$

$$= m \int_{-\infty}^{\infty} [F(x+d) - F(x)]^{m-1} f(x) \, dx.$$  \hspace{1cm} (1)

Proof: Since $m$ specified vertices form a $K_{m,d}$ whenever the greatest distance between any pair does not exceed $d$, the left hand side of (1) is equivalent to this condition. Assume $x_1$ is the minimum realized value. Then, given $X_1 = x_1$, we must have $X_2, ..., X_m \leq x_1 + d$. Since there are $m$ equally likely choices for the minimum, theorem 1 is established.

Note that if $m=1$, the right hand side of (1) is unity.

Theorem 2. The probability that $X_1, X_2, ..., X_m$ form a $K_{m,d}^*$, $m=2, ..., n$, is

$$P\{(1, 2, ..., m) \text{ is a } K_{m,d}^*\} = m(m-1) \int_{-\infty}^{\infty} \int_{x_1}^{x_1+d} [F(x_2) - F(x_1)]^{m-2}$$

$$\cdot [1 - F(x_1 + d) + F(x_2 - d)] f(x_1) f(x_2) \, dx_2 \, dx_1.$$ \hspace{1cm} (2)

Proof. A specified subset of $m$ vertices form a $K_{m,d}^*$, when these $m$ vertices form a $K_{m,d}$ and if $y$ is any other vertex, either $y > \min_{1 \leq i \leq m} x_i + d$ or $y < \max_{1 \leq i \leq m} x_i - d$.

Given $X_1 = x_1$ is the minimum and $X_2 = x_2$ is the maximum of $x_1, x_2, ..., x_m$ and $x_2 - x_1 < d$, the probability that the $m$ vertices form a $K_{m,d}^*$ is $[F(x_2) - F(x_1)]^{m-2}$.
Let $X_1, X_2, ..., X_n$ be independent, identically distributed random variables with common distribution $F_X(x)$, where $F_X(x)$ is absolutely continuous with respect to Lebesgue measure. Let $f_X(x)$ denote the corresponding probability density function. We define a family of graphs $G_{d,n} = (V_n, E_{d,n})$ as follows: For $d > 0$ and $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$, define the vertices of the graph as $V = \{1, 2, ..., n\}$. The pair $(i, j)$ is in $E_{d,n}$, the edges of the graph, whenever $|x_i - x_j| < d$. Since $x_1, x_2, ..., x_n$ are realizations of random variables, the various characteristics of the graphs, such as the number of edges are realizations of random variables and events, such as, the graph is connected, are realizations of random events. We now introduce the definitions and notation that will be employed in the sequel.

For $m \geq 1$, $K_{m,d}$ is a complete subgraph of order $m$, if, for the specified subset $V_m \subset V$, with

$$|V_m| = m, \text{ all } \binom{m}{2} \text{ pairs of elements of } V_m \text{ are in } E_{d,n}. \text{ } K_{1,d} \text{ is a specified vertex.} \text{ } K_{m,d}^* \text{ is a maximal complete subgraph of order } m, \text{ if the vertices of } V_m \text{ form a complete subgraph of order } m \text{ and there is no complete subgraph of order } m+1 \text{ with } K_{m,d}^* \subset K_{m+1,d}, \text{ for any subset } V_{m+1} \text{ of } V, \text{ with } V_m \subset V_{m+1}. \text{ Thus } K_{1,d}^* \text{ is an isolated vertex.}$$

To simplify notation, we will denote $G_{d,n}$ by $G$, $V_n$ by $V$ and $E_{d,n}$ by $E$, when there is no ambiguity concerning $n$ or $d$. Similarly, we will drop the subscript $X$ from the probability density function and the cumulative distribution function.

We now proceed to obtain the probability that a specified set of $m$ vertices form a $K_{m,d}$ or a $K_{m,d}^*$. With no loss of generality, we can assume that these vertices are $\{1, 2, ..., m\}$. 5
\[ [1 - F(x_1 + d) + F(x_2 - d)]^{n-m} \] The marginal probability is obtained by integrating over 
x_1 and x_2. Since there are m(m-1) choices for specifying the minimum and maximum 
values, the conclusion follows.

Theorem 3. The probability that a specified vertex forms a \( K^*_{1,d} \) (i.e. is isolated) is

\[
P\{ \text{1 is isolated} \} = \int_{-\infty}^{\infty} \left\{ 1 - [F(x+d) - F(x-d)] \right\}^{n-1} f(x) \, dx. \tag{3}
\]

Proof. Given \( X_1 = x_1 \), the integrand is the probability that the remaining observations 
are further from \( x_1 \) than \( d \). Multiplying by \( f(x) \) and integrating gives the required 
marginal probability.

A vertex has degree \( \nu \), \( \nu = 0, 1, \ldots, n-1 \), if there are exactly \( \nu \) edges incident with that 
vertex. We now establish the probability that a specified vertex is of degree \( \nu \).

Theorem 4. The probability that a specified vertex has degree \( \nu \), \( \nu = 0, 1, \ldots, n-1 \) is

\[
P\{ \text{vertex 1 has degree } \nu \}
= \int_{-\infty}^{\infty} \binom{n-\nu}{\nu} [F(x+d) - F(x-d)]^\nu \left( 1 - F(x+d) + F(x-d) \right)^{n-\nu-1} f(x) \, dx. \tag{4}
\]

Proof. Vertex 1 has degree \( \nu \) whenever \( \nu \) elements from \( x_2, \ldots, x_n \) are within \( d \) of \( x_1 \) 
and the remaining \( n-\nu-1 \) are further from \( x_1 \) than \( d \). There are \( \binom{n-1}{\nu} \) choices for the 
vertices which are adjacent to vertex 1. Note that if \( \nu = 0 \), this coincides with 
Theorem 3.

Introducing indicator functions and using the inclusion-exclusion formula, we can 
exhibit the exact probability distributions for each of the random variables introduced.
Also, we can then obtain the factorial moments for each of these. Accordingly, we 
define the following random variables. Let
\[ V_{i_1, i_2, \ldots, i_m} = 1 \text{ if the vertices } i_1, i_2, \ldots, i_m \text{ form a } K_{m,d}, \]
\[ \text{where } 1 \leq i_1 < i_2 < \ldots < i_m \leq n, \]
\[ = 0, \text{ otherwise}, \]
\[ W_{i_1, i_2, \ldots, i_m} = 1 \text{ if the vertices } i_1, i_2, \ldots, i_m \text{ form a } K^*_{m,d}, \]
\[ \text{where } 1 \leq i_1 < i_2 < \ldots < i_m \leq n, \]
\[ = 0, \text{ otherwise}, \]
\[ Y_i(\nu) = 1 \text{ if vertex } i \text{ is of degree } \nu, \ 1 \leq i \leq n. \]
\[ = 0, \text{ otherwise.} \]

Let \( L_{m,d} \) be the number of distinct subsets of \( m \) vertices which form a \( K_{m,d} \) and let \( L^*_{m,d} \) be the number of distinct subsets of \( m \) vertices which form a \( K^*_{m,d} \), \( m = 1, 2, \ldots, n \), \( d > 0 \). Also, let \( Y_{n,d}(\nu) \) be the number of vertices of degree \( \nu \), \( \nu = 0, 1, 2, \ldots, n-1. \)

Then

\[ L_{m,d} = \sum V_{i_1, i_2, \ldots, i_m}, \quad (5) \]

and

\[ L^*_{m,d} = \sum W_{i_1, i_2, \ldots, i_m}, \quad (6) \]

the sums running over all distinct subsets \( J_m \) of \( \{1, 2, \ldots, n\} \) with \( |J_m| = m. \)

Similarly,

\[ Y_{n,d}(\nu) = \sum_{i=1}^{n} Y_i(\nu), \quad \nu = 0, 1, \ldots, n-1. \quad (7) \]

Specifically:

\[ L_{1,d} = |V| = n, \]
\[ L_{2,d} = \left| E \right|, \]
\[ L_{3,d} \text{ is the number of triangles in } G. \]
\[ L_{n,d} = 1 \text{ if } V_n \text{ is a complete graph and } 0 \text{ otherwise,} \]
\[ L_{n,d}^* \text{ is the number of isolated vertices of } G, \]
\[ L_{n,d}^* = L_{n,d}. \]

3. ASYMPTOTIC PROPERTIES

In this section, under the assumption \( d \to 0 \) and \( n \to \infty \), (in some cases, only the assumption \( d \to 0 \) is needed), we obtain asymptotic estimates for the probabilities of complete subgraphs of order \( m \), of maximal complete subgraphs of order \( m \) and the probability that a given vertex is of degree \( \nu \).

To accomplish this and to subsequently obtain asymptotic probability distributions for the number of complete subgraphs of order \( m \), the number of maximal complete subgraphs of order \( m \) and the number of vertices of degree \( \nu \), we introduce some preliminary lemmas.

Lemma 1. Let \( k \) and \( m \) be positive integers such that \( 0 < m \leq n \), \( n = 2, 3, \ldots \). Then

\[ \frac{n^{mk}}{(ml)^k} e^{-m^2 (k/(n-m))} \leq \binom{n}{m}^k \leq \frac{n^{mk}}{(ml)^k} \]

and for \( k \leq n/m \),

\[ \frac{n^{mk}}{(ml)^k} e^{-m^k (mk-1)/(n-mk+1)} \leq \frac{n!}{(n-mk)! (ml)^k} \leq \frac{n^{mk}}{(ml)^k}. \]

Thus,

\[ \binom{n}{m}^k \sim \frac{n!}{(n-mk)! (ml)^k}, \text{ for } k = o(\sqrt{n}), \text{ as } n \to \infty \]
Proof. Since \( n(n-1)\ldots(n-m+1) \leq n^m \), the right hand side of (8) is trivially established. To verify the left hand side of (8), write

\[
n(n-1)\ldots(n-m+1) = n^m \prod_{j=0}^{m-1} (1-j/n) \geq n^m 1-m/n)^m. \tag{11}
\]

Since, for \( x \geq -1, \exp(x/(1+x)) < 1+x \), letting \( x = -m/n \), the conclusion follows.

The right hand side of (9) follows upon noting that \( n(n-1)\ldots(n-mk+1) \leq n^{mk} \). To verify the left hand side of (9), write

\[
\frac{n!}{(n-mk)!} = n^{mk} \prod_{i=0}^{mk-1} (1-j/n)
\]

\[
= n^{mk} \sum_{j=0}^{mk-1} \log(1-j/n) = n^{mk} \sum_{j=0}^{mk-1} \sum_{i=1}^{j/n} i
\]

\[
\geq n^{mk} \sum_{j=0}^{mk-1} \sum_{i=1}^{j/n} i \geq n^{mk} \sum_{j=0}^{mk-1} j/(n-j)
\]

\[
\geq n^{mk} \sum_{j=0}^{mk-1} j/(n-mk) \geq n^{mk} e^{-mk/(n-mk+1)},
\]

establishing the lemma.

Lemma 2. Let \( F_X(x) \) be absolutely continuous with respect to Lebesgue measure with probability density function \( f_X(x) \). Let \( x \) be fixed and let \( d \) be a positive constant with \( F_X(x+d) < 1 \) and \( F_X(x-d) > 0 \). Then, if \( f'_X(x) \) exists and is uniformly bounded on \((x-d, x+d)\),
\[ F_X(x + d) - F_X(x - d) = 2df_X(x) + R(d), \]  

(12)

where

\[ |R(d)| \leq d^2 \sup_{x-d < y < x+d} f''(y). \]  

(13)

Similarly, if \( f''(x) \) exists and is uniformly bounded on \( (x-d, x+d) \), then

\[ F_X(x + d) - F_X(x - d) = 2df_X(x) + R_1(d), \]  

(14)

where

\[ |R_1(d)| \leq \frac{d^3}{3} \sup_{x-d < y < x+d} f''(y). \]  

(15)

In the identical manner,

\[ F_X(x + d) - F_X(x) = df_X(x) + R_3(d), \]  

(16)

where

\[ |R_3(d)| \leq \frac{d^2}{2} \sup_{x < y < x+d} f'(y). \]  

(17)

**Proof.** Expand \( F_X(x + d) \) and \( F_X(x - d) \) in a Taylor series about \( x \), obtaining

\[ F_X(x + d) - F_X(x) = df_X(x) + \left( \frac{d^2}{2!} \right) f'(x) + \left( \frac{d^3}{3!} \right) f''(x) + \ldots + \left( \frac{d^k}{k!} \right) f^{(k-1)}(x) + R_{k,d}(x) \]  

(18)
and

\[ F_X(x - d) = F_X(x) - d f_X(x) + \left( \frac{d^2}{2!} \right) f'_X(x) - \left( \frac{d^3}{3!} \right) f''_X(x) + \ldots \pm \left( \frac{d^k}{k!} \right) f^{(k-1)}(x) + R_{k,d}(x), \]  

(19)

where

\[ R_{k,d}(x) = \left( \frac{d^{k+1}}{(k+1)!} \right) f^{(k)}(\theta), \]  

(20)

for some \( \theta \) in \((x, x+d)\) and

\[ R'_{k,d}(x) = \left( \frac{d^{k+1}}{(k+1)!} \right) f^{(k)}(\theta_1) \]  

(21)

for some \( \theta_1 \) in \((x-d, x)\). The conclusion follows upon collecting terms and bounding the corresponding remainders by taking the supremum over \( \theta \).

Lemma 3. Let \( F_X(x) \) be absolutely continuous with respect to Lebesgue measure with probability density function \( f_X(x) \). Let \( x \) be fixed and let \( d \) be a positive constant with \( F_X(x+d) < 1 \) and \( F_X(x-d) > 0 \). Let \( y \) satisfy \( x < y < x + d \). Then, if \( f'_X(x) \) exists and is uniformly bounded on \((x, x+d)\),

\[ F_X(y) - F_X(x) = (y - x) f_X(x) + R_3(d), \]  

(22)

where \( R_3(d) \) is given in (17).

Similarly,

\[ F_X(y - d) - F_X(x + d) = (y - x - 2d) f_X(x) + R_1(d), \]  

(23)

where \( R_1(d) \) is given by (15).
Proof. The proof of lemma 3 is completely analogous to that of lemma 2 and is omitted.

Theorem 5. Let $F_X(x)$ be a cumulative distribution function, absolutely continuous with respect to Lebesgue measure and let $f_X(x)$ be the corresponding probability density function. Let $f_X(x)$ be uniformly continuous on every compact set $K$ and let $f''(x)$ exist and be uniformly bounded on $K$. Then the asymptotic probability $(d \to 0)$ that the vertices $\{1, 2, \ldots, m\}$ form a $K_{m,d}$ is

$$m \, d^{m-1} \int_{-\infty}^{\infty} f^m(x) \, dx.$$  (24)

The asymptotic probability $(d \to 0, n \to \infty, \text{so that } nd \to 0)$ that the vertices $\{1, 2, \ldots, m\}$ form a $K^*_m$ in $G_{n,d}$ is

$$d^{m-1} \int_{-\infty}^{\infty} f^m(x) (m + (n - m) f(x)) [(m-1)d - 2m] \, dx.$$  (25)

The asymptotic probability $(d \to 0, n \to \infty, \text{so that } nd \to 0)$ that vertex 1 is isolated is

$$1 - 2nd \int_{-\infty}^{\infty} f^2(x) \, dx.$$  (26)

The asymptotic probability $(d \to 0, n \to \infty, \text{so that } nd \to 0)$ that vertex 1 has degree $\nu$ is

$$\binom{n-1}{\nu} (2d)^\nu \int_{-\infty}^{\infty} (f(x))^{\nu+1} (1 - 2nd f(x)) \, dx.$$  (27)

Proof. Let $\epsilon > 0$ be chosen. Then there exists an $M = M(\epsilon)$ so that $P\{ |X| > M \} < \epsilon$. Therefore, we can write

$$| \int_{-\infty}^{\infty} [F(x+d) - F(x)]^{m-1} f(x) \, dx - d^{m-1} \int_{-\infty}^{\infty} f^m(x) \, dx |$$
\[ \leq \left| \int_{|x| \geq M} (F(x+d) - F(x))^{m-1} f(x) \, dx \right| - d^{m-1} \int_{|x| \leq M} f^m(x) \, dx \]
\[ + \left| \int_{|x| \leq M} (F(x+d) - F(x))^{m-1} f(x) \, dx - d^{m-1} \int_{|x| \leq M} f^m(x) \, dx \right| \]
\[ \leq 2^{m-1} \epsilon + d^{m-1} \epsilon^m \]

\[ + \sum_{i=0}^{k-1} \int_{x_i}^{x_i+1} \left[ d f(x_i) + R_3(d, x_i) \right]^{m-1} f(x) \, dx - d^{m-1} \int_{|x| \leq M} f^m(x) \, dx, \quad (28) \]

where \( R_3(d, x_i) \) is given in (17). Expanding the second term of (28) in a binomial series, we observe that the first term of the series coincides with the last term and hence (28) may be written as

\[ 2^{m-1} \epsilon + d^{m-1} \epsilon^m \]

\[ + \sum_{i=0}^{k} \sum_{j=1}^{m-1} \int_{|x| \leq M} \binom{m-1}{j} (d f(x))^{m-1-j} (R_3(d, x_i))^j f(x) \, dx. \quad (29) \]

The second term of (28) can be estimated using \( |R_3(d, x_i)| \leq (d^2/2) Q \), where \( Q \) is the uniform bound of \( f'(x) \). Since \( \epsilon \) is arbitrary, we have established that as \( d \to 0 \), the probability that the vertices \{1, 2, ..., m\} form a \( K_{m,d} \) is given by

\[ md^{m-1} \int_{-\infty}^{\infty} f^m(x) \, dx (1 + O(d)) \]

This estimate can be improved using the alternate form (18) and adding the hypothesis that \( f'(x) \) is uniformly bounded on every compact set.

The other conclusions of Theorem 5 are established in a similar manner and are not
specifically provided here.

We will denote the $r$th factorial moment of the random variable $X$ by $\mu_{(r)}(X)$ and the mean and variance of the random variable $X$ by $\mu(X)$ and $\sigma^2(X)$ respectively (when there is little chance of confusion, by $\mu_{(r)}$, $\mu$, and $\sigma^2$ respectively).

4. ASYMPTOTIC POISSON DISTRIBUTIONS

Theorem 6: If $n^m d^{m-1} \to r > 0$ as $n \to \infty$ and $d \to 0$, then $L_{m,d}$ is asymptotically Poisson distributed with mean $\lambda = \left(\frac{n}{m}\right) m d^{m-1} \int_0^\infty e^m (x) \, dx$.

Proof: The $k$th factorial moment of $L_{m,d}$ is given by $\mu_{[k]}(K_{m,d}) = k! S_k$.

Let $A_1$ be the event that the vertices $\{1, 2, \ldots, m\}$ form a $K_{m,d}$. By symmetry, for any subset of $m$ distinct vertices, $\{i_1, i_2, \ldots, i_m\}$, $1 \leq i_1 < i_2 < \ldots < i_m \leq n$, the probability that these vertices form a $K_{m,d} = P(A_1)$. If $\{i_{j_1}, i_{j_2}, \ldots, i_{j_m}\}$, $j = 1, 2, \ldots, k; k = 1, 2, \ldots$ are $k$ disjoint subsets of $m$ distinct vertices, then the probability that each of these forms a $K_{m,d}$ is $\left(\frac{P(A_1)}{k}\right)^k$, since the vertices are realizations of independent random variables.

Now

$$S_k = \sum P\{\bigcap \{A_{ij}\}\};$$

the sum runs over all distinct choices of $k$ subsets of $m$ vertices. We divide the terms in this sum into two sets, those with all $km$ vertices distinct and those with at least one vertex repeated, that is

$$S_k = S_{k_1} + S_{k_2},$$

where the events in $S_{k_1}$ have all vertices distinct and the events in $S_{k_2}$ have at least one of the $km$ vertices repeated in each term. $S_k$ has

$$\left(\binom{n}{m}\right)^k$$
terms, of which

\[ \frac{n!}{(n-mk)! (ml)^k} \]

are in \( S_{k1} \). Then, from Lemma 1, for every fixed \( m \) and \( k \) as \( n \to \infty \), we have

\[ \binom{n}{m}^k = \frac{n!}{(n-mk)! (ml)^k} \left( 1 + O\left( \frac{1}{n} \right) \right) = \frac{n^{mk}}{(ml)^k} \left( 1 + O\left( \frac{1}{n} \right) \right). \tag{31} \]

Then,

\[ S_k = \frac{n^{mk}}{(ml)^k} (P(A_1))^k + S_{k2} \]

It remains to show that \( S_{k2} \to 0 \) as \( n \to \infty \) and \( d \to 0 \), so that \( d^{m-1} n^m \to \tau \). The \( k \)-tuples of sets in \( S_{k2} \) have non-empty intersections with \( km-j \) distinct elements, \( j=1,2,...,km-k \).

From Lemma 1, it is easy to see that the number of selections of elements in \( S_{k2} \) with \( km-j \) distinct elements is asymptotically given by \( c_{jm} n^{km-j} \), for a specified constant \( c_{jm} > 0 \).

Let \( \{i_1, i_2, ..., i_m\} \) and \( \{k_1, k_2, ..., k_m\} \) be two subsets of \( m \) vertices with \( j \) elements in common. Then, the probability that both of these form a \( K_{m,d} \) is

\[ 2 \int_{-\infty}^{\infty} \int_{x}^{x+2d} [F(x+d) - F(x)]^{m-j-1} [F(y) - F(y-d)]^{m-j-1} [F(x+d) - F(y-d)]^j f(x) f(y) dx dy. \]

\[ = c' d^{2m-j-1}, \tag{32} \]

for some positive constant \( c' \).

This can be readily extended to \( k \) subsets of the vertices, where the subsets are partitioned into subcollections according to the number of vertices in common.

However, as in (32) above, the resulting integral has magnitude \( O(d^{km-j-1}) \) and hence each such collection of terms in \( S_{k2} \) has probability tending to zero, whenever \( k>1 \) and \( j<m \). Since there are a finite number of such terms, the conclusion follows.
Theorem 7. If $n \to \infty$ and $d \to 0$ so that $n^{\nu+1} d^{\nu} \to \tau_1 > 0$, then $Y_{n,d}(\nu)$ is asymptotically Poisson distributed with mean 
$n^{\nu+1} d^{\nu} \int_{\infty}^{\infty} [f(x)]^{\nu+1} dx$.

Proof: Let $A_i$ be the event that vertex $i$ has degree $\nu$, $i = 1, 2, \ldots, n$. We proceed by calculating the factorial moments, using the inclusion-exclusion formula. Let $A_{i_1, i_2, \ldots, i_k}(\nu_1, \nu_2, \ldots, \nu_k)$ be the event that vertex $i_1$ has degree $\nu_1$, vertex $i_2$ has degree $\nu_2$, ..., vertex $i_k$ has degree $\nu_k$. To evaluate $S_k$, we need to consider two separate cases.

We let $S_k = S_{k1} + S_{k2}$, where $S_{k1}$ contains the terms corresponding to the case where the $k$-tuples of vertices in those term in $S_k$ are not adjacent. The terms in $S_{k2}$ contain at least one pair of adjacent vertices.

If the vertices $i_1, i_2, \ldots, i_k$ contain no adjacent pairs, then

$$P(A_{i_1}, A_{i_2}, \ldots, A_{i_n}) = \prod P(A_i(\nu_i)).$$  \hspace{1cm} (33)

If $\nu_1 = \nu_2 = \ldots = \nu_k = \nu$, then this is a term in $S_{k1}$ and since the vertices are realizations of independent identically distributed random variables, (33) reduces to $(P(A_1))^k$.

There are

$$\frac{(n-k)!}{(\nu!)^k (n-k\nu)!} \binom{n}{k}$$

such terms. Thus, the magnitude of the terms in $S_{k1}$ is of the form $cn^{k\nu} d^{k\nu}$. It remains to show that the terms in $S_{k2}$ tend to zero under the hypotheses. Select $k$ vertices.

Assume that of the vertices to which they are adjacent, $r$ of the $k$ selected vertices have common adjacencies, but that none of the $k$ selected vertices are adjacent to each other. Let $i_j$ be the number of vertices adjacent to exactly $j$ of the $k$ selected vertices. Then each of these contribute $j$ to the total degree of the $k$ vertices, but should only be counted once in the selection of the vertices to which the $k$ specified vertices are
adjacent, and result in an overcount by $j-1$ in each instance. Hence the contribution of such terms has magnitude $c_j n^{k+1-\sum_{i(j-1)} k\nu-\sum_{j-1}^i}$, which tends to zero as $n\to\infty$ and $d\to0$ with $n^{\nu+1} d^{-\tau} \to \tau > 0$. If in addition, of the $k$ selected vertices, $r_{ij}, 1 \leq i < j \leq k$, is the number of adjacencies of the $i$th vertex to the $j$th vertex, then, it follow that the exponent of $n$ is reduced by twice the number of such adjacencies and the exponent of $d$ is reduced by the number of such adjacencies; hence such terms also tend to zero.

Theorem 8. If $n \to \infty$ and $d \to 0$ so that $n^m d^{m-1} \to \tau > 0$, then $L^* (m_d)$ has an asymptotic Poisson distribution with expected value $\left( \frac{n}{m} \right)^m d^{m-1} \int_{-\infty}^{\infty} (f(x))^m dx$, $m=2,3,\ldots$. If $n \to \infty$ and $d \to 0$ so that $dn/(\ln n) \to \tau > 0$, then the number of isolated vertices $K^* (m_d)$ has an asymptotic Poisson distribution with expected value $n \int_{-\infty}^{\infty} e^{-2ndf(x)} f(x) dx$.

Proof. The proofs are completely analogous to those of the two preceding theorems and the details are omitted.

5. CONCLUDING REMARKS

The purpose of this investigation was to obtain some preliminary results which would be useful in developing statistical tests to determine if clusters are real or simply chance occurrences. The distributions most likely to be assumed by experimenters are the uniform, exponential and the normal distributions, all of which satisfy all the necessary assumptions. The specific choice of a criterion to use in testing would depend on the alternative. Some initial ideas on this have been advanced, but substantially more work is needed. This work is continuing and additional results should be reported soon.

REFERENCES

P. Arabie and L. J. Hubert (1996), An overview of combinatorial data analysis, in Clustering and Classification, P. Arabie, L. J. Hubert and G. De Soete, Editors, World
Scientific, Singapore.


19


