# COMPUTATIONAL ANALYSIS OF THE RANDOM COMPONENTS INDUCED BY A BINARY EQUIVALENCE RELATION

Thesis by

Guy de Balbine

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1968

(Submitted April 17, 1968)

# ACKNOWLEDGEMENT

I wish to express my deepest gratitude to Dr. Joel N. Franklin for his enlightened guidance and constant encouragement throughout the course of this research. This work would not have been without his constructive criticism and challenging questions.

# ABSTRACT

The problem of partitioning into classes by means of a binary equivalence relation is investigated. Several algorithms for determining the number of components in the graph associated with a particular set of elements are constructed and compared. When the classification process operates on independently drawn samples of $n$ distinct elements from a population, the expected number of components is shown to be obtainable recursively for a class of problems called separable; in all cases, estimates are available to reach any desired level of accuracy. Clustering models in Euclidean space are analyzed in detail and asymptotic formulas obtained to complement experiments. Conjectures concerning the general behavior of the expected number of components are presented also. Finally, several computational tools of general interest are improved significantly.

## TABLE OF CONTENTS

# Chapter I

## Introduction

The purpose of this study is to investigate problems of partitioning into classes. Given any binary relation defined for all pairs of elements in the set to be partitioned, a corresponding binary equivalence relation can be defined. Using an undirected graph, the edges of which indicate that the relation holds for adjacent vertices, we obtain a certain number of components, each component corresponding to one of the equivalence classes of the vertex set. The expected number of components is of interest in problems in physical chemistry and in other fields.

Given a graph on $n$ vertices, the number of components can be expressed in terms of the inverse of the row sums of the $n-1$st Boolean power of the graph incidence matrix. This mathematical formulation provides one method for computing the number of components but others, more efficient, also can be constructed. An important problem that we examine is the determination of the expected value of the number of components when the classification process operates on independently drawn samples of $n$ distinct elements from a population. Of course, for some rather simple examples, such as discrete and continuous occupancy problem on the line, we find analytical answers, but in the majority of cases we cannot hope for closed-form solutions and must resort to computational experimentation. For a certain class of problems called separable, such as some random graphs and rooks problems, powerful recursive calculations can be used to obtain expected values. However, if the structure of the relation is far too complex or only statistically known, our more modest aim is to derive estimates for the probable number of components. Depending upon the toil that we are ready to face, various approximate methods ranging from simple a priori estimation to lengthy Monte Carlo sampling are available in order to reach any desired level of accuracy.

Among all binary equivalence relations there are some which deserve special attention since they are closely related to practical problems. In that respect we shall analyze in detail discrete and continuous adjacency problems in 1, 2 or 3 dimensions which form simplified clustering models in Euclidean space.

Throughout this work great care is exercised in using analytical answers, asymptotic expansions and computational estimates in a harmonious conjunction. A result of this comprehensive study is the inference of conjectures concerning the general behavior of the expected number of components.

Finally, the search for efficient specialized algorithms and conclusive experiments has resulted in a variety of significantly improved computational tools which potentially offer a wide applicability.

## 1. 1. Binary Classification. Binary Equivalence Relation

We consider a population $\mathcal{P}$ of N individuals $\left\{a_1, a_2, \ldots, a_N\right\}$; for any pair of elements $a_i, a_j \in \mathcal{P}$, let a connected binary relation $\mathcal{R}$ be defined. Connectedness simply implies that between any two distinct $\mathcal{R}$-members, either $\mathcal{R}$ or $\mathcal{R}^{-1}$ holds. We perform n experiments $\mathcal{E}_1, \ldots, \mathcal{E}_n$, each experiment being defined as the selection of a member of $\mathcal{P}$ according to some prescribed probability law. The space $\mathcal{S}_i$ of experiments $\mathcal{E}_i$ is simply

$$\mathcal{S}_i = \left\{a_1,\ a_2,\ \ldots,\ a_N\right\} = \mathcal{P}$$

and the probability of an elementary event $a_j$ is

$$p_i(a_j),\ i \in \left\{1,\ 2,\ \ldots,\ n\right\}\ ,\ j \in \left\{1,\ 2,\ \ldots,\ N\right\}$$

The outcomes of our combined experiment $\mathcal{E} = \mathcal{E}_1 \times \mathcal{E}_2 \times \ldots \times \mathcal{E}_n$ are ordered n-typles $\mathcal{A} = (a_{i_1}, a_{i_2}, \ldots, a_{i_n})$ forming the space

$$\mathcal{S} = \mathcal{P} \times \ldots \times \mathcal{P} = \mathcal{P}^n$$

The probability of an event $\mathcal{A} \in \mathcal{S}$ is $p(a_{i_1}, a_{i_2}, \ldots, a_{i_n})$ which, if the experiments are independent, becomes

$$p(a_{i_1}, a_{i_2}, \ldots, a_{i_n}) = \prod_{j=1}^{n} p_j(a_{i_j})$$

Each event $\mathcal{A}$ involves n elements of $\mathcal{P}$, distinct or not, which can be grouped into classes by means of the relation $\mathcal{R}$. This classification process is performed under the following conditions:

(1) for any pair $a_i, a_j \in \mathcal{P}$, $a_i \not\equiv a_j$, in a class C there exists some sequence

$$a_i \mathcal{R}^{\pm 1} a_{k_1}, a_{k_1} \mathcal{R}^{\pm 1} a_{k_2}, \ldots, a_{k_\nu} \mathcal{R}^{\pm 1} a_j$$

where $\{a_{k_1}, a_{k_2}, \ldots, a_{k_\nu}\}$ is a subset, possibly empty, of the elements in C.

(2) if an individual in a class is in the relation $\mathcal{R}$ or $\equiv$ with another individual, then the second individual is in the same class as the first one.

The purpose of condition (1) is to accommodate irreflexive and non-reflexive binary relations. By irreflexive we mean that no member of $\mathcal{P}$ bears the relation $\mathcal{R}$ to itself, whereas non-reflexive means that each $\mathcal{R}$-member does not bear the relation $\mathcal{R}$ to itself. Also it establishes the chaining property among elements of the same class. $a_i \mathcal{R}^{\pm 1} a_j$ means either a $\mathcal{R}$ b or b $\mathcal{R}$ a or both. The identity binary relation $\equiv$ is an equivalence relation used to group into the same class identical elements. It is convenient to let $\mathcal{C}(a_i)$ represent that subset of $\mathcal{A}$ which forms the class to which $a_i$ belongs.

Theorem - For any binary relation $\mathcal{R}$, under hypothesis (1) and (2), the classification is unique and the resulting classes are mutually exclusive.

proof: a simple constructive proof proceeds as follows; Construct an undirected graph by adding a vertex for $a_{i_\nu}$ if $a_{i_\nu} \notin \{a_{i_1}, a_{i_2}, \ldots, a_{i_{\nu-1}}\}$. If $a_{i_\nu}$ indeed forms a new vertex, connect it by single edges to all other vertices for which $\mathcal{R}, \mathcal{R}^{-1}$ or both hold. This construction clearly satisfies conditions (1) and (2); each class is obtained by arbitrarily selecting one vertex and picking out all dependent branches until the nodes are exhausted.

Notice that, although we do not require $\mathcal{R}$ to be an equivalence relation, conditions (1) and (2) ensure that the classification is done in terms of an equivalence relation $\mathcal{R}'$ derived from $\mathcal{R}$ and such that

$$a_i \mathcal{R}' a_j \quad \Longleftarrow \quad a_i \mathcal{R} a_j \quad \text{or} \quad a_j \mathcal{R} a_i \quad \text{or both}$$

$$a \mathcal{R}' a \qquad \forall a \in \mathcal{P}$$

$$a_i \mathcal{R}' a_j \quad \text{and} \quad a_j \mathcal{R}' a_k \quad \Longrightarrow \quad a_i \mathcal{R}' a_k$$

Of course, $\mathcal{R}'$ depends on the particular sample chosen since it is defined to be transitive. In practice, we perform a series of elementary pairwise tests with a symmetric reflexive relation, transitivity only being obtained indirectly.

# CHAPTER II

## Class Counting Algorithms

In a very limited number of cases, we can find analytical solutions and possibly asymptotic estimates for the number of classes given a population $\mathcal{P}$, an equivalence relation $\mathcal{R}'$ and a sampling distribution over $\mathcal{P}$. Most of the time, however, the alternative has to be chosen which requires computing sample statistics.

Several algorithms to determine the number of classes, given the elements of some sample $\mathcal{A}$, can be proposed. They all have, as a common characteristic, their generality of use since they are applicable to any equivalence relation $\mathcal{R}'$ and finite sample $\mathcal{A}$ from a finite or infinite population $\mathcal{P}$. As a matter of fact, their only interface with the outside world is at the following level: are the $i^{th}$ and $j^{th}$ element of the current sample in the relation $\mathcal{R}'$ with each other?

In order to appreciate the merits of these algorithms, we can use criteria based upon their speed (count of operations), storage requirements and accuracy. Accuracy here refers to cases in which the number of classes is only estimated without actually performing the complete counting. The result is in general no longer an integer but a certain range of integer values which includes the true answer and the accuracy is then measured by the extent of the range. In the present chapter we are strictly concerned with exact counting algorithms but in chapter ( IV ) approximate methods will be presented too.

An essential choice to make is to select a data structure representation for the connectivity information among elements of $\mathcal{A}$. Of course, one may construct the 0-1 incidence matrix but this requires testing a priori all $\binom{n}{2}$ pairs of distinct elements $(a_i, a_j)$ for the relation $\mathcal{R}'$ to hold. Also, if the probability that $a_i \mathcal{R}' a_j$ holds is small, any storage representation which accommodates all of the $\binom{n}{2}$ entries will appear extremely wasteful. On the other hand, if we choose to only keep track of the less likely outcomes, here those pairs for which $a_i \mathcal{R}' a_j$, the retrieval effort is bound to be significantly larger than before. Yet, we can develop an algorithm which does not necessarily require the full incidence matrix and which tests $a_i \mathcal{R}' a_j$ at most once. This list structure algorithm is presented last and is probably the most useful. Nevertheless, we feel that the two following algorithms have interesting characteristics in their own right which amply justify their analysis.

## 2.1. Algorithm 1: Iterative Computation on the Incidence Matrix

The incidence matrix A of the undirected graph $G_n$ with n vertices, single edges and no loops is

$$\begin{aligned} a_{ij} &= 1 - \delta_{ij} \qquad &\text{if} \qquad & v_i \mathcal{R}' v_j \\ &= 0 \qquad &\text{if} \qquad & v_i \not\mathcal{R}' v_j \end{aligned} \tag{2.1}$$

The binary relation $\mathcal{R}'$ is an equivalence relation derived from $\mathcal{R}$ (see section [ 1.1 ]) and A is thus an nxn symmetric matrix with 0 diagonal elements.

Let $a_{ij}^k$ be the (i, j) element of the $k^{th}$ power of A ; physically, $a_{ij}^k$ represents the number of paths of length k from

$v_i$ to $v_j$. Since the shortest path between any two vertices in $G_n$ cannot be of length greater than $n-1$, $v_i$ and $v_j$ belong to the same class if and only if

$$\sum_{k=0}^{n-1} \beta_k \, a_{ij}^k > 0 \qquad (2.2)$$

for any set of coefficients $\beta_k$ such that

$$\beta_k \geqslant 1 \qquad k = 0, 1, \ldots, n-1$$

In particular, consider the matrix

$$D = A + I$$

where $I$ is the identity matrix of order $n$. Its $(n-1)^{st}$ power has the expression

$$D^{n-1} = (A+I)^{n-1} = \sum_{k=0}^{n-1} \binom{n-1}{k} A^k \qquad (2.3)$$

which is of the type ( 2.2 ). Consequently, $\mathcal{C}(v_i) \equiv \mathcal{C}(v_j)$, $\mathcal{C}(v)$ designating the class or set of vertices to which the vertex $v$ belongs, if and only if $d_{ij}^{n-1} > 0$.

Let $\vec{U}$ be the column vector with all $n$ components equal to 1. For any vector $\vec{V}$ with non-zero components $v_i$ define as $\vec{V}^{-1}$ the inverse vector having components $\frac{1}{v_i}$.

<u>Theorem</u> - The number of classes of a graph $G_n$ with incidence matrix A is

$$c = \vec{U}' \cdot \left[ \min \left( \vec{U}\,\vec{U}' \,,\ (A+I)^{n-1} \right) \cdot \vec{U} \right]^{-1} \qquad (2.4)$$

proof: as previously shown, the (i, j) element of $(A+I)^{n-1}$ is either $\geq 1$ or $0$, depending upon whether or not there exists a path from $v_i$ to $v_j$. Taking the minimum of $(d_{ij}^{n-1}, 1)$ yields a (0-1) matrix, indeed a truth matrix for the relation $\mathcal{R}'$ on $G_n$. Applying this transformation to $\vec{U}$, produces a row sum vector, the $i^{th}$ component of which is the size of the class of which $v_i$ is a member. In the last operation to be performed, each row is counted with a weight inversely proportional to the size of the class to which it belongs, thereby yielding $c$ by simply taking the scalar product of a unit row vector with the inverse of the row sum vector as defined above.

Remark on the computation of $(A+I)^{n-1}$: For the purpose of obtaining the number of classes, the intermediate matrices $A, A^2, A^3, \ldots, A^{n-2}$ need not be explicitly computed and a shortcut may be used. As a matter of fact, the next paragraph is not only relevant to matrix powers but also to other transformations, provided they are associative.

Let the binary representation of $n-1$ be

$$n-1 = \sum_{\nu=0}^{m} b_\nu \, 2^\nu$$

$$b_m = 1 \quad , \quad b_\nu \in \{0, 1\} \qquad \nu = 0, 1, \ldots, m-1$$

then

$$D^{n-1} = \prod_{\nu=0}^{m} D^{b_\nu 2^\nu} \tag{2.5}$$

operation which involves $\tau = m + m' - 1$ matrix multiplications, if there are exactly $m'$ non-zero $b_\nu$'s. This number $\tau$ satisfies

$$\lfloor \log_2(n-1) \rfloor - 1 \leq \tau \leqslant 2 \lfloor \log_2(n-1) \rfloor - 1 \qquad (2.6)$$

which is significantly better than the $n-2$ multiplications by $D$ in the straightforward approach. Of course, unless we are in the special case $n-1=2^m$, the present scheme necessitates two matrices to be kept all along, namely, $D^{2^\nu}$ and $\prod_{i=0}^{\nu} D^{b_i 2^i}$. But it is certainly not a very serious objection, especially for symmetric matrices, in which case the above two matrices can conveniently be housed in the upper and lower halves of an $(n+1 \times n)$ matrix.

Let us now return to the specific case of class computation. As we have already pointed out, we only need to know whether $d_{ij}^{n-1}$ is $0$ or $\geqslant 1$ or equivalently whether $d_{ij}^{N}$ is, for $N > n-1$. Choosing $N = 2^{\lceil \log_2(n-1) \rceil}$ simplifies greatly our task since we can now dispense with one matrix and only calculate $D, D^2, D^4, \cdots, D^N$. Furthermore, let us assume that at the $\nu^{th}$ stage, i.e., when computing $D^{2^\nu}$ from $D^{2^{\nu-1}}$, we still have $d_{i,j}^{2^{\nu-1}} = 0$. We must then form the scalar product of the $i^{th}$ row of $D^{2^{\nu-1}}$ with its $j^{th}$ column, but as soon as this product becomes greater than $1$ we need not go any further; we just set $d_{ij}^{2^\nu} = 1$. Also, we have tacitly assumed from the beginning that every matrix product is performed in place, the updated values $d_{ij} = 1$ being used as soon as they are found. Consequently, we never even really compute $D^2, D^4, \ldots, D^N$ but at the completion of the $\nu^{th}$ stage we obtain a matrix $\tilde{D}^{2^\nu} \geqq D^{2^\nu}$. So much the better! Finally, the algorithm terminates when either $N$ is reached or a complete sweep of the matrix produces no update since $D^{2^{\nu-1}} = D^{2^\nu} \Longrightarrow D^{2^\nu} = D^N$.

As we now show, the steps towards efficiency described in the previous paragraph reduce significantly $\mu = \mathcal{O}(n^4)$ , number of operations necessary for the straightforward computation of $D^{n-1}$.

a) computing $D^{2^{\lceil \log_2(n-1)\rceil}}$ instead of $D^{n-1}$ yields $\mu = \mathcal{O}(n^3 \log_2 n)$ .

b) let $\alpha_\ell$ be the number of pairs of distinct vertices $\{v_i, v_j\}$ which are at the distance $\ell$ . In the product $D^{2^\nu} \times D^{2^\nu}$ , only $\binom{n}{2} - \sum_{\ell=1}^{2^\nu} \alpha_\ell$ scalar products need be evaluated; to reach $D^N$ , the number of scalar products is expressed by

$$\tau = \lceil \log_2(n-1)\rceil \binom{n}{2} - \sum_{\nu=0}^{\lceil \log_2(n-1)\rceil - 1} \sum_{\ell=1}^{2^\nu} \alpha_\ell$$

$$= \sum_{\nu=0}^{\lceil \log_2(n-1)\rceil - 1} \sum_{\ell=2^\nu+1}^{n-1} \alpha_\ell = \sum_{\ell=2}^{n-1} \lfloor \log_2 \ell \rfloor \alpha_\ell \tag{2.7}$$

so that the operational count becomes

$$\mu = \sum_{\ell=2}^{n-1} \lfloor \log_2 \ell \rfloor \alpha_\ell \, \mathcal{O}(n^2) \tag{2.8}$$

Let $c_\nu$ be the number of classes of size $\nu$ , i=1,2,...,n ; then

$$\sum_{\nu=1}^{n} \binom{\nu}{2} c_\nu = \sum_{\ell=1}^{n-1} \alpha_\ell \tag{2.9}$$

which is maximum for a unique class of size n ; in fact, the worst circumstance must correspond to a graph with the smallest number of edges, that is a tree on n vertices, since adding any edge can only decrease or leave unchanged the distance between any two vertices. Furthermore, proceeding by induction on the number of edges, that tree must be a linear chain; the adjunction of an edge to a linear chain maximizes ( 2.8 ) or ( 2.9 ) if it is made at the end of the chain producing the sequence of distances $\alpha_\ell = n-\ell$, $\ell = 1, 2, \ldots, n-1$; as

$$\sum_{\ell=1}^{n-1} \alpha_\ell = \binom{n}{2}$$

any other sequence can be constructed by repeatedly borrowing 1 from some $\alpha_{\ell_1}$ and adding 1 to $\alpha_{\ell_2}$ such that $\ell_1 > \ell_2$ but $\lfloor \log_2 \ell_1 \rfloor \geqslant \lfloor \log_2 \ell_2 \rfloor$ so that the maximum number of scalar products is expressed by

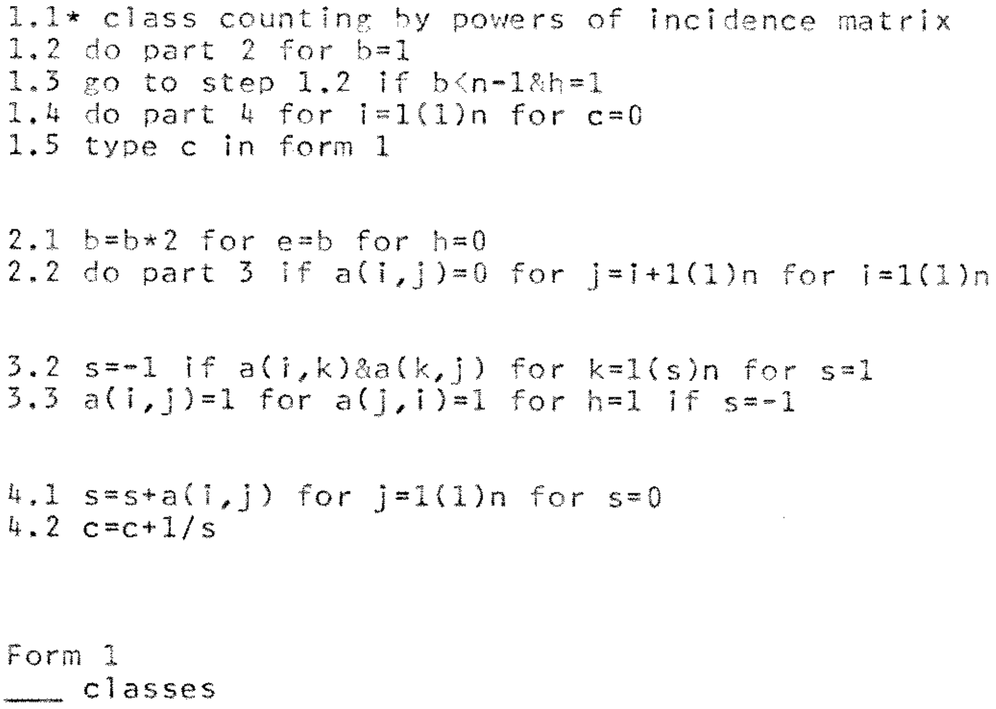
$$\tau = \sum_{\ell=2}^{n-1} \lfloor \log_2 \ell \rfloor (n-\ell) < \int_1^n (n-x) \log_2 x \, dx \qquad (2.10)$$

Using

$$\int_1^n (n-x)\log_2 x \, dx = \frac{1}{\log 2}\left[\frac{n^2}{2}\log n - \frac{(n-1)(3n-1)}{4} - (n-\frac{1}{2}) \log 2\right]$$

the number of scalar products will then be at worst $\frac{n^2}{2} \log_2 n + \mathcal{O}(n^2)$. Each such operation involves at most n tests which are performed as logical intersection (AND) operations rather than multiplications.

We have thus shown that in the worst possible case of a linear chain, the number of tests to perform is bounded above by $\frac{n^3}{2} \log_2 n$.



```
1.1* class counting by powers of incidence matrix
1.2 do part 2 for b=1
1.3 go to step 1.2 if b<n-1&h=1
1.4 do part 4 for i=1(1)n for c=0
1.5 type c in form 1


2.1 b=b*2 for e=b for h=0
2.2 do part 3 if a(i,j)=0 for j=i+1(1)n for i=1(1)n


3.2 s=-1 if a(i,k)&a(k,j) for k=1(s)n for s=1
3.3 a(i,j)=1 for a(j,i)=1 for h=1 if s=-1


4.1 s=s+a(i,j) for j=1(1)n for s=0
4.2 c=c+1/s


Form 1
____ classes
```

Figure 2.1

However, in practice the average number of tests is much smaller equal to $\beta \sum_{\ell=2}^{n-1} \lfloor \log_2 \ell \rfloor \alpha_\ell n$, $\frac{1}{n} \leq \beta \leq 1$, β representing the average number of tests performed during a scalar product, before a match is encountered. In practice β is small and the algorithm is quite simple so that it proves more efficient than using (Wrathall)

$$\frac{1}{I - \varepsilon A} = I + \varepsilon A + \varepsilon^2 A^2 + \ldots$$

that is, performing an actual matrix inversion to obtain the non-zero entries in the final incidence matrix.

Figure (2. 1) shows an implementation of algorithm 1 in CITRAN (CIT translator) which exhibits its simplicity.

## 2. 2. Algorithm 2: Recursive Computation on the Incidence Matrix

We take the opportunity here to comment briefly on the interrelationship between programming languages and the algorithms they host. Often, we find that the convenient way of attacking a given problem from a numerical standpoint does not correspond to the intuitive approach we would use if we were asked to obtain the solution in a few simple cases. In some instances, this discrepancy can be explained by our inability to think along the lines of an efficient but too complex algorithm. Most of the time, however, we tend to formulate the problem in terms of those basic processes which are most natural to the programming language used. For instance, a language like FORTRAN is especially suited for iterative calculations but has no built-in recursion capability, whereas, the contrary is true for McCarthy's LISP 1. 0. Consequently, changing the host language and performing a literal translation of an algorithm may not necessarily yield the same algorithm as we would obtain had we directly used the latter language. If neither language is a subset of the other one, this will in general be the case. By way of illustration, we now examine a "natural" algorithm.

Given a graph $G_n$ with $n$ vertices, the separation of its vertex set into disjoint connected classes can be achieved by selecting one of the remaining vertices and "pulling out" all of its dependent ones, that is, removing the connected subgraph of the class to which it belongs, and so on until no vertex remains. If we think in terms of a rigid physical structure, the selection process is a single operation, all nodes of the subgraph being moved at once. If the structure is now articulated with loose joints, the removal may be considered to take place in stages. First, all vertices at a distance 1 from the chosen vertex are marked; then all vertices not previously marked and having an edge in common with one of the last vertices marked are themselves selected, and so forth. We have thus introduced a hierarchical structure with respect to an arbitrary node acting as the root of each subgraph; each vertex at a distance d from the root is responsible for collecting the vertices to whom it is directly connected and which have not yet been marked. This intuitive procedure can be easily implemented as a recursive program operating on the upper half of the incidence matrix. For any $a_{ij} = 1$, row i and column j are scanned one position at a time, always moving away from $a_{ij}$, under the following rules (assuming we are presently moving along row i)

- if $a_{ik} = 0$, continue;
- if $a_{ik} = 1$ and is already marked as a member of the same class as $a_{ij}$, do not examine any further elements along this direction;

```
1.01* recursive class counting algorithm
1.1 V(k)=(k=1) for H(k)=(k=1) for k=n(-1)1
1.2 do part 2 if a(i,j)=1 for j=i(1)n for i=1(1)n
1.3 type k-1 in form 2


2.1 do part 3 for I(1)=i for J(1)=j for k=k+1 for l=1


3.03 a(I(l),J(l))=k
3.05 go to step 3.26 if u¬=l&u¬=0 for u=V(J(l))
3.1 do part 4 for E(l)=I(l)-1(s)1 for s=-1
3.2 do part 4 for E(l)=I(l)+1(s)J(l) for s=1
3.26 go to step 3.6 if u¬=l&u¬=0 for u=H(I(l))
3.3 do part 5 for E(l)=J(l)-1(s)I(l) for s=-1
3.4 do part 5 for E(l)=J(l)+1(s)n for s=1
3.6 l=l-1


4.02 done if a(E(l),J(l))=0
4.05 go to step 4.5 for s=-s if a(E(l),J(l))=k
4.1 done if 0<H(E(l))<l
4.2 H(E(l))=l+1 if H(E(l))=0
4.3 do part 3 for l=l+1 for I(l+1)=E(l) for J(l+1)=J(l)
4.5 done


5.02 done if a(I(l),E(l))=0
5.05 go to step 5.5 for s=-s if a(I(l),E(l))=k
5.1 done if 0<V(E(l))<l
5.2 V(E(l))=l+1 if V(E(l))=0
5.3 do part 3 for l=l+1 for I(l+1)=I(l) for J(l+1)=E(l)
5.5 done


Form 2
___ classes
```

Figure 2.2

• if $a_{ik} = 1$ but column $k$ has already been assigned for scanning by some element $a_{\ell k}$, $\ell \neq i$, continue; otherwise, assign column $k$ for scanning by $a_{ik}$ and initiate the scan in that column.

This process starts with $a(1,1)$, proceeds to find all of the vertices in its class, then ends up at $a(1,1)$; the diagonal elements are examined one at a time, until one is found which has not been assigned to a class yet. The algorithm starts again with this new element until all diagonal positions are assigned to specific classes.

To show the conciseness of the algorithm when written in an appropriate language, figure ( 2.2 ) is a listing of the corresponding CITRAN (CIT translator) program, parts 3, 4 and 5 being reentrant.

The number of tests to be performed is approximately constant $\mathcal{O}(n^2)$ but we must take into account the greater complexity of the algorithm.

## 2.3. Algorithm 3: Class Counting by Means of List Structures

Definition - A contraction $H$ of a graph $G_n$ is a graph whose vertices are connected subgraphs of $G_n$ forming a partitioning of the vertex set of $G_n$.

Two vertices of $H$ are adjacent if the corresponding subgraphs have adjacent vertices in $G_n$.

Lemma - The number of classes is invariant under a contraction operation.

proof: let $a_i$ and $a_j$ be two distinct vertices of $G_n$. Each one corresponds to unique subgraphs $H_i$ and $H_j$ in the partitioning of $G_n$, therefore, to unique vertices $h_i$ and $h_j$ of $H$. If $a_i \mathcal{R}' a_j$,

$a_i \epsilon H_i$ , $a_j \epsilon H_j \Rightarrow h_i \mathcal{R}' h_j$ . If $h_i \mathcal{R}' h_j$ , $H_i$ and $H_j$ are connected subgraphs with at least one edge in common, thus $H_i \cup H_j$ is connected and $a_i \mathcal{R}' a_j$ .

Definition - A contraction H of $G_n$ is minimal if its vertex set is independent. The number of classes in $G_n$ is equal to the number of vertices in its minimal contraction.

An elementary contraction is the replacement of a pair of adjacent vertices by a single vertex.

Any contraction of $G_n$ can be realized as a sequence of elementary contractions.

The minimal contraction of G is obtained after exactly n-c contractions.

We now describe how a very simplified ring structure can be advantageously used to obtain the minimal contraction of $G_n$ . The elements $a_{i_\nu} \epsilon \mathcal{A}$, $\nu = \{1, 2, \ldots, n\}$, will always be referenced by means of their index $\nu$ . A ring is a linked subset of $\mathcal{A}$ such that each element contains the index of the next ring element. The first element of a ring has a lower index than its predecessor.

At any stage of the process, a ring link and a class identifier are associated with each element. Members of the same class are chained to form a ring; at the beginning, all elements belong to single element rings (i.e., their ring link is equal to their index). The class identifier of an element is the index of the beginning of the ring to which it belongs.

Each element is tested in turn with every element of higher index for the relation $\mathcal{R}'$ to hold, provided that they do not already belong to the same class as a result of the transitivity property of $\mathcal{R}'$. Whenever a matching element is found, the new class obtained is assigned a class identifier equal to the minimum of the pair of identifiers. The next step of the elementary contraction is to merge both rings while assigning the new class identifier to all elements encountered.

After repeating this procedure n times, we obtain c classes, c being equal to n minus the number of mergers performed. The actual classes can be trivially obtained by walking around the rings, starting with their element of lowest index.

Notice that the links were carefully chosen in order to make the merging process simple and efficient; the new class identifier is the index at which merging is to start; it then proceeds by updating ring links and class identifiers of both rings until their first element is reached.

Algorithm 3 differs from the previous ones in the fact that it makes use of the transitivity property of $\mathcal{R}'$ to perform as few tests as possible. This may be of special importance if the relation $\mathcal{R}'$ is costly to evaluate. At the same time, since the whole incidence matrix is never needed, the amount of storage required is kept to a small value, namely, 3 n words.

The number of operations involved can be analyzed as follows: in the worst possible case, every contraction associates a single vertex of $G_n$ with a subgraph $H_\nu$ of $G_n$. The corresponding ring merging requires therefore $2(\nu+1)$ operations. For c classes of sizes $n_1, n_2, \ldots, n_c$ the ring operation count is bounded by

```
1.1* Class counting by list processing
1.3 C(j)=j for L(j)=j for j=1(1)n
1.4 do part 2 for l=k+1(1)n for k=1(1)n-1 for c=n
1.5 type c in form 4


2.1* Does relation R hold?
2.12 done if C(k)=C(l)
2.13* part 20 sets f=1 if R holds, 0 otherwise;
2.14 do part 20
2.16 done if f=0
2.2  p=(C(k)=m) for  m=min(C(k),C(l)) for c=c-1
2.22 R(p)=C(l) for R(1-p)=C(k)
2.23 R(0)=L(R(0)) for d=R(0) for p=0
2.24 p=1-p if R(p)>R(1-p)
2.26 C(R(1))=m if p=1
2.28 t=L(d) for d=R(p) for L(d)=R(p)
2.3  go to step 2.24 for R(p)=t if t>d
2.32 L(d)=R(p) for p=1-p
2.34 done if p=0
2.36 t=L(R(p)) for C(R(p))=m
2.38 go to step 2.36  for R(p)=t if t>R(p)
2.4  L(R(p))=m

Form 4
   __   classes found
```

```
Relation R: 2 integers are multiples  (10 numbers)

Sample :
 84   46   37   27   20   88    3   10   72   84

   5  classes found
Class   1:      84     27      3     72     84
Class   2:      46
Class   3:      37
Class   4:      20     10
Class   5:      88
```

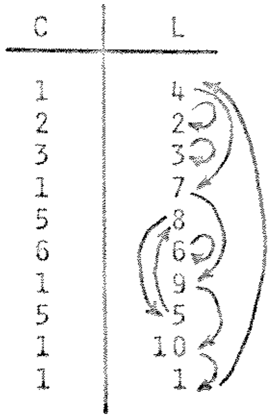


Figure 2.3

$$\sum_{i=1}^{c} \sum_{\nu=1}^{n_i-1} 2(\nu+1) = n - 2c + \sum_{i=1}^{c} n_i^2 \tag{2.11}$$

Besides, a number of tests $a_i \mathcal{R}' a_j$ are performed; if we distinguish between positive and negative tests, the test being positive if $a_i \mathcal{R}' a_j$, negative if $a_i \not\mathcal{R}' a_j$, all of the negative tests have to be carried out whereas the number of positive tests is dependent upon the ordering of the elements in the sample $\mathcal{A}$ (we assume here that $a_i \not\mathcal{R}' a_j$ cannot be deduced from the outcome of previous tests involving $a_i$ or $a_j$ separately). The number of negative tests performed is always

$$\sum_{i=1}^{c-1} n_i \sum_{j=i+1}^{c} n_j = \frac{1}{2}\left(n^2 - \sum_{i=1}^{c} n_i^2\right) \tag{2.12}$$

whereas the minimum number of positive tests corresponds to the total number of branches necessary to span each class with a tree, that is $(n-c)$ tests. The minimum number of tests is therefore

$$\frac{1}{2}\left(n^2 - \sum_{i=1}^{c} n_i^2\right) + n - c$$

or as a fraction of the total number of tests

$$\alpha = \frac{\frac{1}{2}\left(n^2 - \sum_{i=1}^{c} n_i^2\right) + n - c}{\frac{1}{2} n(n-1)} = 1 - \frac{\sum_{i=1}^{c} n_i^2 + 2c - 3n}{n(n-1)} \tag{2.13}$$

The total number of operations $\tau$ is bounded by

$$\tau \leqslant n-2c+\sum_{i=1}^{c} n_i^2 + \frac{n(n-1)}{2} = \frac{n(n+1)}{2} - 2c + \sum_{i=1}^{c} n_i^2 \qquad (2.14)$$

but in general it will be much smaller. Indeed the lower bound of the ring operation count is

$$2\sum_{i=1}^{c}\sum_{\nu=1}^{\lfloor \log_2 n_i \rfloor} 2^{\nu}\left\lfloor \frac{n_i}{2^{\nu}} \right\rfloor \geqslant 2\sum_{i=1}^{c}\sum_{\nu=1}^{\lfloor \log_2 n_i \rfloor} (n_i - 2^{\nu})$$

$$\geqslant 2\sum_{i=1}^{c}\left(n_i\lfloor \log_2 n_i \rfloor - 2^{\lfloor \log_2 n_i \rfloor + 1} + 2\right)$$

$$\geqslant 2\sum_{i=1}^{c} n_i\lfloor \log_2 n_i \rfloor - 4(n-c) \qquad (2.15)$$

which restricts $\tau$ to be in the range

$$\frac{1}{2}\left(2 - \sum_{i=1}^{c} n_i^2\right) + 2\sum_{i=1}^{c} n_i\lfloor \log_2 n_i \rfloor - 4(n-c) \leqslant \tau \leqslant \frac{n(n+1)}{2} + \sum_{i=1}^{c} n_i^2 - 2c \qquad (2.16)$$

A point of interest is the variation of $\alpha$ with the ordering of sample elements. For a fixed number of classes $c$, $\alpha$ is minimum when $\sum_{i=1}^{c} n_i^2$ is maximum subject to $\sum_{i=1}^{c} n_i = n$. In c-dimensional Euclidean space, the solutions are points with integer coordinates greater than or equal to 1 in the hyperplane $\sum_{i=1}^{c} n_i - n = 0$. They are of the form $(1, 1, \ldots, 1, n-c+1)$ or any permutation of the vector elements by symmetry; the corresponding $\alpha_{\min}$ is by (2.13)

$$\alpha_{\min} = 1 - \frac{c-1 + (n-c+1)^2 + 2c-3n}{n(n-1)} = 1 - \frac{(n-c)(n-c-1)}{n(n-1)} \qquad (2.17)$$

The algorithm will pay off if $\alpha$ is close to $\alpha_{\min}$ and $\alpha_{\min}$ is itself small, which requires having as few classes as possible, the classes being as large as possible. Consequently, if we have the possibility of ordering the elements within the sample drawn, we should rank them according to the probability that they are in the relation $\mathcal{R}'$ with all of the other sample elements. Notice by the way, that the ordering only matters within each class; as we already pointed out, we cannot avoid performing all negative tests anyway. That problem is somewhat akin to the connector problem of communication theory. Between any two nodes of a given network, a direct connecting line can be established at a given cost and the problem is to draw the cheapest network, the cost of a tree being the sum of the costs of its edges. A minimal cost tree can be simply constructed by choosing at each step, the cheapest connector until a spanning tree is obtained. This spanning tree is called an economy tree and can be shown to be of minimal cost.

The same method could be applied if we were to evaluate $a_i \mathcal{R}' a_j$ for all $a \in \mathcal{A}$, assigning a cost 1 if $a_i \not\mathcal{R}' a_j$, 0 if $a_i \mathcal{R}' a_j$. But, if we try to minimize the number of times $\mathcal{R}'$ has to be computed, only a probabilistic cost can be assigned to each element $a_i \in \mathcal{P}$. This cost can be taken to be

$$p_i = 1 - \frac{1}{N} \sum_{\substack{a_j \in \mathcal{P} \\ a_i \mathcal{R}' a_j}} p(a_j \in \mathcal{A})$$

In the graph $G_N$, this is equivalent to assigning to each vertex a certain weight, depending upon which vertices it has edges in common with. If the sample $\mathcal{A}$ is obtained by uniformly sampling $\mathcal{P}$ with replacement, the weight $w_i$ of $a_i$ is simply related to the local degree $\rho_i$ of $a_i$ by

$$w_i = \frac{1 + \rho_i}{N}$$

and the elements of $\mathcal{A}$ should then be ordered by decreasing weight to make $\alpha$ optimal.

In practice, even though we may not explicitly know the local degrees in the graph $G_N$, we may still be able to estimate $w_i$. For instance, take the simple case where the relation $\mathcal{R}$ is:two integers in the range $[1,N]$ are multiples of each other. $w_i$ is then a decreasing monotone sequence if the integers in the sample $\mathcal{A}$ form an ascending sequence.

## 2.4. Multiple Size Sampling

In the previous paragraphs, we described how the ring algorithm can be used most efficiently by ordering the sample elements. Alternatively, if we do not perform any rearrangement, the algorithm can be used to compute at the same time the number of classes in samples $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$, the $\nu^{th}$ sample being a subset of $\mathcal{A}_n$ with elements $\left\{a_{i_1}, a_{i_2}, \ldots, a_{i_\nu}\right\}$. This constitutes the greatest advantage of algorithm 3 and makes it about an order of magnitude faster than other algorithms for this particular type of multiple sampling.

# CHAPTER III

## Recursive Determination of the Number of Components

### 3.1. Random Graph

Let $G_n$ be a complete graph on $n$ vertices with no slings (a sling is an edge connecting a vertex to itself) and single edges (no edges in parallel). $G_n$ forms an n-clique in Berge's notation. Let $V(G_n)$ be the vertex set of $G_n$, $E(G_n)$ its set of edges; $e(G_n)$ designates the number of elements in $E(G_n)$, i.e., the number of edges.

Our experiment consists of selecting a spanning subgraph $H_n$ of $G_n$ such that

$$V(H_n) \equiv V(G_n)$$

$$E(H_n) \subseteq E(G_n)$$

the probability of choosing any particular subgraph $G_n$ being only dependent upon its number of edges

$$p(H=H_n) = \frac{f\left(e(H_n)\right)}{\binom{\binom{n}{2}}{e(H_n)}} \qquad \text{where} \qquad \sum_{i=0}^{\binom{n}{2}} f(i) = 1 \tag{3.1}$$

The outcome of the experiment is the number of components in the subgraph $H$, i.e., the minimum number of connected subgraphs which span $H$. The probability that the number of components be equal to $\ell$ is of course

$$p(c=\ell) = \sum_{\substack{\forall H_n \ni H_n \subseteq G_n \\ c(H_n)=\ell}} \frac{f\left(e(H_n)\right)}{\binom{\binom{n}{2}}{e(H_n)}} \tag{3.2}$$

A particular case of interest corresponds to the choice

$$f(e) = \binom{\binom{n}{2}}{e} p^{e}(1-p)^{\binom{n}{2}-e} \qquad 0 \leqslant p \leqslant 1 \tag{3.3}$$

which arises when selecting $k$ distinct edges out of $E(G_n)$, each edge having the same probability of selection. We define

<u>Definition</u> - A random n-graph with $e$ edges is a subgraph of a complete graph $G_n$ with exactly $e$ distinct edges obtained by sampling uniformly and without replacement $E(G_n)$.

We now present a computational method to find the probability distribution of the number of components in a random graph.

<u>Theorem</u> - The probability $p(n, c)$ that a random n-graph has $c$ components is given by the recurrence formula

$$p(n, c) = \frac{1}{c} \sum_{\nu=1}^{n-1} \binom{n}{\nu} (1-p)^{\nu(n-\nu)} p(\nu, 1) p(n-\nu, c-1) \tag{3.4}$$

$$c = 2,\ 3,\ \ldots,\ n\ ;\ n \geqslant 2$$

$$p(n, 1) = 1 - \sum_{c=2}^{n} p(n, c) \tag{3.5}$$

proof: consider a random n-graph with $c$ components, obtained by sampling from an n-clique, each edge being chosen with probability $p=p(2,1)$. If $c \geqslant 2$, this graph can be broken into two disjoint subgraphs with, respectively, $\nu$ vertices, one component and $n-\nu$ vertices, $c-1$ components provided that no edge crosses the partition, which occurs with probability $(1-p)^{\nu(n-\nu)}$. For the family of partitions satisfying the same constraint we get

$$\binom{n}{\nu} p^{\nu(n-\nu)} p(\nu, 1) p(n-\nu, c-1)$$

after multiplying by the number of subsets of $\nu$ vertices. These partial results still have to be summed over all possible $\nu$'s, however in the process of singling out every connected subgraph we count each configuration as many times as it has components, that is exactly c times. The theorem follows.

It is interesting to notice that these recurrence formulas must be used in a precise sequence. The computation of the probability that a random graph be strongly connected is based upon the knowledge of the probability distribution for more than one component. Therefore, we proceed as follows:

i) assume all terms $p(k, \nu)$ known for $1 \leqslant k \leqslant \nu \leqslant n-1$

ii) compute $p(n, \nu)$ $\quad 2 \leqslant \nu \leqslant n-1$

iii) subsequently obtain $p(n, 1)$

iv) repeat steps i) through iii) for $n = n+1$

Theorem - The probability $p(n, c, e)$ that a random n-graph with e edges has c components is given by the recurrence formulas

$$p(n, c, e) = \frac{1}{c\binom{\binom{n}{2}}{e}} \sum_{\substack{\nu=1 \\ \nu(n-\nu) \leq \binom{n}{2} - e}}^{n-1}$$

$$\times \sum_{\lambda=0}^{e} \binom{\binom{\nu}{2}}{\lambda}\binom{\binom{n-\nu}{2}}{e-\lambda} p(\nu, 1, \lambda) p(n-\nu, c-1, e-\lambda) \qquad (3.6)$$

$$e = 0, 1, \ldots, \binom{n}{2}; \quad c=2, \ 3, \ldots, n;$$

$$p(n, 1, e) = 1 - \sum_{\nu=2}^{n} p(n, c, \nu) \qquad (3.7)$$

proof: this formula is most easily deduced from formula (3.4), since the probability $p(n, c, e)$ is in fact equal to the coefficient of $p^{e}(1-p)^{\binom{n}{2} - e}$ in $p(n, c)$ divided by the sum of these coefficients for $c = 1, \ 2, \ \ldots, \ n-1$, i. e., $\binom{\binom{n}{2}}{e}$. Alternatively,

$$p(n, c) = \sum_{e=0}^{\binom{n}{2}} p^{e}(1-p)^{\binom{n}{2} - e}\binom{\binom{n}{2}}{e} p(n, c, e) \qquad (3.8)$$

To relieve the burden of notation, let us define

$$N = \binom{n}{2}$$

$$p = p(2, 1)$$

$$q = p(2, 2) = 1-p$$

The probabilities $p(n, c)$ have the following polynomial representation in terms of $p$ and $q$,

$$p(n, c) = \sum_{\nu=0}^{N} \binom{N}{\nu} p(n, c, \nu)\, p^{\nu} q^{N-\nu} \tag{3.9}$$

Another simplification yet can be achieved if we use the new integer function

$$g(n, c, e) = p(n, c, e) \binom{N}{e} \tag{3.10}$$

which satisfies the recurrence formula

$$g(n, c, e) = \frac{1}{c} \sum_{\substack{\nu=1 \\ \nu(n-\nu) \leqslant N-\nu}}^{n-1} \sum_{\lambda=0}^{e} g(\nu, 1, \lambda) g(n-\nu, c-1, e-\lambda) \tag{3.11}$$

It should be clear that those integer coefficients $g(n, c, e)$ actually represent the number of graphs with exactly $e$ edges and $c$ components among all possible $2^N$ subgraphs of $G_n$. From a numerical standpoint, they are most conveniently evaluated for reasonable $N$ as long as we stay within the range of integer arithmetic. Table ( 3.1) shows the values of $p(n, c)$ for $n=1, 2, \ldots, 6$.

Lemma - For any random n-graph, the following relationship holds

$$n-c \leqslant e \leqslant \binom{n-c+1}{2} \tag{3.12}$$

proof: first, examine the inequality on the left; let $n_i$ be the number of vertices in the $i^{th}$ component, thus

$$\sum_{i=1}^{c} n_i = n$$

The minimum number of edges in the $i^{th}$ component is reached for any tree on $n_i$ vertices, the number of edges being then $n_i-1$ ; summing over all components we get

$$e \geqslant \sum_{i=1}^{c} (n_i-1) = n-c$$

Similarly, for the inequality on the right, the maximum number of edges in the $i^{th}$ component corresponds to an $n_i$-clique and has $\binom{n_i}{2}$ edges so that

$$e \leqslant \sum_{i=1}^{c} \binom{n_i}{2} = \frac{1}{2} \left( \sum_{i=1}^{c} n_i^2 - n \right)$$

Now, the maximum of $\sum_{i=1}^{c} n_i^2$ subject to the constraint $\sum_{i=1}^{c} n_i=n$ is attained for a vector of the form $(1, 1, \ldots, 1, n-c+1)$ or any permutation of its elements. See a more general proof in section (4.1) . Finally

$$e \leqslant \frac{1}{2} \left( c-1 + (n-c+1)^2 - n \right) = \binom{n-c+1}{2} \qquad (3.13)$$

This lemma has a direct practical application to the computation of recurrence formulas (3.11) for $g(n, c, e)$. Indeed, by sharpening the limits of summation, it becomes no longer necessary to store and refer to null values of $g(n, c, e)$ ; these would actually be a considerable nuisance if we were to apply formulas (3.11) straightforwardly.

After introducing the bounds just computed, we obtain

$$g(n, c, e) = \frac{1}{c} \sum_{\substack{\nu=1 \\ \nu(n-\nu) \leqslant N-\nu}}^{n-1} \times \sum_{\lambda=\max\left[\nu-1,\, e-\binom{n-\nu-c+2}{2}\right]}^{\min\left[e-n+\nu+c-1,\, \binom{\nu}{2}\right]} g(\nu, 1, \lambda) g(n-\nu, c-1, e-\lambda) \qquad (3.14)$$

$$1 \leqslant c \leqslant n \;\; ; \;\; n-c \leqslant e \leqslant \binom{n-c+1}{2}$$

$g(n, c, e) = 0$ , otherwise

The great savings in computing time and storage space, definitely justify the slight increase in complexity.

## 3.2. Expected Number of Components

The polynomials ( 3.9 ) involve monomials of the form $p^{\alpha}q^{\beta}$ which can be transformed to a single variable $p$ or $q$ , simply by making use of the binomial theorem. This is especially interesting when either $p$ or $q$ is small compared to $1$ and when we compute the expected number of classes. As a matter of fact, we get

$$\begin{aligned} p(n, c) &= \sum_{\nu=0}^{N} \binom{N}{\nu} p(n, c, \nu) p^{\nu} \sum_{k=0}^{N-\nu} \binom{N-\nu}{k} p^{k}(-1)^{k} \\ &= \sum_{\nu=0}^{N} \sum_{k=0}^{N-\nu} (-1)^{k} \binom{N}{\nu} \binom{N-\nu}{k} p(n, c, \nu) \qquad (3.15) \end{aligned}$$

which, after reversing the order of summation, yields

$$p(n, c) = \sum_{k=0}^{N} p^k \left[ \sum_{\nu=0}^{k} \binom{N}{\nu}\binom{N-\nu}{k-\nu} p(n, c, \nu)\, (-1)^{k-\nu} \right]$$

$$= \sum_{k=0}^{N} p^k \left[ \sum_{\nu=0}^{k} (-1)^{k-\nu} \binom{N-\nu}{N-k} g(n, c, \nu) \right] \qquad (3.16)$$

In the variable $q$, the computation proceeds in an identical fashion to give the polynomial

$$p(n, c) = \sum_{k=0}^{N} q^k \left[ \sum_{\nu=0}^{k} (-1)^{k-\nu} \binom{N-\nu}{N-k} g(n, c, N-\nu) \right] \qquad (3.17)$$

which could actually be derived directly from (3.16) since exchanging the roles of $p$ and $q$ is equivalent to permuting the monomials $p^{\alpha}q^{N-\alpha}$ and $p^{N-\alpha}q^{\alpha}$, which in turn means exchanging the coefficients $g(n, c, \nu)$ and $g(n, c, N-\nu)$. The corresponding graph interpretation is simply to replace $H_n$ by its complement graph in $G_n$.

If we use the single variable polynomials, we get for the expected number of components in a random n-graph the following expressions

$$\mathcal{E}\left(c_n(p)\right) = \sum_{k=0}^{N} p^k \left[ \sum_{\nu=0}^{k} (-1)^{k-\nu} \binom{N-\nu}{N-k} \sum_{c=1}^{n} c g(n, c, \nu) \right] \qquad (3.18)$$

$$\mathcal{E}\left(c_n(q)\right) = \sum_{k=0}^{N} q^k \left[ \sum_{\nu=0}^{k} (-1)^{k-\nu} \binom{N-\nu}{N-k} \sum_{c=1}^{n} c g(n, c, N-\nu) \right] \qquad (3.19)$$

For the actual numerical evaluation of these results, we would obviously apply the last lemma, as we formerly did, and use the following formulas

$$\mathcal{E}\left(c_n(p)\right) = \sum_{k=0}^{N} p^k \left[\sum_{c=1}^{n} c \sum_{\nu=n-c}^{\min\left(\binom{n-c+1}{2},\, k\right)} (-1)^{k-\nu}\binom{N-\nu}{N-k} g(n, c, \nu)\right] \quad (3.20)$$

$$\mathcal{E}\left(c_n(q)\right) = \sum_{k=0}^{N} q^k \left[\sum_{c=1}^{n} c \sum_{\nu=N-\binom{n-c+1}{2}}^{\min(N-n+c,\, k)} (-1)^{k-\nu} \binom{N-\nu}{N-k} g(n, c, N-\nu)\right] \quad (3.21)$$

The polynomials p(n, c) can be found in table ( 3. 1) and the expected value of the number of components is given in tables ( 3. 2 ) and ( 3. 3 ). Figure ( 3. 1 ) shows the variation of the expected value for various probability levels from 0 to 1 by increments of .05. It is convenient to define a clustering coefficient $\rho$ by

$$\rho = \frac{\frac{n}{c} - 1}{n - 1}$$

which variation can be found in figure ( 3. 2 ) for n ranging between 2 and 20.

$$p(1,1) = 1$$

$$p(2,1) = p$$
$$p(2,2) = q$$

$$p(3,1) = p^3 + 3p^2q$$
$$p(3,2) = 3pq^2$$
$$p(3,3) = q^3$$

$$p(4,1) = p^6 + 6p^5q + 15p^4q^2 + 16p^3q^3$$
$$p(4,2) = 4p^3q^3 + 15\,p^2q^4$$
$$p(4,3) = 6pq^5$$
$$p(4,4) = q^6$$

$$p(5,1) = p^{10} + 10p^9q + 45p^8q^2 + 120p^7q^3 + 205\,p^6q^4 + 222p^5q^5 + 125p^4q^6$$
$$p(5,2) = 5p^6q^4 + 30p^5q^5 + 85p^4q^6 + 110p^3q^7$$
$$p(5,3) = 10p^3q^7 + 45p^2q^8$$
$$p(5,4) = 10pq^9$$
$$p(5,5) = q^{10}$$

$$p(6,1) = p^{15} + 15p^{14}q + 105p^{13}q^2 + 455p^{12}q^3 + 1365p^{11}q^4 + 2997p^{10}q^5 + 4945p^9q^6 + 6165p^8q^7$$
$$5700p^7q^8 + 3660p^6q^9 + 1296p^5q^{10}$$
$$p(6,2) = 6p^{10}q^5 + 60p^9q^6 + 270p^8q^7 + 735p^7q^8 + 1330p^6q^9 + 1617p^5q^{10} + 1080p^4q^{11}$$
$$p(6,3) = 15p^6q^9 + 90p^5q^{10} + 285p^4q^{11} + 435p^3q^{12}$$
$$p(6,4) = 20p^3q^{12} + 105p^2q^{13}$$
$$p(6,5) = 15pq^{14}$$
$$p(6,6) = q^{15}$$

Table 3.1

RANDOM GRAPH: AVERAGE NUMBER OF COMPONENTS

$$c_1(p) = 1$$
$$c_2(p) = 2 - p$$
$$c_3(p) = 3 - 3p + p^3$$
$$c_4(p) = 4 - 6p + 4p^3 + 3p^4 - 6p^5 + 2p^6$$
$$\begin{aligned} c_5(p) = 5 &- 10p + 10p^3 + 15p^4 - 18p^5 - 60p^6 + 130p^7 - 105p^8 + 40p^9 \\ &- 6p^{10} \end{aligned}$$
$$\begin{aligned} c_6(p) = 6 &- 15p + 20p^3 + 45p^4 - 18p^5 - 330p^6 + 60p^7 + 2445p^8 - 6485p^9 \\ &+ 8712p^{10} - 7206\ p^{11} + 3925p^{12} - 1350p^{13} + 270p^{14} - 24p^{15} \end{aligned}$$
$$\begin{aligned} c_7(p) = 7 &- 21p + 35p^3 + 105p^4 + 42p^5 - 980p^6 - 1950p^7 + 11760p^8 \\ &+ 12355p^9 - 182721p^{10} + 589281p^{11} - 1128820p^{12} + 1502550p^{13} \\ &- 1471305p^{14} + 1084104p^{15} - 603435p^{16} + 250950p^{17} - 75810p^{18} \\ &+ 15750p^{19} - 2016p^{20} + 120p^{21} \end{aligned}$$

Table 3.2

RANDOM GRAPH: AVERAGE NUMBER OF COMPONENTS

$$c_1(q) = 1$$
$$c_2(q) = 1 + q$$
$$c_3(q) = 1 + 3q^2 - q^3$$
$$c_4(q) = 1 + 4q^3 + 3q^4 - 6q^5 + 2q^6$$
$$c_5(q) = 1 + 5q^4 + 10q^6 - 10q^7 - 15q^8 + 20q^9 - 6q^{10}$$
$$\begin{aligned} c_6(q) = 1 &+ 6q^5 + 15q^8 - 5q^9 - 60q^{11} + 25q^{12} + 90q^{13} - 90q^{14} \\ &+ 24q^{15} \end{aligned}$$
$$\begin{aligned} c_7(q) = 1 &+ 7q^6 + 21q^{10} - 21q^{11} + 35q^{12} - 105q^{14} - 105q^{16} + 420q^{17} \\ &- 630q^{19} + 504q^{20} - 120q^{21} \end{aligned}$$
$$\begin{aligned} c_8(q) = 1 &+ 8q^7 + 28q^{12} - 28q^{13} + 56q^{15} + 35q^{16} - 168q^{17} + 112q^{18} \\ &- 280q^{19} - 210q^{20} + 560q^{21} + 140q^{22} + 1680q^{23} - 3150q^{24} \\ &- 1176q^{25} + 5040q^{26} - 3360q^{27} + 720q^{28} \end{aligned}$$

Table 3.3

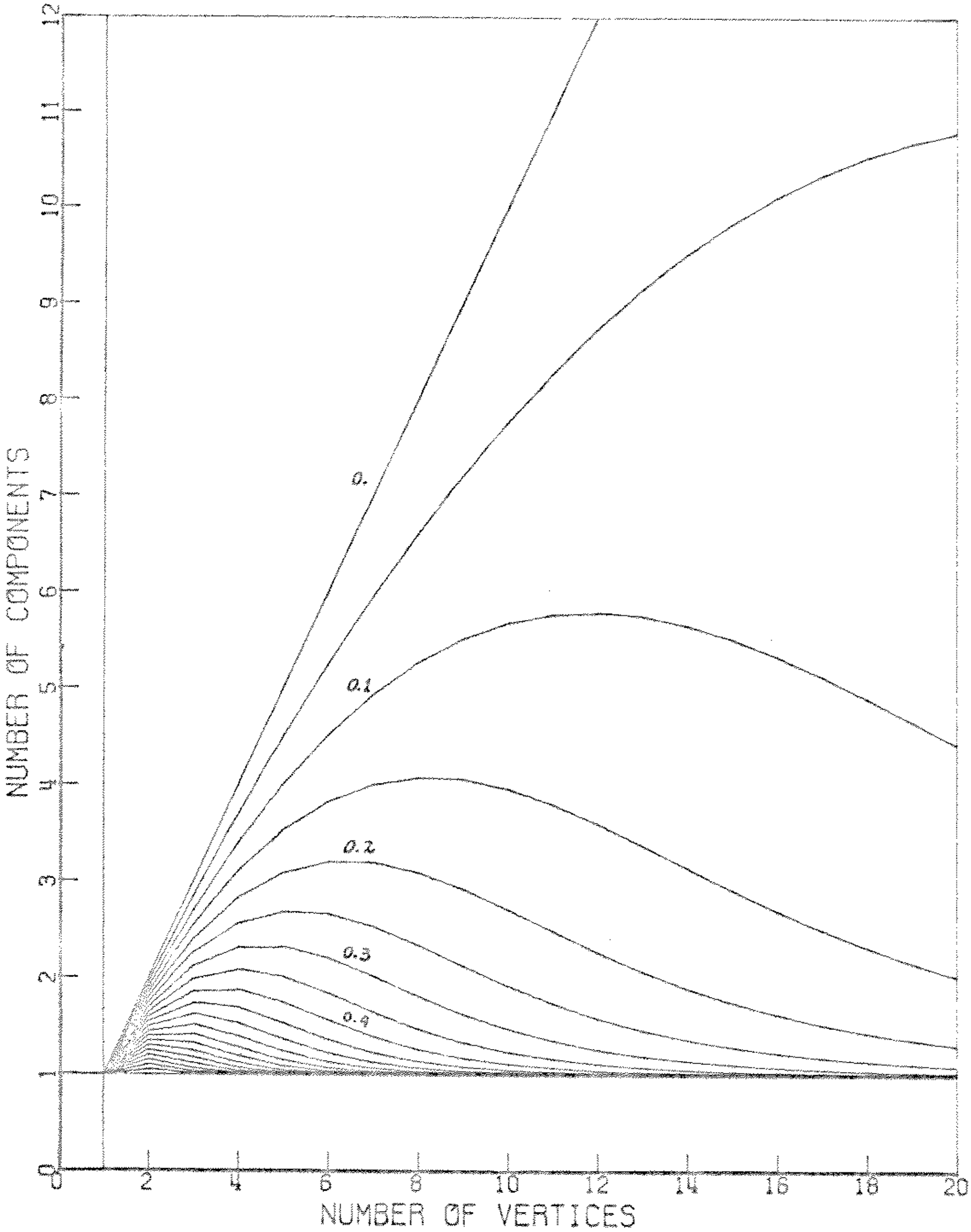

NUMBER OF COMPONENTS
0
1
2
3
4
5
6
7
8
9
10
11
12
NUMBER OF VERTICES
0
2
4
6
8
10
12
14
16
18
20
0.
0.1
0.2
0.3
0.4

Figure 3.1

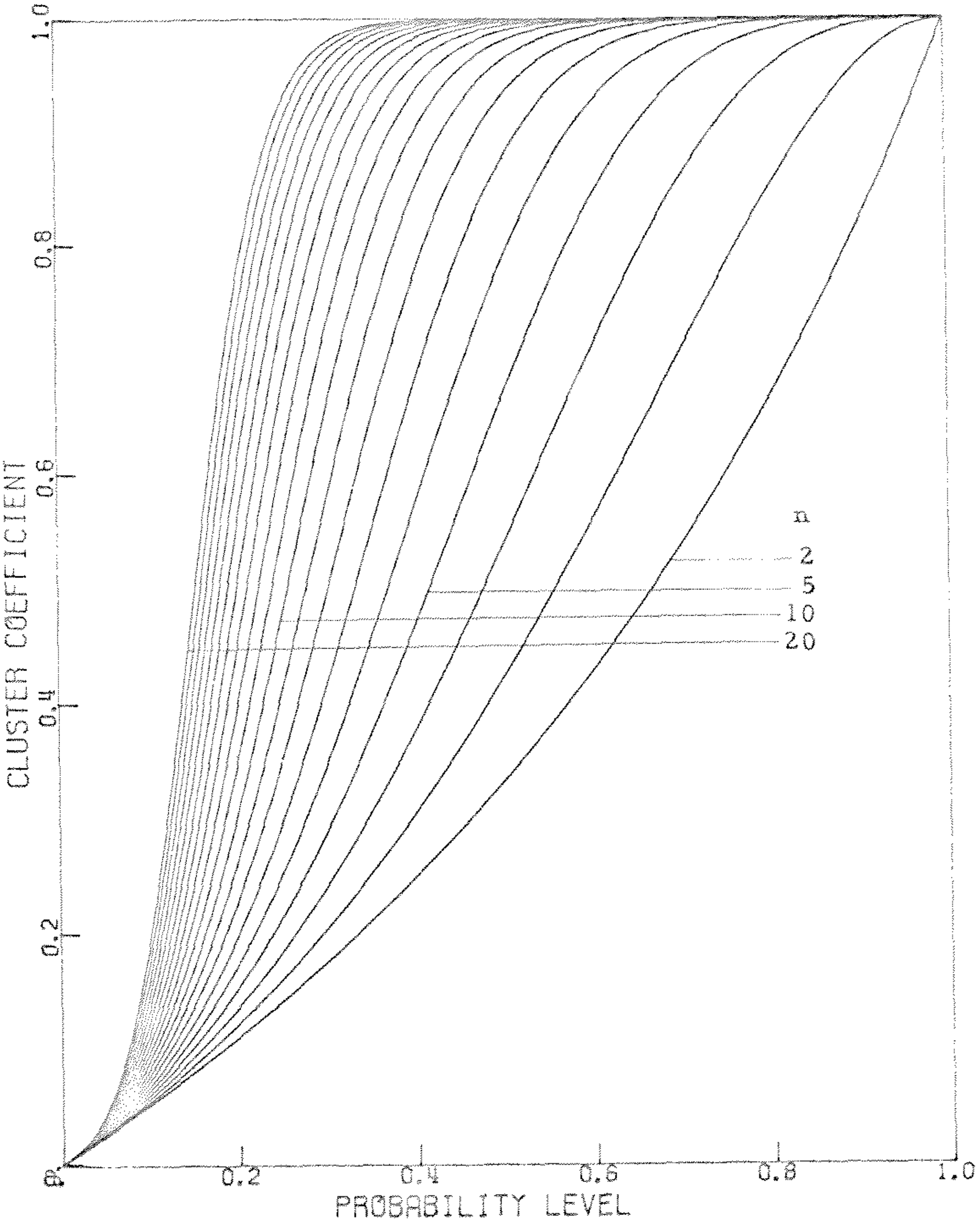

1.0
0.8
0.6
0.4
0.2
CLUSTER COEFFICIENT
0
0.2
0.4
0.6
0.8
1.0
PROBABILITY LEVEL
n
2
5
10
20

Figure 3.2

When either p or q becomes small, the variation of the number of classes can be easily estimated using the following lemma.

Lemma - The expected number of components in a random n-graph is

$$\mathcal{E}\left(c_n(p)\right) = n - \binom{n}{2}p + \binom{n}{3}p^3 + \mathcal{O}(n^4p^4) \qquad (3.22)$$

$$\mathcal{E}\left(c_n(q)\right) = 1 + nq^{n-1} + \mathcal{O}\left(n^2q^{2(n-2)}\right) \qquad (3.23)$$

proof: as a starting point we use expressions (3.20) and (3.21) for the expected values. From our lemma, together with some simple geometrical considerations we know that

$$g(n, 1, N-\nu) = \binom{N}{\nu} \qquad 0 \leqslant \nu \leqslant n-2$$

$$g(n, n-\nu, \nu) = \binom{N}{\nu} \qquad 0 \leqslant \nu \leqslant 2$$

We now compute the coefficients of $p^k$ then $q^k$.

i) polynomial in p :

$$\mathcal{E}\left(c_n(p)\right) = \sum_{k=0}^{N} C_{p,k}\, p^k$$

$$C_{p,0} = ng(n, n, 0) = n$$

$$C_{p,1} = \sum_{c=1}^{n} c \sum_{\nu=n-c}^{\min\left(\binom{n-c+1}{2},1\right)} (-1)^{\nu-1} \binom{N-\nu}{N-1} g(n,c,\nu)$$

$$= -\, n \binom{N}{N-1}\, g(n,n,0) + (n-1)\binom{N-1}{N-1}\, g(n,n-1,1)$$

$$= -\, n N + (n-1)N = -\binom{n}{2}$$

$$C_{p,2} = n\binom{N}{N-2} g(n,n,0) - (n-1)\binom{N-1}{N-2} g(n,n-1,1) + (n-2)\binom{N-2}{N-2} g(n,n-2,2)$$

$$= n\,\frac{N(N-1)}{2} - (n-1)(N-1)N + (n-2)\,\frac{N(N-1)}{2} = 0$$

$$C_{p,3} = -n\binom{N}{N-3} g(n,n,0) + (n-1)\binom{N-1}{N-3} g(n,n-1,1) - (n-2)\binom{N-2}{N-3} g(n,n-2,2)$$

$$+ (n-2)\binom{N-3}{N-3} g(n,n-2,3) + (n-3)\binom{N-3}{N-3} g(n,n-3,3) = g(n,n-2,3)$$

but this term is merely the number of triangles which can be formed with three edges and n vertices, i. e., $C_{p,3} = \binom{n}{3}$. Although we do not explicitly compute $C_{p,4}$, it is clear from the structure of (3.20) that $C_{p,4} = \mathcal{O}(n^4)$ which yields formula (3.22).

ii) polynomial in q:

$$\mathcal{E}\left(c_n(q)\right) = \sum_{k=0}^{N} C_{q,k}\, q^k$$

$$C_{q,0} = g(n,1,0) = 1$$

If $1 \leqslant j < n-1$ we have

$$C_{q,\,j} = \sum_{c=1}^{n} c \sum_{\nu=N-\binom{n-c+1}{2}}^{j} (-1)^{j-\nu}\binom{N-\nu}{N-j}\, g(n,\,c,\,N-\nu)$$

$$= \sum_{\nu=0}^{n-1} (-1)^{j-\nu}\binom{N-\nu}{N-j} g(n,\,1,\,N-\nu)$$

$$= (-1)^{-j} \sum_{\nu=0}^{j} (-1)^{\nu}\binom{N-\nu}{N-j}\binom{N}{\nu} = (-1)^{-j}\binom{0}{j} = 0$$

$$C_{q,\,n-1} = g(n,\,2,\,N-n+1) = n \qquad\qquad n \geqslant 3$$

since it corresponds to one isolated vertex and an $n-1$ clique.

Similarly when $n-1 < j \leqslant 2n-3$, equation (3.21) yields

$$c_{q,\,j} = \sum_{\nu=0}^{k} (-1)^{k-\nu}\binom{N-\nu}{N-k} g(n,\,1,\,N-\nu)$$

$$+\ 2 \sum_{\nu=n-1}^{k} (-1)^{k-\nu}\binom{N-\nu}{N-k} g(n,\,2,\,N-\nu)$$

$$= \sum_{\nu=0}^{k} (-1)^{k-\nu}\binom{N-\nu}{N-k}\binom{N}{\nu}$$

$$= 0$$

| $n,c$ \ $q^{\nu}$ | 1 | $q$ | $q^2$ | $q^3$ | $q^4$ | $q^5$ | $q^6$ | $q^7$ | $q^8$ | $q^9$ | $q^{10}$ | $q^{11}$ | $q^{12}$ | $q^{13}$ | $q^{14}$ | $q^{15}$ | $q^{16}$ | $q^{17}$ | $q^{18}$ | $q^{19}$ | $q^{20}$ | $q^{21}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3, 1 | 1 | 3 | | | | | | | | | | | | | | | | | | | | |
| 3, 2 | | | 3 | | | | | | | | | | | | | | | | | | | |
| 3, 3 | | | | 1 | | | | | | | | | | | | | | | | | | |
| 4, 1 | 1 | 6 | 15 | 16 | | | | | | | | | | | | | | | | | | |
| 4, 2 | | | | 4 | 15 | | | | | | | | | | | | | | | | | |
| 4, 3 | | | | | | 6 | | | | | | | | | | | | | | | | |
| 4, 4 | | | | | | | 1 | | | | | | | | | | | | | | | |
| 5, 1 | 1 | 10 | 45 | 120 | 205 | 222 | 125 | | | | | | | | | | | | | | | |
| 5, 2 | | | | | 5 | 30 | 85 | 110 | | | | | | | | | | | | | | |
| 5, 3 | | | | | | | | 10 | 45 | | | | | | | | | | | | | |
| 5, 4 | | | | | | | | | | 10 | | | | | | | | | | | | |
| 5, 5 | | | | | | | | | | | 1 | | | | | | | | | | | |
| 6, 1 | 1 | 15 | 105 | 455 | 1365 | 2997 | 4945 | 6165 | 5700 | 3660 | 1296 | | | | | | | | | | | |
| 6, 2 | | | | | | 6 | 60 | 270 | 735 | 1330 | 1617 | 1080 | | | | | | | | | | |
| 6, 3 | | | | | | | | | | 15 | 90 | 285 | 435 | | | | | | | | | |
| 6, 4 | | | | | | | | | | | | | 20 | 105 | | | | | | | | |
| 6, 5 | | | | | | | | | | | | | | | 15 | | | | | | | |
| 6, 6 | | | | | | | | | | | | | | | | 1 | | | | | | |
| 7, 1 | 1 | 21 | 210 | 1330 | 5985 | 20349 | 54257 | 116175 | 202755 | 290745 | 343140 | 331506 | 258125 | 156555 | 68295 | 16807 | | | | | | |
| 7, 2 | | | | | | | 7 | 105 | 735 | 3185 | 9576 | 21189 | 35595 | 45990 | 45360 | 32417 | 13377 | | | | | |
| 7, 3 | | | | | | | | | | | | 21 | 210 | 945 | 2625 | 5005 | 6762 | 5250 | | | | |
| 7, 4 | | | | | | | | | | | | | | | | 35 | 210 | 735 | 1295 | | | |
| 7, 5 | | | | | | | | | | | | | | | | | | | 35 | 210 | | |
| 7, 6 | | | | | | | | | | | | | | | | | | | | | 21 | |
| 7, 7 | | | | | | | | | | | | | | | | | | | | | | 1 |

Table ( 3.4 ): $g(n, c, \nu)$

## 3.3. Arbitrary Edge Distribution

The previous results do not exclusively apply to the random graph problem where the probability distribution of the number of edges is binomial but also to more general distributions provided that the probability of any given configuration be uniquely determined by its number of edges, all $2^N$ configurations being feasible. Indeed, knowing $g(n, c, e)$ which is the number of configurations forming $c$ classes for fixed $e$, it is a simple matter to obtain $p(n, c)$ for some edge distribution $f(n, \nu)$ such that $\sum_{\nu=0}^{N} f(n, \nu) = 1$. The probability of $c$ components becomes

$$p(n, c) = \sum_{\nu=0}^{N} f(n, \nu)\, p(n, c, \nu) = \sum_{\nu=0}^{N} \frac{f(n, \nu)\, g(n, c, \nu)}{\binom{N}{\nu}}$$

$$= \sum_{\nu=n-c}^{\binom{n-c+1}{2}} \frac{f(n, \nu)\, g(n, c, \nu)}{\binom{N}{\nu}} \qquad (3.24)$$

and the average number of components is

$$\mathcal{E}(c_n) = \sum_{c=1}^{n} \sum_{\nu=n-c}^{\binom{n-c+1}{2}} c \frac{f(n, \nu)\, g(n, c, \nu)}{\binom{N}{\nu}} \qquad (3.25)$$

All of these values are readily obtained from the table (3.4) and extensions of that table.

## 3.4. Discrete Separable Problems

The approach taken in the case of the random graph can be extended to encompass a wider class of problems that we shall call "separable". The applicability of the method is based upon our ability to express any configuration with $c$ classes as the sum of pairs of configuration with, respectively, $\nu$ and $c-\nu$ classes, $1 \leq \nu \leq c-1$. The set of $n$ points is thus broken into subsets of $k$ and $n-k$ points forming $\nu$ and $c-\nu$ classes. All other parameters describing the configuration must be summed and the method will succeed if all of these sums can be expressed in terms of simpler configurations exclusively, i.e., configurations already evaluated. Notice that the probability of having a single class is always found last by taking the difference between the total number of configurations and the number of configurations having 2, 3, ... or $n$ classes for a fixed set of parameters.

We now investigate in some detail an example of separable problem.

## 3.5. Rooks Problem On a Chessboard

Given an $n \times m$ rectangular board $B_{n,m}$ with positions $b_{ij}$, $1 \leq i \leq n$, $1 \leq j \leq m$, we can speak of the number of rooks classes formed by $k$ marks on $B_{n,m}$; any two marks $b_{i_1,j_1}$, $b_{i_\ell,j_\ell}$ are in the same class if and only if there exists some sequence of marks

$$b_{i_1 j_1}\ b_{i_2 j_2}\ \dots\ b_{i_\ell j_\ell}$$

$$\left.\begin{array}{l}\text{for which either } i_\nu = i_{\nu+1} \text{ and } j_\nu \neq j_{\nu+1} \\ \text{or } i_\nu \neq i_{\nu+1} \text{ and } j_\nu = j_{\nu+1}\end{array}\right\} \quad 1 \leq \nu < \ell$$

Let $g(n, m, c, k)$ be the number of distinct configurations of $k$ marks on $B_{nm}$ forming $c$ rooks classes. Let $d(n, m, c, k)$ be the corresponding count when the configurations are constrained to occupy exactly $n$ rows and $m$ columns; such configurations are called dense. It is obvious that by symmetry

$$g(n, m, c, k) = g(m, n, c, k)$$

$$d(n, m, c, k) = d(m, n, c, k)$$

On $B_{nm}$, the number of configurations can be expressed in terms of dense configurations by

$$g(n, m, c, k) = \sum_{i=1}^{n} \sum_{j=1}^{m} \binom{n}{i}\binom{m}{j} d(i, j, c, k) \qquad (3.26)$$

Later, this expansion in terms of dense configurations will be given sharper limits by weeding out the null $d$ values.

Any configuration forming $c$ classes, $c \geq 2$, can actually be decomposed into a pair of patterns on smaller boards $B_{n_1 m_1}$ and $B_{n-n_1\, m-m_1}$ called sub-boards, which together comprise all of the $k$ marks (that is, the complement of the union of the sub-boards in $B_{nm}$ contains no mark) and furthermore, each class is completely included in only one of the two sub-boards. In general this decomposition is not unique.

Theorem - The number $g(n, m, c, k)$ of configurations of $k$ marks forming $c$ classes on $B_{nm}$ is given by the recurrence relations

$$g(n, m, c, k) = \frac{1}{c} \sum_{n_1=1}^{n-1} \sum_{m_1=1}^{m-1} \sum_{k_1=1}^{k-1} \sum_{n_2=1}^{n-n_1} \sum_{m_2=1}^{m-m_1} \times \binom{n_1+n_2}{n_1}\binom{m_1+m_2}{m_1}\binom{n}{n_1+n_2}\binom{m}{m_1+m_2} \times d(n_1, m_1, 1, k_1)\, d(n_2, m_2, c-1, k-k_1) \tag{3.27}$$

$$2 \leqslant c \leqslant \min(n, m)$$

$$g(n, m, 1, k) = \binom{nm}{k} - \sum_{\nu=2}^{\min(n,m)} g(n, m, \nu, k) \tag{3.28}$$

$$d(n, m, c, k) = g(n, m, c, k) - \sum_{\substack{i=c \\ ij < nm}}^{n} \sum_{j=c}^{m} \binom{n}{i}\binom{m}{j} d(i, j, c, k) \tag{3.29}$$

$$1 \leqslant k \leqslant nm \quad , \qquad 1 \leqslant c \leqslant \min(n, m)$$

$$g(1, 1, 1, 1) = d(1, 1, 1, 1) = 1 \tag{3.30}$$

proof: let us first establish how to put $B_{nm}$ in a canonical order. Pick the first mark $b_{i_1 j_1}$ encountered when scanning successively the first row, then first column, second row then second column, ....

The mark $s_{i_1 j_1}$ is a member of some class which has row set $\{i_1 i_2 \ldots i_{n_1}\}$ and column set $\{j_1 j_2 \ldots j_{m_1}\}$. Appropriate permutations are then applied to the rows and columns in order to produce row sum and column sum vectors

$$\{i_1, i_2, \ldots, i_{n_1}, i_1', \ldots, i_{n-n_1}'\}$$

$$\{j_1, j_2, \ldots, j_{m_1}, j_1', \ldots, j_{m-m_1}'\}$$

satisfying $\quad i_1 < i_2 < \ldots < i_{n_1}\,,\;\; i_1' < i_2' < \ldots < i_{n-n_1}'$

$$j_1 < j_2 < \ldots < j_{m_1}\,,\;\; j_1' < j_2' < \ldots < j_{m-m_1}'$$

The sub-board $B_{n_1 m_1}$ now contains a unique class forming a dense configuration. If we repeat this process for $B_{n-n_1\, m-m_1}$ until all classes have been transformed into their dense equivalent, we say that $B_{nm}$ is now in canonical form. Designate by $B_{n_2\, m_2}$ the dense equivalent of $B_{n-n_1\, m-m_1}$. Each dense configuration can now be expanded into a number of equivalent configurations by appropriately permuting rows and columns independently. However, to forbid generating the same configuration in several ways, the permutations must preserve the order of the rows and columns of $B_{n_1 m_1}$ among themselves; this also holds for those of $B_{n_2 m_2}$ and for the empty rows and columns $B_{n-n_1-n_2,\, m-m_1-m_2}$. The coefficient in the summation is then simply the product of the number of ordered partitions $(n_1, n_2, n-n_1-n_2)$ of the row and column set of $B_{n_1 m_1}$ which is

$$\frac{n!}{n_1!n_2!(n-n_1-n_2)!}\ \frac{m!}{m_1!m_2!(m-m_1-m_2)!}=\binom{n_1+n_2}{n_1}\binom{m_1+m_2}{m_1}\binom{n}{n_1+n_2}\binom{m}{m_1+m_2}$$

Still, some identical configurations are multiply counted since $B_{n_1m_1}$ is going to represent successively each one of the $c$ classes on $B_{nm}$. Thus, formulas (3.27) and (3.28) follow immediately. Formula (3.29) derives the number of dense configurations on $B_{nm}$ by removing from the total number of configurations of $k$ marks forming $c$ classes, those configurations which have a dense representation on proper sub-boards of $B_{nm}$.

Lemma - The only possibly non-vanishing terms $g(n, m, c, k)$ and $d(n, m, c, k)$ occur under the following conditions

$$\begin{array}{lll} g(n, m, c, k) & \text{if} & c \leqslant k \leqslant nm - (c-1)(n+m-c) \\ & & 1 \leqslant c \leqslant \min(n, m) \\ & & \\ d(n, m, c, k) & \text{if} & n+m-c \leqslant k \leqslant nm - (c-1)(n+m-c) \\ & & 1 \leqslant c \leqslant \min(n, m) \end{array}$$

proof: obvious from geometrical considerations.

Using that lemma, formulas (3.27) through (3.29) can be made more efficient computationally by sharpening the limits of summation, thus reducing greatly the numerical toil. We obtain

$$g(n, m, c, k)=\frac{1}{c}\sum_{n_1=1}^{n-1}\ \sum_{m_1=1}^{m-1}\ \sum_{k_1=n_1+m_1-1}^{n_1m_1}\ \sum_{n_2=c-1}^{n-n_1}\ \sum_{m_2=c-1}^{\min(m-m_1,\, c-1+k-k_1-n_2)}$$

$$\times\binom{n_1+n_2}{n_1}\binom{m_1+m_2}{m_1}\binom{n}{n_1+n_2}\binom{m}{m_1+m_2}$$

$$\times\ d(n_1, m_1, 1, k_1)\ d(n_2, m_2, c-1, k-k_1) \qquad (3.31)$$

$$2 \leqslant c \leqslant \min(n, m)$$

$$g(n, m, 1, k) = \binom{nm}{k} - \sum_{c=2}^{\min(n, m)} g(n, m, c, k) \tag{3.32}$$

$$d(n, m, c, k) = g(n, m, c, k) - \sum_{\substack{i=c \\ ij <}}^{n} \sum_{\substack{j=c \\ nm}}^{\min(m, c+k-i)} \binom{n}{i}\binom{m}{j} d(i, j, c, k) \tag{3.33}$$

$$g(1, 1, 1, 1) = d(1, 1, 1, 1) = 1 \tag{3.34}$$

These formulas must of course be used in a definite sequence. Assume all values $g(n, m, c, k)$ and $d(n, m, c, k)$ known for $nm \leq N^2$.

Using formula (3.31), we can compute $g(n, m, c, k)$ for $n = N+1$, $m = 1, 2, \ldots, N+1$ and all permissible values of $c$ and $k$. This process requires only sub-boards $nm \leq N^2$ and is therefore successful.

Then (3.32) gives $g(n, m, 1, k)$ for all the newly computed nxm boards. Finally (3.33) produces $d(n, m, c, k)$ which will be required in the computation of larger boards. All values of $d$ and $g$ for $nm \leq (N+1)^2$ are now known, which completes the induction proof.

It is clear that by symmetry we only need to compute explicitly $g(n, m, c, k)$ and $d(n, m, c, k)$ for $n \geq m$.

Of course, the previous results can be interpreted as the probability that $k$ marks distributed at random on $B_{nm}$ form $c$ rooks classes if we introduce

$$p(n, m, c, k) = \frac{g(n, m, c, k)}{\binom{nm}{k}} \qquad (3.35)$$

and formulas (3.31) through (3.34) can then be rewritten in terms of $p(n, m, c, k)$.

We can also proceed as in the random graph case and derive class polynomials by assigning to each configuration of $k$ marks a probability $\binom{nm}{k} p^k (1-p)^{nm-k}$. Let $q = 1-p$, we get

$$\begin{aligned} p(n, m, c) &= \sum_{k=c}^{nm-(c-1)(n+m-c)} \binom{nm}{k} p^k q^{nm-k} p(n, m, c, k) \\ &= \sum_{k=c}^{nm-(c-1)(n+m-c)} p^k q^{nm-k} g(n, m, c, k) \qquad (3.36) \end{aligned}$$

Transformation to a single variable polynomial is straightforward by means of

$$\begin{aligned} \sum_{i=0}^{n} a_i p^i q^{n-i} &= \sum_{i=0}^{n} p^i \sum_{j=0}^{i} a_j (-1)^{i-j} \binom{n-j}{i-j} \\ &= \sum_{i=0}^{n} q^i \sum_{j=0}^{i} a_{n-j} (-1)^{i-j} \binom{n-j}{i-j} \end{aligned}$$

We obtain the following polynomials in $p$ or $q$, expressing the expected number of classes:

## Table 3.5

## RECTANGULAR BOARD, ROOKS POLYNOMIALS

$p(1,1,0) = q$

$p(1,1,1) = p$

$p(2,1,0) = q^2$

$p(2,1,1) = 2pq + p^2$

$p(2,2,0) = q^4$

$p(2,2,1) = 4pq^3 + 4p^2q^2 + 4p^3q + p^4$

$p(2,2,2) = 2p^2q^2$

$p(3,1,0) = q^3$

$p(3,1,1) = 3pq^2 + 3p^2q + p^3$

$p(3,2,0) = q^6$

$p(3,2,1) = 6pq^5 + 9p^2q^4 + 14p^3q^3 + 15p^4q^2 + 6p^5q + p^6$

$p(3,2,2) = 6p^2q^4 + 6p^3q^3$

$p(3,3,0) = q^9$

$p(3,3,1) = 9pq^8+18p^2q^7+42p^3q^6+81p^4q^5+117p^5q^4+84p^6q^3+36p^7q^2+9p^8q+p^9$

$p(3,3,2) = 18p^2q^7 + 36p^3q^6 + 45p^4q^5 + 9p^5q^4$

$p(3,3,3) = 6p^3q^6$

$p(4,1,0) = q^4$

$p(4,1,1) = 4pq^3 + 6p^2q^2 + 4p^3q + p^4$

$p(4,2,0) = q^8$

$p(4,2,1) = 8pq^7+16p^2q^6+32p^3q^5+56p^4q^4+56p^5q^3+28p^6q^2+8p^7q+p^8$

$p(4,2,2) = 12p^2q^6 + 24p^3q^5 + 14p^4q^4$

$p(4,3,0) = q^{12}$

$p(4,3,1) = 12pq^{11}+30p^2q^{10}+88p^3q^9+237p^4q^8+528p^5q^7+834p^6q^6+780p^7q^5$
$495p^8q^4+220p^9q^3+66p^{10}q^2+12p^{11}q+p^{12}$

$p(4,3,2) = 36p^2q^{10}+108p^3q^9+222p^4q^8+264p^5q^7+90p^6q^6+12p^7q^5$

$p(4,3,3) = 24p^3q^9 + 36p^4q^8$

$p(4,4,0) = q^{16}$

$p(4,4,1) = 16pq^{15}+48p^2q^{14}+176p^3q^{13}+620p^4q^{12}+1968p^5q^{11}+5040p^6q^{10}$
$+9664p^7q^9+12228p^8q^8+11296p^9q^7+7992p^{10}q^6+4368p^{11}q^5$
$+1820p^{12}q^4+560p^{13}q^3+120p^{14}q^2+16p^{15}q+p^{16}$

Table 3.5a

RECTANGULAR BOARD, ROOKS POLYNOMIALS

$p(4,4,2) = 72p^2q^{14}+288p^3q^{13}+888p^4q^{12}+1968p^5q^{11}+2896p^6q^{10}+1776p^7q^9$
$\quad +642p^8q^8+144p^9q^7+16p^{10}q^6$

$p(4,4,3) = 96p^3q^{13}+288p^4q^{12}+432p^5q^{11}+72p^6q^{10}$

$p(4,4,4) = 24p^4q^{12}$

$p(5,1,0) = q^5$

$p(5,1,1) = 5pq^4+10p^2q^3+10p^3q^2+5p^4q+p^5$

$p(5,2,0) = q^{10}$

$p(5,2,1) = 10pq^9+25p^2q^8+60p^3q^7+140p^4q^6+222p^5q^5+210p^6q^4+120p^7q^3$
$\quad +45p^8q^2+10p^9q+p^{10}$

$p(5,2,2) = 20p^2q^8+60p^3q^7+70p^4q^6+30p^5q^5$

$p(5,3,0) = q^{15}$

$p(5,3,1) = 15pq^{14}+45p^2q^{13}+155p^3q^{12}+525p^4q^{11}+1533p^5q^{10}+3580p^6q^9$
$\quad +5805p^7q^8+6285p^8q^7+4990p^9q^6+3003p^{10}q^5+1365p^{11}q^4$
$\quad +455p^{12}q^3+105p^{13}q^2+15p^{14}q+p^{15}$

$p(5,3,2) = 60p^2q^{13}+240p^3q^{12}+660p^4q^{11}+1320p^5q^{10}+1425p^6q^9+630p^7q^8$
$\quad +150p^8q^7+15p^9q^6$

$p(5,3,3) = 60p^3q^{12}+180p^4q^{11}+150p^5q^{10}$

## Table 3.6

## AVERAGE NUMBER OF CLASSES

$$c(1,1) = p$$

$$c(2,1) = 2p-p^{2}$$

$$c(2,2) = 4p-4p^{2}+p^{4}$$

$$c(3,1) = 3p-3p^{2}+p^{3}$$

$$c(3,2) = 6p-9p^{2}+2p^{3}+3p^{4}-p^{6}$$

$$c(3,3) = 9p-18p^{2}+6p^{3}+9p^{4}-18p^{7}+18p^{8}-5p^{9}$$

$$c(4,1) = 4p-6p^{2}+4p^{3}-p^{4}$$

$$c(4,2) = 8p-16p^{2}+8p^{3}+4p^{4}-4p^{6}+p^{8}$$

$$\begin{aligned} c(4,3) = {} & 12p-30p^{2}+16p^{3}+15p^{4}+6p^{6}-72p^{7}+39p^{8}+100p^{9}-144p^{10}+72p^{11} \\ & -13p^{12} \end{aligned}$$

$$\begin{aligned} c(4,4) = {} & 16p-48p^{2}+32p^{3}+28p^{4}+48p^{6}-288p^{7}+84p^{8}+208p^{9}+1608p^{10} \\ & -5472p^{11}+7576p^{12}-5856p^{13}+2664p^{14}-672p^{15}+73p^{16} \end{aligned}$$

$$c(5,1) = 5p-10p^{2}+10p^{3}-5p^{4}+1$$

$$c(5,2) = 10p-25p^{2}+20p^{3}+2p^{3}+2p^{5}-10p^{6}+5p^{8}-p^{10}$$

$$\begin{aligned} c(5,3) = {} & 15p-45p^{2}+35p^{3}+15p^{4}+3p^{5}+20p^{6}-180p^{7}+15p^{8}+550p^{9}-573p^{10} \\ & -210p^{11}+775p^{12}-600p^{13}+210p^{14}-29p^{15} \end{aligned}$$

Table 3.7

AVERAGE NUMBER OF CLASSES

$c(1,1) = 1-q$

$c(2,1) = 1-q^2$

$c(2,2) = 1+2q^2-4q^3+q^4$

$c(3,1) = 1-q^3$

$c(3,2) = 1+6q^3-12q^4+6q^5-q^6$

$c(3,3) = 1+9q^4-42q^6+54q^7-27q^8+5q^9$

$c(4,1) = 1-q^4$

$c(4,2) = 1+14q^4-32q^5+24q^6-8q^7+q^8$

$c(4,3) = 1+12q^5+6q^6-24q^7-96q^8+240q^9-210q^{10}+84q^{11}-13q^{12}$

$c(4,4) = 1+16q^6-16q^7+66q^8-96q^9-152q^{10}+48q^{11}+972q^{12}-1760q^{13}$
$+1344q^{14}-496q^{15}+73q^{16}$

$c(5,1) = 1-q^5$

$c(5,2) = 1+30q^5-80q^6+80q^7-40q^8+10q^9-q^{10}$

$c(5,3) = 1+15q^6+15q^7-30q^8-45q^9-210q^{10}+855q^{11}-1110q^{12}+705q^{13}$
$-225q^{14}+29q^{15}$

$$\mathscr{E}(c_{nm}) = \sum_{c=1}^{\min(n,m)} c \sum_{i=c}^{nm} p^i \sum_{j=0}^{i} g(n, m, c, j)\,(-1)^{i-j} \binom{nm-j}{i-j} \qquad (3.37)$$

$$= \sum_{c=1}^{\min(n,m)} c \sum_{i=0}^{nm} q^i \sum_{j=0}^{\min(i,\, nm-c)} g(n, m, c, nm-j)(-1)^j \binom{nm-j}{i-j} \qquad (3.38)$$

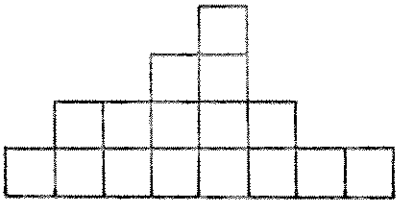
Tables ( 3. 5 ) and (3. 5a) list the rooks polynomials for each class while tables (3. 6) and ( 3. 7 ) contains the corresponding class polynomials.

## 3. 6. Interface Counting

Often however, the foregoing method may not be applicable if the problem is not of the separable type. Notwithstanding, even if an elegant analytical approach seems remote, numerical treatment can yield fast and accurately, answers for small cases; some insight into the behavior of the general solutions may hopefully be gained from these results. We illustrate these remarks with some rooks problem.

Suppose that the board $B_{nm}$ has a set of arbitrary restricted positions and the rooks relationship holds across them. We can think of building $B_{nm}$ by successively appending 1 row to $B_{im}$ , $1 \leq i \leq n-1$ ; the characteristic feature of this problem is that, for the purpose of class counting, configurations of $k$ marks forming $c$ classes on $B_{im}$ can be described in terms of $g(i, c, k, \vec{W})$ where $\vec{W}$ is an m-dimensional column assignment vector. The $j^{th}$ component of $\vec{W}$ is $w_j = \nu$ if the $j^{th}$ column belongs to the $\nu^{th}$ class; classes are ranked according to the order in which their first column is encountered when successively looking at columns 1 through $m$. When $r$ marks are distributed in the unrestricted positions of the newly appended row, a configuration with $k+r$ marks, $c'$ classes and column assignment vector $\vec{W}'$ results; this configuration is thus counted in the term $g(i+1, c', k+r, \vec{W}')$.

Computations of that type are bound to be rather lengthy since they simply provide an ordered way to review all feasible configurations. However, significant improvements can be achieved if it is possible to lump together groups of configurations. For instance, consider a rooks problem on a semi-triangular board



whereby row $i+1$ has the same column set as row $i$ plus $m(i+1)-m(i)$ extra positions. The new feature of this problem lies in that configurations of $k$ marks and $c$ classes on $B_{i,m(i)}$ can be counted in terms of $g(i, c, k, \vec{V})$ where

$$\vec{V} = \begin{pmatrix} v_o \\ v_1 \\ v_c \end{pmatrix}, \quad v_1 \leq v_2 \leq \ldots \leq v_c, \quad \sum_{j=1}^{c} v_j = m(i) \quad ;$$

here $v_o$ designates the number of empty columns and $v_j$, $j=1,\ldots,c$

the number of columns occupied by the $j^{th}$ class when those classes have been ranked according to the number of columns they own. As a new row containing $r$ marks is appended, we have to take into consideration all $\prod_{j=0}^{c} \binom{v_j}{r_j}$ ways in which $r_j$ marks can fall into the set of $v_j$ columns, $j = 0, 1, \ldots, c$, subject to the constraints

$$\sum_{j=0}^{c} r_j = r$$

$$r_j \leqslant v_j \qquad j = 0, 1, \ldots, c$$

Let $r_{j_1}, r_{j_2}, \ldots, r_{j_s}$ be the only non-zero $r$ values for some particular choice of $r's$. Classes $j_1, j_2, \ldots, j_s$ collapse to a single class and

$$g\left(i+1,\ c+1-s,\ k+r, \begin{pmatrix} v_o'' \\ v_1'' \\ \cdots \\ v_{c+1-s}'' \end{pmatrix}\right) \longleftarrow g\left(i, c, k, \begin{pmatrix} v_o \\ v_1 \\ \cdots \\ v_c \end{pmatrix}\right) \prod_{j=0}^{c} \binom{v_j}{r_j} \qquad (3.39)$$

where

$$v_o'' = v_o - r_o$$

$$\left.\begin{array}{lll} v_j' = v_j & \text{if} & r_j = 0 \\ v_j' = 0 & \text{if} & r_j \neq 0 \end{array}\right\} j = 1, 2, \ldots, c$$

$$v_{c+1}' = v_{j_1} + v_{j_2} + \ldots + v_{j_s} + r_o$$

and $v_1'', \ldots, v_{c+1-s}''$ is the sequence obtained from $v_1', v_2', \ldots, v_{c+1}'$ by ordering its elements into non-decreasing order, null components being removed. The arrow in ( 3.39) means that the term on the right contributes additively to the value of the term on the left which had an initial null value.

The basic step we have just described is to be repeated for $0 \leq r \leq m(i)$ and for any such $r$, all partitions into $c+1$ groups or less must be generated. Because of the special role played by the $v_o$ empty columns, it is convenient to split $r$ into $r_o$ and $r-r_o$ marks; this latter group is broken into $r_{j_1}, r_{j_2}, \ldots, r_{j_s}$ marks subject to the above constraints; this decomposition can be carried out in $\alpha(r, r_o; \vec{S}, \vec{V})$ ways where

$$\alpha(r, r_o; \vec{S}, \vec{V}) = \binom{v_o}{r_o} \underbrace{\sum_{r_{j_1}=1}^{v_{j_1}} \sum_{r_{j_2}=1}^{v_{j_2}} \cdots \sum_{r_{j_s}=1}^{v_{j_s}}}_{\sum_{\nu=1}^{s} r_{j_\nu} = r-r_o} \prod_{\nu=1}^{s} \binom{v_{j\nu}}{r_{j\nu}} \qquad (3.40)$$

which would numerically be computed using the formula

$$\alpha(r, r_o; \vec{S}, \vec{V}) = \binom{v_o}{r_o} \sum_{r_{j_1}=1}^{\min(v_{j_1}, r-r_o)} \; \sum_{r_{j_2}=1}^{\min(v_{j_2}, r-r_o-r_{j_1})} \cdots \sum_{r_{j_{s-1}}=1}^{\min(v_{j_s}, r-r_o-r_{j_1}-\ldots-r_{j_{s-2}})}$$

$$\times \binom{v_{j_s}}{r-r_o-r_{j_1}-\ldots-r_{j_{s-1}}} \prod_{\nu=1}^{s-1} \binom{v_{j\nu}}{r_{j\nu}} \qquad (3.41)$$

| n, c \ k | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2, 1 | 3 | 2 | 1 | | | | | | | | | | | | | | | | | | |
| 2, 2 | | 1 | | | | | | | | | | | | | | | | | | | |
| 3, 1 | 6 | 8 | 11 | 11 | 6 | 1 | | | | | | | | | | | | | | | |
| 3, 2 | | 7 | 8 | 4 | | | | | | | | | | | | | | | | | |
| 3, 3 | | | 1 | | | | | | | | | | | | | | | | | | |
| 4, 1 | 10 | 20 | 45 | 91 | 147 | 165 | 114 | 45 | 10 | 1 | | | | | | | | | | | |
| 4, 2 | | 25 | 60 | 96 | 94 | 45 | 6 | | | | | | | | | | | | | | |
| 4, 3 | | | 15 | 22 | 11 | | | | | | | | | | | | | | | | |
| 4, 4 | | | | 1 | | | | | | | | | | | | | | | | | |
| 5, 1 | 15 | 40 | 125 | 381 | 1038 | 2336 | 4061 | 5133 | 4586 | 2924 | 1358 | 455 | 105 | 15 | 1 | | | | | | |
| 5, 2 | | 65 | 240 | 661 | 1397 | 2146 | 2157 | 1279 | 419 | 79 | 7 | | | | | | | | | | |
| 5, 3 | | | 90 | 292 | 515 | 497 | 217 | 23 | | | | | | | | | | | | | |
| 5, 4 | | | | 31 | 52 | 26 | | | | | | | | | | | | | | | |
| 5, 5 | | | | | 1 | | | | | | | | | | | | | | | | |
| 6, 1 | 21 | 70 | 280 | 1141 | 4389 | 15026 | 43748 | 103860 | 194157 | 280403 | 314162 | 278449 | 198864 | 115305 | 54135 | 20341 | 5985 | 1330 | 210 | 21 | 1 |
| 6, 2 | | 140 | 700 | 2751 | 9051 | 24473 | 51878 | 81658 | 90763 | 69871 | 38165 | 15451 | 4626 | 975 | 129 | 8 | | | | | |
| 6, 3 | | | 350 | 1792 | 5670 | 12438 | 18485 | 17110 | 8938 | 2442 | 389 | 30 | | | | | | | | | |
| 6, 4 | | | | 301 | 1176 | 2212 | 2112 | 862 | 72 | | | | | | | | | | | | |
| 6, 5 | | | | | 63 | 114 | 57 | | | | | | | | | | | | | | |
| 6, 6 | | | | | | 1 | | | | | | | | | | | | | | | |

g(n, c, k)

Table 3. 8

Table 3.9

AVERAGE NUMBER OF COMPONENTS

$$c_1(p) = p$$

$$c_2(p) = 3p - 2p^2$$

$$c_3(p) = 6p - 8p^2 + 2p^3 + p^4$$

$$c_4(p) = 10p - 20p^2 + 10p^3 + 3p^4 - 4p^7 + 2p^8$$

$$\begin{aligned} c_5(p) = {} & 15p-40p^2+30p^3+3p^4+2p^5+8p^6-46p^7+10p^8+39p^9+6p^{10}-56p^{11} \\ & +38p^{12}-8p^{13} \end{aligned}$$

$$\begin{aligned} c_6(p) = {} & 21p - 70p^2+70p^3-7p^4+14p^5+54p^6-238p^7-4p^8+71p^9+1233p^{10} \\ & -1822p^{11}-798p^{12}+3370p^{13}-1733p^{14}-1806p^{15}+2922p^{16}-1704p^{17} \\ & +484p^{18}-56p^{19} \end{aligned}$$

Table 3.10

AVERAGE NUMBER OF COMPONENTS

$$c_1(q) = 1 - q$$

$$c_2(q) = 1 + q - 2q^2$$

$$c_3(q) = 1 + 4q^2 - 6q^3 + q^4$$

$$c_4(q) = 1 + 6q^3 + 3q^4 - 28q^5 + 28q^6 - 12q^7 + 2q^8$$

$$\begin{aligned} c_5(q) = {} & 1 + 7q^4+2q^5+14q^6-46q^7-95q^8+341q^9-390q^{10}+224q^{11}-66q^{12} \\ & +8q^{13} \end{aligned}$$

$$\begin{aligned} c_6(q) = {} & 1+8q^5+q^6+41q^8-37q^9-120q^{10}-276q^{11}+515q^{12}+2202q^{13}-7031q^{14} \\ & +8910q^{15}-6258q^{16}+2568q^{17}-580q^{18}+56q^{19} \end{aligned}$$

Of course, when

$$r-r_o \leq \min_{\nu=1,\dots,s} v_{j\nu}$$

the coefficient $\alpha(r, r_o; \vec{S}, \vec{V})$ can simply be expressed using Stirling numbers of the 2nd kind as $\binom{v_o}{r_o}\left\{\begin{matrix} r-r_o \\ s \end{matrix}\right\}$.

An example of the application of this method, the number of configurations on a triangular board $\left(m(i+1)=m(i)+1\ ,\ 1 \leq i \leq 5\ ,\ m(1) = 1\right)$ with $k$ marks and $c$ classes is presented in table ( 3.8). From these coefficients, we derive the expected number of classes as a function of the probability $p$ that any given position on the board be selected. The resulting polynomials in $p$ and $q=1-p$ are shown in table ( 3.9) and (3.10).

## 3.7. Implementation of the Interface Algorithm

Without describing the chores of actual machine computation, a few remarks are still worth making. The integer function $g$ in formula (3.39) is indexed by the vector $\vec{V}$ which is of variable dimension. Although one could assume that $\vec{V}$ has a fixed maximum dimension with a number of zero components, a more refined (thus less costly) approach is to consider this problem as the computation of a functional and implement it with standard list processing techniques. With each triplet $\{n, k, c\}$ is associated a tree of vectors $\vec{V}$ such that the branching decision on the $\ell^{th}$ level is based upon the component $v_\ell$ once a given sequence $v_1, v_2, \ldots v_{\ell-1}$ has been encountered; then the actual g values are at the terminal nodes. Processes are provided for creating and updating trees as new $\vec{V}$'s are generated. One may of course look at those trees as sub-trees of one unique tree, the first three levels of which contain the n, k, c links, respectively.

# CHAPTER IV

## Estimation of the Number of Components

In the preceding chapter, combinatorial solutions were found which led to an exact determination of the expected value of the number of components. Unfortunately, this situation only occurs in rather simple instances, in most cases the problem is either too complex or uncompletely specified so that we can only aim at estimates of the probable number of components.

From a graph theoretic standpoint, we first investigate the relationship between bounds on the local degrees and bounds for the number of edges and components. If we actually choose a particular sequence of local degrees, we can either derive bounds for the expected number of components or perform a Monte Carlo sampling of the space of graphs with prescribed degree sequence in order to obtain the expected value with any degree of accuracy.

Last, we present a conjecture for the general behavior of the expected number of components provided that the sampled space $\mathcal{P}_N$ forms a connected graph. If this is not the case, the result of sampling $\mathcal{P}_N$ can be simply expressed in terms of the individual outcomes in each one of its components.

## 4.1. A Priori Estimates of the Number of Classes

Let us examine under what conditions we can derive estimates for the number of classes without actually performing a complete counting. This can be especially useful if the number of vertices of the graph is extremely large, making the full computation too costly or if the graph is only known from a statistical standpoint, for instance, by the distribution of its local degrees.

We first determine the relationship between the number of vertices, edges and components in a graph $G_n$.

Theorem - Let $G_n$ be a graph with single edges and no loops having $n$ vertices and $c$ connected components. If the local degrees $\rho$ for all vertices satisfy $\rho_\ell \leq \rho \leq \rho_u$ where $1 \leq \rho_\ell \leq \rho_u$, then the number of components is bounded by

$$1 \leqslant c \leqslant \left\lfloor \frac{n}{\rho_\ell + 1} \right\rfloor \tag{4.1}$$

and the number of edges by

$$\max\left(n-c, \left\lceil \frac{\rho_\ell n}{2} \right\rceil\right) \leqslant e(n,c) \leqslant (\rho_\ell + 1)^2 \left(\frac{\nu + 1}{2}\right) + (\rho_u + 1)^2 \left(\frac{c-\nu}{2}\right)$$

$$+ \nu(c-\nu-1)\,(\rho_\ell + 1)(\rho_u + 1) - n\nu(\rho_\ell + 1) - n(c-\nu-1)(\rho_u + 1) + \binom{n}{2} \tag{4.2}$$

where

$$\nu = \left\lceil \frac{c(\rho_u + 1) - n}{\rho_u - \rho_\ell} \right\rceil - 1 \tag{4.3}$$

proof: i) upper bound: let $n_i$ be the size of the $i^{th}$ connected component. The number of edges is then equal to

$$e(n,c) = \sum_{i=1}^{c} \frac{n_i(n_i-1)}{2} = \frac{1}{2}\left(\sum_{i=1}^{c} n_i^2 - n\right) \tag{4.4}$$

We show that $e$ is maximum for the choice

$$\{n_i\} = \underbrace{\{\rho_\ell+1,\dots,\rho_\ell+1,}_{\nu}\ n-\nu(\rho_\ell+1)-(c-\nu-1)(\rho_u+1),\underbrace{\rho_u+1,\dots,\rho_u+1\}}_{c-\nu-1}$$

Let us assume that there exist two classes of size $n_{i_1}$ and $n_{i_2}$ such that

$$\rho_\ell+1 < n_{i_1} \leq n_{i_2} < \rho_u+1$$

These are replaced by two classes of size

$$n'_{i_1} = \max(\rho_\ell+1, n_{i_1}+n_{i_2}-\rho_u-1)$$

and

$$n'_{i_2} = \min(n_{i_1}+n_{i_2}-\rho_\ell-1, \rho_u+1)$$

Their sum remains constant since

$$\begin{aligned} n'_{i_1}+n'_{i_2} &= (\rho_\ell+1)+(n_{i_1}+n_{i_2}-\rho_\ell-1) \quad \text{if} \quad n_{i_1}+n_{i_2}-\rho_\ell-1 \leq \rho_u+1 \\ &= (\rho_u+1)+(n_{i_1}+n_{i_2}-\rho_u-1) \quad \text{if} \quad n_{i_1}+n_{i_2}-\rho_u-1 \geqslant \rho_\ell+1 \end{aligned}$$

The corresponding variation in the number of edges is

$$\begin{aligned} \delta e = \frac{1}{2}(n'^2_{i_1}+n'^2_{i_2}-n^2_{i_1}-n^2_{i_2}) &= \frac{1}{2}\left((\rho+1)^2+\left(n_{i_1}+n_{i_2}-(\rho+1)\right)^2-n^2_{i_1}-n^2_{i_2}\right) \\ &= (\rho+1)^2-(\rho+1)(n_{i_1}+n_{i_2})+n_{i_1}n_{i_2} \end{aligned} \tag{4.5}$$

where

$$\rho = \rho_\ell \qquad \text{if} \qquad n_{i_1} + n_{i_2} \leqslant \rho_u + \rho_\ell + 2$$
$$\rho = \rho_u \qquad \text{otherwise}$$

We see that $\delta e \geqslant 0$ if $\rho + 1 \notin \left] n_{i_1}, n_{i_2} \right[$ , which was our hypothesis.

In the definition of the solution vector $\{n_i\}$ the value of $\nu$ is determined by the inequality

$$\rho_\ell + 1 \leq n - \nu(\rho_\ell + 1) - (c - \nu - 1)(\rho_u + 1) < \rho_u + 1$$

or equivalently

$$\frac{c(\rho_u + 1) - n}{\rho_u - \rho_\ell} - 1 \leq \nu < \frac{c(\rho_u + 1) - n}{\rho_u - \rho_\ell} \tag{4.6}$$

so that

$$\nu = \left\lceil \frac{c(\rho_u + 1) - n}{\rho_u - \rho_\ell} \right\rceil - 1 \tag{4.7}$$

The maximum number of edges is then obtained as

$$e_{\max} = \frac{1}{2}\left[\nu(\rho_\ell + 1)^2 + (c - \nu - 1)(\rho_u + 1)^2 + \Big(n - \nu(\rho_\ell + 1) - (c - \nu - 1)(\rho_u + 1)\Big)^2 - n\right]$$
$$= \binom{n}{2} + (\rho_\ell + 1)^2\binom{\nu + 1}{2} + (\rho_u + 1)^2\binom{c - \nu}{2} - n\Big[\nu(\rho_\ell + 1) + (c - \nu - 1)(\rho_u + 1)\Big] + \nu(c - \nu - 1)(\rho_\ell + 1)(\rho_u + 1)$$

ii) lower bound: let $\rho(a_i)$ be the degree of vertex $a_i$. It is clear that for a connected graph

$$\sum_{i=1}^{n} \rho(a_i) = 2e$$

Consequently for $\rho_\ell > 1$

$$\text{if} \quad \rho_\ell \mid 2 \quad \text{or} \quad n \mid 2 \qquad e_{\min} = \frac{\rho_\ell n}{2} \qquad \text{if } \rho_u \geqslant \rho_\ell$$

$$\text{if} \quad \rho_\ell \nmid 2 \quad \text{and} \quad n \nmid 2 \qquad e_{\min} = \frac{\rho_\ell (n-1)}{2} + \frac{\rho_\ell + 1}{2} = \frac{\rho_\ell n + 1}{2} \qquad \text{if } \rho_u \geqslant \rho_\ell + 1$$

therefore

$$e_{\min} = \left\lceil \frac{\rho_\ell n}{2} \right\rceil$$

Two cases have not been considered yet, namely, $\rho_\ell = 0$ and $\rho_\ell = 1$.

But

$$e = 0 \qquad \text{for } n = 1 \qquad \text{if} \qquad 0 \leqslant \rho_\ell \leqslant \rho_u \leqslant 1$$

and

$$e = n-1 \quad \text{for any } n \qquad \text{if} \qquad \rho_u > 1$$

so that

$$e_{\min}(n, 1) = \max\left(n-1, \left\lceil \frac{\rho_\ell n}{2} \right\rceil\right) \qquad \rho_u > 1 \qquad (4.8)$$

since the only cases of interest require $\rho_u > 1$ for edges to exist.

For $c$ classes, we have therefore

$$e(n, c) \geqslant \sum_{i=1}^{c} \max\left(n_i - 1, \left\lceil \frac{\rho_\ell n_i}{2} \right\rceil\right)$$

$$\geqslant \max\left(n-c, \left\lceil \frac{\rho_\ell n}{2} \right\rceil\right) \qquad (4.9)$$

<u>Corollary</u> - If $\rho_\ell = \rho_u > 1$, a graph $G$ exists if $n\rho_\ell \mid 2$ and then

$$1 \leqslant c \leqslant \frac{n\rho_\ell}{2} \qquad (4.10)$$

$$e = \frac{n\rho_\ell}{2} \qquad (4.11)$$

proof:

$$\sum_{i=1}^{c} \sum_{j=1}^{n_i} \rho_\ell = 2e$$

$$\frac{1}{2} \sum_{i=1}^{c} n_i \rho_\ell = e$$

For each class we require $n_i\rho_\ell \mid 2$ which implies $n_i\rho_\ell \geqslant 2$ so that

$$n\rho_\ell \geqslant 2c$$

<u>Theorem</u> - The number of classes is bounded by

$$\max\left(1, \sum_i \delta_{1, r_i}\right) \leqslant c \leqslant \left\lfloor \sum_i \frac{1}{r_i} \right\rfloor \tag{4.12}$$

where

$$r_i = 1 + \sum_j a_{ij}$$

proof: the lower bound is obvious; whenever a row sum $r_i$ is equal to 1, the corresponding vertex is a one element class; for the upper bound, we assign to each vertex a weight equal to the inverse of its local degree. If vertex $v_i$ belongs to a class of size $s_i$ then

$$\frac{1}{s_i} \leqslant \frac{1}{r_i} \leqslant 1$$

so that

$$c = \sum_i \frac{1}{s_i} \leqslant \left\lfloor \sum_i \frac{1}{r_i} \right\rfloor \leqslant \sum_i \frac{1}{r_i} \tag{4.13}$$

## 4.2. Calculation of Bounds for the Expected Number of Classes

Since all row sums distributions are invariant under any permutation of the $n$ vertices of $G_n$, we may assume, without loss of generality that

$$k_1 \leqslant k_2 \leqslant \cdots \leqslant k_n$$

and similarly for all the other row sums sequences. Let

$$F(k_1, k_2, \ldots, k_n) = P\left\{ r_{i_1} \leqslant k_1, \ldots, r_{i_n} \leqslant k_n \right\}$$

be the joint distribution of the row sums $r_{i_1} \leqslant r_{i_2} \leqslant \cdots \leqslant r_{i_n}$ ; let

$$f(k_1, k_2, \ldots, k_n) = P\left\{ r_{i_1} = k_1, \ldots, r_{i_n} = k_n \right\}$$

be their joint density. If

$$R = \sum_i \frac{1}{r_i} \; ,$$

we have

$$P\left\{ \rho \leqslant R \right\} = 1 - \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} \cdots \sum_{k_n=k_{n-1}}^{r_n} f(k_1, k_2, \ldots, k_n) \qquad (4.14)$$

We now determine bounds for the expected value of $c$ by calculating the expected values of both sides of inequality (4.12) as follows

$$\mathcal{E}(\lfloor \rho \rfloor) = \sum_{k_1=1}^{n} \sum_{k_2=k_1}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} \left\lfloor \frac{1}{k_1} + \frac{1}{k_2} + \cdots + \frac{1}{k_n} \right\rfloor f(k_1, k_2, \ldots, k_n) \qquad (4.15)$$

$$\mathcal{E}\left\{\max\left(1,\ \sum_i \delta_{i,r_i}\right)\right\} = \sum_{k_1=2}^{n} \sum_{k_2=k_1}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(k_1, k_2, \ldots, k_n)$$

$$+ \sum_{k_2=2}^{n} \sum_{k_3=k_2}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1,\ k_2, \ldots, k_n)$$

$$+ \ldots$$

$$+ (n-1) \sum_{k_n=2}^{n} f(1, 1, \ldots, 1, k_n)$$

$$+ n\, f(1, 1, \ldots, 1) \qquad (4.16)$$

This last expression can be transformed using

$$\sum_{k_1=2}^{n} \sum_{k_2=k_1}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(k_1, k_2, \ldots, k_n)$$

$$= 1 - \sum_{k_2=1}^{n} \sum_{k_3=k_2}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, k_2, \ldots, k_n)$$

to give

$$\left\{ \max\left(1, \sum_i \delta_{1,r_i}\right)\right\}$$

$$= 1 - \sum_{k_2=1}^{n} \sum_{k_3=k_2}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, k_2, \ldots, k_n)$$

$$+ \sum_{k_2=2}^{n} \sum_{k_3=k_2}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, k_2, \ldots, k_n) + \ldots + n f(1, 1, \ldots, 1)$$

$$= 1 - \sum_{k_3=1}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, 1, k_3, \ldots, k_n)$$

$$+ 2 \sum_{k_3=2}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, 1, k_3, \ldots, k_n) + \ldots + n f(1, 1, \ldots, 1)$$

$$= 1 + \sum_{k_3=2}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, 1, k_3, \ldots, k_n)$$

$$- \sum_{k_4=1}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, 1, 1, k_4, \ldots, k_n) + \ldots + n f(1, 1, \ldots, 1)$$

$$= 1 + \sum_{k_3=2}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, 1, k_3, \ldots, k_n) + \ldots$$

$$+ (n-2) \sum_{k_n=2}^{n} f(1, 1, \ldots, 1, k_n) + (n-1) f(1, 1, \ldots, 1) \qquad (4, 17)$$

so that if we let

$$B_1(n) = 1 + \sum_{k_3=2}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} f(1, 1, k_3, \ldots, k_n) + \ldots + (n-1)f(1, 1, \ldots, 1) \tag{4.18}$$

$$B_2(n) = \sum_{k_1=1}^{n} \sum_{k_2=k_1}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} \left\lfloor \sum_{i} \frac{1}{k_i} \right\rfloor f(k_1, k_2, \ldots, k_n) \tag{4.19}$$

we have bounded the expected value of the number of classes by

$$1 \leq B_1(n) \leq \mathcal{E}(c_n) \leq B_2(n) \leq n \tag{4.20}$$

Still, the upper bound $B_2(n)$ can be refined and replaced by $B_2'(n)$ in the form

$$B_2'(n) = \sum_{k_1=1}^{n} \sum_{k_2=k_1}^{n} \cdots \sum_{k_n=k_{n-1}}^{n} \left\lfloor 1 + \sum_{i=1}^{n-k_n} \frac{1}{k_i} \right\rfloor f(k_1, k_2, \ldots, k_n) \tag{4.21}$$

To justify this step, consider a particular set of row sums $\{k_1, k_2, \ldots, k_n\}$ ; the largest class certainly contains no fewer than $k_n$ members so that, at least $k_n$ among the $k_i$'s must be greater than or equal to $k_n$ ; therefore we may replace $\sum_i \frac{1}{k_i}$ by

$$\sum_{i=1}^{n-k_n} \frac{1}{k_i} + \sum_{i=n-k_{n+1}}^{n} \frac{1}{k_n} = 1 + \sum_{i=1}^{n-k_n} \frac{1}{k_i}$$

and still preserve the inequality on the right. Now, if $k_n < n$ , the same reasoning may be applied again in order to decrease $B_2'(n)$ .

Thus, let us define the transformation $\mathcal{C}$ by

$$\mathcal{C}\left(a+\sum_{i=1}^{\nu}\frac{1}{k_i}\right) = a+1+\sum_{i=1}^{\nu'}\frac{1}{k_i} \tag{4.22}$$

if $\exists\, \nu' \ni \nu' = \left\{\max_{1\leq j\leq \nu} j \;\middle|\; k_j \leq j \leq \nu - k_\nu\right\}$

otherwise

$$\mathcal{C}\left(a+\sum_{i=1}^{\nu}\frac{1}{k_i}\right) = a+1 \tag{4.23}$$

<u>Theorem</u> $1 \leq B_1(n) \leq \mathcal{E}(c_n) \leq B_2''(n) \leq n$

where $B_1(n)$ is given by (4.18) and

$$B_2''(n) = \sum_{k_1=1}^{n}\sum_{k_2=k_1}^{n}\cdots\sum_{k_n=k_{n-1}}^{n}\mathcal{C}^n\left(\sum_{i=1}^{n}\frac{1}{k_i}\right)f(k_1,k_2,\ldots,k_n) \tag{4.24}$$

proof: designate by $s_1 \leq s_2 \leq \ldots \leq s_c$ the size of each of the c classes; we then estimate the sequence

$$s_{i_1} \leq s_{i_2} \leq \cdots \leq s_{i_n}$$

by

$$r_1 \leq r_2 \leq \cdots \leq r_n$$

But for these values, $r_\nu \leq s_{i_\nu}$, $\nu = 1,2,\ldots,n$ as we can show by induction on $\nu$. Clearly $r_n \leq s_{i_n}$ ; assume now $r_j \leq s_{i_j}$, $j=\nu, \nu+1,\ldots,n$, then

i) if $s_{i_{\nu-1}} = s_{i_\nu} \Rightarrow r_{\nu-1} \leq r_\nu \leq s_{i_{\nu-1}}$

ii) if $s_{i_{\nu-1}} < s_{i_\nu}$ $\exists$ only $n-\nu+1$ $s'_s \ni s > s_{i_{\nu-1}}$

but if $r_{\nu-1} > s_{i_{\nu-1}} \Rightarrow \exists$ $n-\nu+2$ $r'_s \ni r \geq r_{\nu-1} > s_{i_{\nu-1}}$

which is a contradiction. Consequently $r_{\nu-1} \leq s_{i_{\nu-1}}$.

Let us now examine the application of transformation $\mathcal{C}$ to the initial class estimate

$$\tilde{c} = \sum_{\nu=1}^{n} \frac{1}{r_\nu} \geq \sum_{\nu=1}^{n} \frac{1}{s_{i_\nu}} \qquad (4.25)$$

The largest class is of size $s_{i_n} \geq r_n$. We may replace $r_n$ of the $r'_s$ by $r_n$ and still preserve the above inequality provided we select the $r_n$ largest $r'_s$ in the sequence. Furthermore, we can even be more shrewd when determining the size of the largest class in the subsequence. We remark that

$$r_\nu > \nu \implies s_{i_\nu} = s_{i_{\nu+1}} = \ldots = s_{i_{r_\nu}}$$

since $s_{i_\nu} \geq r_\nu > \nu$ excludes the possibility that $s_i$ be the start of a new class $s_{i_\nu}, s_{i_{\nu-1}}, \ldots$; therefore $s_{i_\nu} = s_{i_{\nu+1}}$ and the same reasoning may be applied to $s_{i_{\nu+1}} > \nu$ until we reach $s_{i_{r_\nu}} \geq r_\nu$. Consequently, the estimate of the size of the largest remaining class is determined by the smallest integer $j \geq 0$ such that

$$r_{n-r_n-j} \leq n-r_n-j$$

which is equivalent to saying that the largest class must be of size at least $r_n+j$. The same argument is valid for $\sum_{i=1}^{r_{n-r_n-j}} \frac{1}{r_i}$ and so forth until $r_1$ has been assigned to a class.

## 4.3. Local Degrees Distribution in Subgraphs of $G_n$

Let $v_n(d)$ be the number of vertices of degree $d$ in $G_n$. Let $\{G_k\}$ be the family of subgraphs of $G_n$ with $k$ vertices. Our experiment is the selection of a vertex from one member of $\{G_k\}$ with outcome $d \in \{0, 1, \dots, k-1\}$. Let $p_k(d)$ be the probability that the vertex has degree $d$, it is related to $p_n(\nu)$ by

$$p_k(d) = \binom{k-1}{d} \sum_{\nu=d}^{n-1} \frac{\binom{n-k}{\nu-d}}{\binom{n-1}{\nu}} p_n(\nu) \tag{4.26}$$

*result which can also be expressed in an equivalent form*

$$p_k(d) = \binom{k-1}{d} \sum_{\nu=d}^{n-1} \frac{(n-k)!\,\nu!\,(n-\nu-1)!}{(\nu-d)!\,(n-k-\nu+d)!\,(n-1)!} p_n(\nu)$$

$$= \binom{k-1}{d} \sum_{\nu=d}^{n-1} \frac{\binom{\nu}{d}\binom{n-\nu-1}{k-d-1}}{\binom{n-1}{k-1}\binom{k-1}{d}} p_n(\nu)$$

$$= \frac{1}{\binom{n-1}{k-1}} \sum_{\nu=d}^{n-1} \binom{\nu}{d}\binom{n-\nu-1}{k-d-1} p_n(\nu) \tag{4.27}$$

## 4.4. Examples for Particular Distributions $p_n(d)$

We look formally at some distributions $p_n(d)$ without worrying about the realizability of $v_n(d)$ as a graph.

1) uniform distribution

$$p_n(\nu) = \frac{1}{\nu_2 - \nu_1 + 1} \qquad \nu_1 \leq \nu \leq \nu_2 \quad , \qquad \text{otherwise } 0$$

$$p_k(d) = \frac{1}{\binom{n-1}{k-1}(\nu_2 - \nu_1 + 1)} \sum_{\nu=\max(d,\nu_1)}^{\min(n-1,\nu_2)} \binom{\nu}{d}\binom{n-\nu-1}{k-d-1}$$

when $\nu_2 = n-1$ and $\nu_1 \leq d$

$$p_k(d) = \frac{\binom{n}{k}}{\binom{n-1}{k-1}(n-\nu_1)} = \frac{n}{k(n-\nu_1)}$$

which is also uniform.

2) binomial distribution with parameter $\alpha$

$$p_n(\nu) = \binom{n-1}{\nu} \alpha^{\nu} (1-\alpha)^{n-\nu-1}$$

$$p_k(d) = \binom{k-1}{d} \sum_{\nu=d}^{n-1} \binom{n-k}{\nu-d} \alpha^{\nu} (1-\alpha)^{n-\nu-1}$$

$$= \binom{k-1}{d} \sum_{\nu=0}^{n-d-1} \binom{n-k}{\nu} \alpha^{\nu+d} (1-\alpha)^{n-\nu-d-1}$$

since $d \leq k-1$ the summation can be changed to

$$p_k(d) = \binom{k-1}{d} \sum_{\nu=0}^{n-k} \binom{n-k}{\nu} \alpha^{\nu}(1-\alpha)^{n-k-\nu} \alpha^{d}(1-\alpha)^{k-d-1}$$

$$= \binom{k-1}{d} \alpha^{d} (1-\alpha)^{k-d-1}$$

which is also binomial with the same parameter.

Since the binomial distribution acts as the kernel of the transformation, we define the ratios

$$R_n(d) = \frac{p_n(d)}{\binom{n-1}{d}} \qquad (4.28)$$

which, by formula (4.26), satisfy the recurrence relations

$$R_k(d) = \sum_{\nu=d}^{n-1} \binom{n-k}{\nu-d} R_n(\nu) = \sum_{\nu=0}^{n-k} \binom{n-k}{\nu} R_n(d+\nu) \qquad (4.29)$$

or alternatively

$$R_k(d) = R_{k+1}(d) + R_{k+1}(d+1) \qquad (4.30)$$

Computation of the $R_k$ given $R_n$ is of course a simple matter using a triangular tableau

$$\begin{tikzcd}
R_n(0) \arrow[d] & R_n(1) \arrow[dl] \arrow[d] & R_n(2) \arrow[dl] & \cdots & R_n(n-2) \arrow[d] & R_n(n-1) \arrow[dl] \\
R_{n-1}(0) \arrow[d] & R_{n-1}(1) \arrow[dl] \arrow[d] & R_{n-1}(2) \arrow[dl] & \cdots & R_{n-1}(n-2) \arrow[ddllll, dash] & \\
R_{n-2}(0) & R_{n-2}(1) & R_{n-2}(2) & \cdots & & \\
R_1(0) & & & & &
\end{tikzcd}$$

Example - for a square $\ell \times \ell$ board, two elementary squares being connected if they share an edge, there are

$(\ell-2)^2$ points with degree 4

$4(\ell-2)$ points with degree 3

4 points with degree 2

so that a tableau R can be built starting from

$$p_{n^2}(4) = \left(1 - \frac{2}{\ell}\right)^2$$

$$p_{n^2}(3) = \frac{4(\ell-1)}{\ell^2}$$

$$p_{n^2}(2) = \frac{4}{\ell^2}$$

$$p_{n^2}(i) = 0 \qquad i \notin \{2, 3, 4\}$$

It is shown in table (4.1).

Table 4.1

Local Degrees Distribution for a 4x4 Board

| | Value | | | | |
|---|---|---|---|---|---|
| 1 | 1.000000 | 0. | 0. | 0. | 0. |
| 2 | 0.800000 | 0.200000 | 0. | 0. | 0. |
| 3 | 0.630952 | 0.338095 | 0.030952 | 0. | 0. |
| 4 | 0.489560 | 0.424176 | 0.082967 | 0.003297 | 0. |
| 5 | 0.372711 | 0.467399 | 0.147253 | 0.012454 | 0.000183 |
| 6 | 0.277473 | 0.476190 | 0.216117 | 0.029304 | 0.000916 |
| 7 | 0.201099 | 0.458242 | 0.282967 | 0.054945 | 0.002747 |
| 8 | 0.141026 | 0.420513 | 0.342308 | 0.089744 | 0.006410 |
| 9 | 0.094872 | 0.369231 | 0.389744 | 0.133333 | 0.012821 |
| 10 | 0.060440 | 0.309890 | 0.421978 | 0.184615 | 0.023077 |
| 11 | 0.035714 | 0.247253 | 0.436813 | 0.241758 | 0.038462 |
| 12 | 0.018864 | 0.185348 | 0.433150 | 0.302198 | 0.060440 |
| 13 | 0.008242 | 0.127473 | 0.410989 | 0.362637 | 0.090659 |
| 14 | 0.002381 | 0.076190 | 0.371429 | 0.419048 | 0.130952 |
| 15 | 0. | 0.033333 | 0.316667 | 0.466667 | 0.183333 |
| 16 | 0. | 0. | 0.250000 | 0.500000 | 0.250000 |

## 4.5. Canonical Graph Representation

The study of properties of graphs numerically often requires manipulating large families of such graphs and recognizing isomorphic graphs as being several instances of the same member. For undirected graphs $G_n$ with no loops and single edges, each member is uniquely represented by its symmetric incidence matrix.

For classification purposes, one wishes to find an integer function which maps the family G into the positive integers. The mapping function should possess some essential properties:

i) uniqueness of the image of a graph and its isomorphs

ii) computational simplicity of the mapping function

iii) computational feasibility of the inverse mapping

A classical function which appears repetitively in the literature is the permanent. For a symmetric incidence matrix A it is defined as:

$$\text{per}\,(A) = \sum_{\mathcal{J}} a_{1\,i_1}\, a_{2\,i_2} \cdots a_{n\,i_n} \tag{4.31}$$

The summation in (4.31) extends over the set $\mathcal{J}$ of all permutations of the integers $\{1, 2, \ldots, n\}$. Clearly, the permanent has property (i) since it remains invariant under any permutation of the rows and columns. Property (ii) has to be examined more carefully. Instead of generating the set $\mathcal{J}$, it is more efficient to apply the principle of inclusion and exclusion to the permanent calculation and obtain the following theorem (Ryser).

Theorem - Let $A_n$ be a square nxn matrix and $A_r$ be an nxr matrix obtained from $A_n$ by selecting $r$ of its columns. Let $S(A_r)$ be the product of the row sums of $A_r$. Then

$$\text{per}\ (A) = \sum_{i=0}^{n-1} (-1)^i \sum_{\{A_{n-i}\}} S(A_{n-i}) \qquad (4.32)$$

We estimate the number of operations required to compute the permanent by

$$\omega = \sum_{i=0}^{n-1} \binom{n}{i} \left[ n(n-i-1) + k \right]$$

where addition and multiplication are considered as identical and $k$ takes into account the number of operations required to produce the next combination. An approximate value of $\omega$ is

$$\omega \simeq 2^{n-1}(n^2+k) \simeq n^2 2^{n-1} \quad \text{for} \quad n >> 1$$

Unfortunately property iii) does not hold. Given the permanent of a symmetric nxn matrix, we cannot recover the original matrix short of an endless trial and error approach.

Therefore, motivation exists to seek a different function of the incidence matrix having the three properties stated above, and offering practical advantages for numerical computation.

## 4.6. Binary Invariants

Definition - An incidence matrix is said to be monotonic if its sequences of row and column sums are monotonic non-increasing sequences.

Let $R = \left\{ r_1, r_2, \ldots, r_n \right\}$ be the sequence of row sums of the symmetric incidence matrix $A$ in monotonic form. The r's assume only $d$ distinct values and $\rho_k$ is the number of row sums having the common value $k$.

Definition - The binary sum of a symmetric matrix $A$ with zero diagonal elements is

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} \, 2^{\frac{(n - i - 1)(n - i)}{2} + n-j} \qquad (4.33)$$

Definition - The binary invariant of $A$ is the minimum of the binary sum of $A$ taken over all monotonic isomorphs of $A$.

$$\beta = \min_{\left\{ \hat{A} \right\}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2^{\frac{(n-i+1)(n-i)}{2} + n-j} \, \hat{a}_{ij} \qquad (4.34)$$

The set $\left\{ \hat{A} \right\}$ is formed by first making $A$ monotonic, then applying to the rows and columns of the resulting matrix all of the permutations which permutes rows with identical row sums among themselves and similarly for columns. These permutations form a proper subgroup of the symmetric group on $n$ objects if $d > 1$. This subgroup is the product of the symmetric groups on $\rho_o, \rho_1, \rho_2, \ldots, \rho_{n-1}$ objects; it is of degree $n$ and order $\rho_o! \rho_1! \rho_2! \ldots \rho_{n-1}!$.

Notice that the minimum taken over all isomorphs of $A$ is not used since it would require finding the minimum of a set with $n!$ elements in every case. By requiring the incidence matrices to be monotonic, the evaluation of $\beta$ is easier the larger $d$ is.

The number of operations involved is approximately

$$\frac{n^2}{2} \prod_{i=0}^{n-1} \rho_i!$$

Clearly, this number may become inordinately large when some of the $\rho_i$'s are large and the permanent then turns out to be more readily computable. Both the permanent and a binary sum are required to allow classification and recovery of the matrix $A$.

Using binary invariants gives us a simple and powerful way of numerically handling problems which involve undirected graphs with single edges and no loops. These graphs being the only ones that we are concerned with in class counting, we have not extended the definition to encompass more general types of graphs since that would entail losing some computational efficiency. However, extensions are straightforward; for instance, directed graphs with single edges and single slings would simply be encoded using the binary representation of the full matrix.

## 4.7. Finding the Set of Non-Isomorphic Graphs with Prescribed Local Degrees

We now show that once a member of the set is known, the other elements can be easily obtained, using a method similar to Ryser's [27].

Let $\mathcal{A}(\vec{\rho})$ be the set of all nxn incidence matrices which are symmetric, monotonic and have $\rho_i$ row sums equal to i (as usual the diagonal elements are zero).

Consider some matrix $A \in \mathcal{A}(\rho)$. We define an interchange matrix for A to be a 2x2 submatrix of A of the form

$$\alpha_o = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ or } \alpha_1 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

where the elements of $\alpha_o$ and $\alpha_1$ are restricted to be off diagonal elements of A.

The replacement of $\alpha_i$ by $\alpha_{1-i}$ (accompanied, of course, by the interchange of their conjugates to preserve the symmetry of A) leaves $\vec{\rho}$ unchanged so that $\mathcal{A}(\vec{\rho})$ is closed under an arbitrary sequence of elementary interchanges.

Theorem - The graph induced by the interchange operation on the elements of $\mathcal{A}(\vec{\rho})$ is strongly connected.

proof: let $A_1, A_2 \in \mathcal{A}(\vec{\rho})$ and $\hat{A}_1 \neq \hat{A}_2$. We are going to transform independently the incidence matrices $\hat{A}_1$ and $\hat{A}_2$ so that their first rows become

$$a_{12} = a_{13} = \ldots = a_{1\,r_1+1} = 1$$

$$a_{1\,r_1+2} = \ldots = a_{1n} = 0$$

Take $\hat{A}_1$ for instance. Let the first zero element in the first row be $a_{1j_1}$ with $a_{1j_2} = 1$ for $j_2 > j_1$, otherwise $\hat{A}_1$ is already in the desired form. We seek now to exchange $a_{1j_1}$ with $a_{1j_2} = 1$, $j_2 > j_1$. At the same time we must find a row $i > 1$ for which $a_{ij_1} = 1$ and $a_{ij_2} = 0$, $i \neq j_2$. Such a row always exists. If it did not, then column $j_2$ would have as many 1's as column $j_1$ plus an extra one in the first row implying

$$r_{j_2} \geqslant r_{j_1} + 1$$

but this is certainly false from the assumption of monotonicity. After this transformation has been applied repetitively to $\hat{A}_1$ and $\hat{A}_2$, both these matrices have identical first row and column which can now be removed, thereby modifying their common degree vector. The resulting matrices are put in monotonic form, their first row contains the same number of 1's and another sequence of interchanges is performed on each one. This process terminates when the resulting matrices become scalars.

Therefore, we have proved that it is possible to go from any element $A_1$ to any other $A_2$ by a sequence of interchanges, or equivalently that the graph with the elements of $\mathcal{A}(\vec{\rho})$ as vertices, is strongly connected.

Example - Let

$$\vec{\rho} = \begin{pmatrix} 0 \\ 3 \\ 2 \\ 1 \\ 1 \end{pmatrix}$$

a starting matrix is

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \qquad \beta(A) = 1008640$$

By successive interchanges we find seven non-isomorphic graphs

| | binary invariant | aliases | | |
|---|---|---|---|---|
| 1 | 913408(4) | 976896(4) | 1008640(4) | |
| 2 | 1303040(12) | | | |
| 3 | 1430016(2) | 1493504(2)<br>1747456(2) | 1525248(2)<br>1779200(2) | 1683968(2) |
| 4 | 1433856(2) | 1499264(2)<br>1761296(2) | 1531968(2)<br>1794056(2) | 1695776(2) |
| 5 | 1874432(4) | 1906176(4) | 1969664(4) | |
| 6 | 1878272(1) | 1880192(1)<br>1910016(1)<br>1921032(1)<br>1983504(1) | 1886240(1)<br>1912896(1)<br>1975424(1)<br>1984520(1) | 1888272(1)<br>1917984(1)<br>1976384(1) |
| 7 | 1892356(4) | 1925122(4) | 1990657(4) | |

Let $\alpha_i$ be the number of distinct monotonic matrices in class i. The probability to find a matrix of class i when sampling uniformly over the set of all matrices whose monotonic transforms belong to $\mathcal{A}(\overline{\rho})$ is proportional to the number of aliases plus 1 in class i ; indeed the number of distinct A's is

$$\frac{n!}{\prod_{i=0}^{n-1} \rho_i!} \alpha_i$$

so that

$$p(A \in C_i) = \frac{\alpha_i}{\sum_j \alpha_j}$$

However, sampling by successive interchanges does not yield the same result since there exists a non-zero correlation between successive samples (the correlation vanishes if $p(A_i | A_{i-1}) = p(A_i)$ which, in general, is not the case since there may not even exist a single interchange which transforms a prescribed $A$ into another one).

| | | observed frequency | $p(A \in C_i)$ |
|---|---|---|---|
| 1 | 913408 | 12.5 | 8.8 |
| 2 | 1303040 | 3.1 | 2.9 |
| 3 | 1430016 | 15.9 | 17.6 |
| 4 | 1433856 | 17.2 | 17.6 |
| 5 | 1874432 | 7.4 | 8.8 |
| 6 | 1878272 | 31.6 | 35.3 |
| 7 | 1892356 | 12.3 | 8.8 |

This is illustrated by exhibiting the class transition matrices

$$P_{ij}(\ell) = P\left(A_{\nu+\ell} \in C_j \middle| A_\nu \in C_i\right)$$

$$P(1) = \begin{pmatrix} 15.2 & 0. & 23.2 & 15.2 & 0. & 35.2 & 11.2 \\ 0. & 0. & 32.3 & 67.8 & 0. & 0. & 0. \\ 12. & 10.7 & 26.4 & 8.2 & 17.6 & 24.6 & 0. \\ 11.1 & 8.1 & 9.9 & 22.7 & 0. & 21.6 & 26.8 \\ 0. & 0. & 44.6 & 0. & 10.8 & 44.7 & 0. \\ 11.4 & 0. & 8.9 & 14.6 & 12. & 38. & 15.2 \\ 25.2 & 0. & 0. & 27.7 & 0. & 35. & 12.2 \end{pmatrix}$$

whereas

$$P(5) = \begin{pmatrix} 14.5 & 3.2 & 16.1 & 10.5 & 6.4 & 36.2 & 12.9 \\ 16.2 & 3.2 & 3.2 & 13. & 13.0 & 32.4 & 19.5 \\ 10.7 & 1.3 & 15.2 & 18.3 & 8.8 & 33.5 & 12.0 \\ 8.8 & 2.3 & 11.7 & 18.7 & 6.4 & 39.7 & 12.3 \\ 13.6 & 2.7 & 15.0 & 16.3 & 10.9 & 25.8 & 16.3 \\ 13.4 & 4.5 & 17.5 & 17.5 & 7.0 & 27.7 & 12.4 \\ 13.9 & 3.3 & 20.4 & 22.1 & 5.7 & 27.0 & 8.2 \end{pmatrix}$$

## 4.8. Finding an Initial Graph

The proof of the last theorem is constructive in the sense that it leads to an algorithm for deciding whether or not $\mathcal{A}(\vec{\rho}) = \emptyset$. In the latter case, the algorithm constructs one feasible incidence matrix.

Initially we are given some row sum vector R to which corresponds a certain $\vec{\rho}$, the incidence matrix having all zero entries (in the following we only refer to the upper half of that matrix).

i) first, check that the sum of the row sums is even. If this test fails, clearly no solution graph exists.

ii) then, permute the incidence matrix into a monotonic form (i.e., $r_1 \geqslant r_2 \geqslant \ldots \geqslant r_n$) and process each row at a time:

> if the row sum is null go the next row;
>
> else, let $d_{max}$ be the maximum row sum;
>
> set that particular row sum equal to 0 and decrease by 1 the next $d_{max}$ rows with highest possible row sums;
>
> if this forces some row sum to become negative no solution graph exists; at the same time we set $a\big(\min(i,j), \max(i,j)\big)$ where i is the row picked as $d_{max}$ and j corresponds to each row sum which has been decreased by 1 ;
>
> After all rows have been processed without early termination, the upper half of a is the incidence matrix.

In terms of graph transformations, the foregoing construction simply says to select one of the vertices with maximum degree; call the vertex $a_{i_1}$, its degree being $d_{i_1}$; connect $d_{i_1}$ other vertices to $a_{i_1}$ selecting them in order of decreasing degrees. Then repeat the construction on the graph $G'$ obtained from G by removing $a_{i_1}$ and its $d_{i_1}$ adjacent edges.

## 4.9. Class Estimation in Terms of Transitional Probabilities

We shall now examine the behavior of the expected number of equivalence classes when random samples of increasing size are selected from a population $\mathscr{P}$ of N individuals. If we perform an experiment to sample $\mathscr{P}$ without replacement, the number of classes in the sample of $n+1$ elements is conditioned by the $n$ elements chosen earlier, regardless of their order; still, there exist $\binom{N}{n}$ sets of transition probabilities which express the probability that the particular sample $S_{n+1}$ will have $c_{n+1}$ classes if $S_n$ had $c_n$ classes itself. However, if we think of the experiment as producing the number of classes in a sample of size $n$, averaged over the space of all subsets of $\mathscr{P}$ containing $n$ elements, there exists a unique set of transition probabilities from $c_n$ to $c_{n+1}$. The discrete random variables $c_n$ take the values $1, 2, \ldots, n$ and

$$P\left\{c_n = i_n \,\middle|\, c_{n-1} = i_{n-1}, \ldots, c_1 = i_1\right\} = P\left\{c_n = i_n \,\middle|\, c_{n-1} = i_{n-1}\right\}$$

so that the sequence $c_n$ forms an inhomogeneous Markov chain.

We introduce the density

$$p_i(n) = P\left\{c_n = i\right\}$$

and the transition probabilities

$$P_{ij}(n, \ell) = P\left\{c_n = i \,\middle|\, c_\ell = j\right\} \qquad n > \ell$$

which satisfy

$$p_i(n) = \sum_{j=1}^{\ell} P_{ij}(n, \ell)\, p_j(\ell)$$

$$\sum_{i=1}^{n} p_i(n) = 1$$

$$\sum_{i=1}^{n} P_{ij}(n, \ell) = 1$$

together with the discrete Chapman-Kolmogorov equation

$$P_{ij}(n, \ell) = \sum_{k=1}^{r} P_{ik}(n, r)\, P_{kj}(r, \ell) \qquad n > r > \ell$$

The expected number of classes is

$$\mathcal{E}(c_n) = \sum_{i=1}^{n} i\, p_i(n) \tag{4.35}$$

and the expected value of its forward difference

$$\Delta c_n = c_{n+1} - c_n$$

is

$$\mathcal{E}(\Delta c_n) = \sum_{i=1}^{n+1} i\left[p_i(n+1) - p_i(n)\right]$$

$$= \sum_{i=1}^{n+1} i\left[\sum_{j=1}^{n} P_{ij}(n+1,n)\,p_j(n) - p_i(n)\right]$$

$$= \sum_{j=1}^{n} p_j(n)\left(\sum_{i=1}^{n+1} i\,P_{ij}(n+1,n) - j\right)$$

$$= \sum_{j=1}^{n} p_j(n)\sum_{i=1}^{j+1}(i-j)\,P_{ij}(n+1,n) \tag{4.36}$$

formula which is immediately clear.

Theorem

$$\delta_{\nu 0} \leq \sum_{j=1}^{\nu} p_j(\nu)\;\; P_{j+1\,j}\,(\nu+1,\nu) \leq \frac{(n-\nu)(n-\nu-1)}{n(n-1)} \tag{4.37}$$

proof: let $G_n$ have $v_d$ vertices of degree $d$ , $d = 1,2,\ldots,n-1$ . For a sample of $\nu$ vertices, $0 \leq \nu \leq n$ , we express in two ways the probability that the $(\nu+1)$st selected element increases the number of classes by 1

$$E_+ = \sum_{j=1}^{\nu} p_j(\nu)\,P_{j+1\,j}(\nu+1,\nu) = \frac{1}{n}\sum_{d=1}^{n-\nu-1} v_d\,\frac{\binom{n-d-1}{\nu}}{\binom{n-1}{\nu}} \tag{4.38}$$

which reveals the influence of the sample size if we transform to

$$E_+ = \frac{1}{n} \sum_{d=1}^{n-\nu-1} v_d \frac{\binom{n-\nu-1}{d}}{\binom{n-1}{d}} \qquad (4.39)$$

$E_+$ is thus a monotone decreasing function of $\nu$ and bounds for $E_+$ can be derived by examining two limiting cases: $G_{n1}$ an n-clique and $G_{n2}$ a linear chain of $n$ vertices. We look for the maximum and minimum of $E_+$ over the space of all connected graphs on $n$ vertices. The coefficient of $v_d$, as seen from the first expression (4.38) of $E_+$, is a monotone decreasing function of $d$. The minimum of $E_+$ is obviously achieved for $\vec{V}_1 = (0, 0, \ldots, n)$ where $\vec{V} = (v_1, v_2, \ldots, v_{n-1})$. That graph is an n-clique for which $E_+ = \delta_{0\nu}$. Similarly, the maximum of $E_+$ must correspond to a vector of degrees $\vec{V}_2 = (2, n-2, 0, \ldots, 0)$ since the connection constraint of the graph precludes using such vectors as $(n, 0, \ldots, 0)$. The following argument shows why $\vec{V}_2$ corresponds to the maximum. The maximum must occur for some tree with $n$ vertices so that

$$\sum_{d=1}^{n-1} d\, v_d = 2(n-1)$$

Pick any vertex; there is at least one edge attached to it by connectivity. Select each incident edge in turn; if it is incident to a vertex of degree higher than 1, detach it from that vertex and attach it again to any vertex of lowest possible degree. This procedure can be repeated indefinitely but once we reach a linear chain, the transformation is equivalent to a rearrangement of the vertices of the chain. Replacing now in (4.39) $v_d$ by $v_1=2$, $v_2=n-2$ we obtain

$$\delta_{\nu 0} \leqslant E_+(n, \nu) \leqslant \frac{(n-\nu)(n-\nu-1)}{n(n-1)} \tag{4.40}$$

Case $C_N = 1$

Conjecture - The expected number of components $\xi(c_n)$ for an $n$ subgraph obtained by uniformly sampling without replacement the vertex set of a connected graph $G_N$ is such that

$$\xi(c_{n-1}) \geqslant \xi(c_n) \Rightarrow \xi(c_n) \geqslant \xi(c_{n+1}) \tag{4.41}$$

Since $c_1 = c_N = 1$, the behavior of $\xi(c_n)$ is then well determined; it increases monotonically starting at $c_1 = 1$ and as soon as it starts to decrease, it must continue to do so until it reaches 1. Notice that $\xi(c_n)$ is in general not convex. Although we have shown earlier that the birth process for components is monotonically decreasing with $n$, there still remains the possibility that the death process decreases suddenly so that the overall result is an increase in the expected number of components.

A partial analysis can be carried out as follows. Assume

$$\xi(c_{n-1}) > \xi(c_n)$$

and consider the result of adding two vertices $x$ and $y$ to a configuration with $c_{n-1}$ classes. Let $\Delta x$ be the variation in the number of components when adding $x$ alone, $\Delta y$ for $y$ alone and similarly $\Delta(x, y)$ when adding both $x$ and $y$. We distinguish between cases:

i) $\Delta x > 0, \Delta y > 0$ or both

$$\text{if} \quad x \not\mathcal{R} y \qquad \Delta(x, y) = \Delta x + \Delta y$$

$$\text{if} \quad x \mathcal{R} y \qquad \Delta(x, y) = \Delta x + \Delta y - 1$$

so that

$$\Delta(x, y) \leq 1 + \min(\Delta x, \Delta y) \tag{4.42}$$

ii) $\Delta x \leq 0$ and $\Delta y \leq 0$

then

$$\Delta(x, y) \leq \min(\Delta x, \Delta y) \tag{4.43}$$

Let $\alpha$ designate the probability of creating a new component

$$\alpha = p(\Delta x = 1) = p(\Delta y = 1)$$

$$\beta = p(x \mathcal{R} y)$$

We have the inequality

$$\begin{aligned} \left(\Delta(x, y)\right) &\leq \xi\left(\min(\Delta x, \Delta y)\right) + \alpha\beta(2-\alpha) - \left(1 - \frac{n-1}{N}\right)^2 \xi(\Delta x) \\ &\leq \xi(\Delta x) - (1-\alpha)\left(\xi(\Delta x) + q\right) + \alpha\beta(2-\alpha) - \left(1 - \frac{n-1}{N}\right)^2 \xi(\Delta x) \end{aligned} \tag{4.44}$$

where $q$ satisfies

$$\mathcal{E}(\Delta x) = \alpha - (1 - \alpha)q$$

Expression ( 4. 4 4) can be transformed to yield

$$\mathcal{E}\Big(\Delta(x, y)\Big) \leq \mathcal{E}(\Delta x)\left[1 + \alpha - \left(1 - \frac{n-1}{N}\right)^2\right] + \alpha\,(1 - 2\beta) - \alpha^2\,(1 - \beta) \tag{4. 45}$$

thus the conjecture is at least correct when $\beta \geq \frac{1}{2}$ if $\alpha > \left(1 - \frac{n-1}{N}\right)^2$.

The condition is explicitly

$$\mathcal{E}(\Delta x)\left[1 - \frac{1}{\alpha}\left(1 - \frac{n-1}{N}\right)^2\right] + 1 + \alpha\beta - \alpha - 2\beta \leq 0 \tag{4. 46}$$

Case $C_N > 1$

If the graph corresponding to $\mathcal{P}_N$ is not connected, the expected number of components can be derived from the knowledge of the individual outcomes when sampling each component of $\mathcal{P}_N$ independently.

Let $\xi\left(c_i(n_i)\right)$ be the expected number of component when selecting uniformly without replacement $n_i$ vertices from the $i^{th}$ component of $\mathcal{P}_N$ such that

$$0 \leqslant n_i \leqslant s_i \quad , \quad \sum_{i=1}^{C_N} s_i = N$$

Then

$$\xi(c_n) = \frac{\displaystyle\sum_{n_1=0}^{\min(n,\,r_1)} \; \sum_{n_2=0}^{\min(n-n_1,r_2)} \cdots \sum_{n_{C_{N-1}}=0}^{\min(n-n_1-\cdots-n_{C_{N-2}},\,r_{C_{N-1}})} E}{\displaystyle\sum \sum \cdots \sum_{n_1+n_2+\cdots+n_{C_N}=n} \binom{r_1}{n_1}\binom{r_2}{n_2}\cdots\binom{r_{C_N}}{n_{C_N}}} \tag{4.47}$$

where

$$E = \binom{r_1}{n_1}\cdots\binom{r_{C_N}}{n_{C_N}}\left[\xi\left(c_1(n_1)\right)+\cdots+\xi\left(c_{C_N}(n_{C_N})\right)\right] \tag{4.48}$$

If we pose now

$$W\,(r_1,\ r_2,\ \ldots,\ r_{C_N}\ ;\ n_1,\ n_2,\ \ldots,\ n_{C_N}) \ = \ \frac{\prod_{i=1}^{C_N}\binom{r_i}{n_i}}{\binom{N}{n}} \qquad (4.49)$$

then we have simply

$$\xi(c_n) = \sum_{n_1+n_2+\ldots+n_{C_N}=n} W(r_1, r_2, \ldots, r_{C_N}; n_1, n_2, \ldots, n_{C_N}) \sum_{i=1}^{C_N} \xi\Big(c_i(n_i)\Big) \qquad (4.50)$$

the summation being performed over all points with positive integer coordinates in the hyperplane

$$n_1 + n_2 + \ldots + n_{C_N} \ = \ n$$

The weight function is simply a hypergeometric probability distribution in $C_N$ dimensions.

Case $C_N = 2$

Let

$$F_2(\vec{r}; n) = \sum_{n_1=0}^{n} \frac{\binom{r_1}{n_1}\binom{r_2}{n-n_1}}{\binom{N}{n}} \left[ c_1(n_1) + c_2(n-n_1) \right] \tag{4.51}$$

where $F_2(\vec{r}; n)$ means $\xi(c_n)$ evaluated when $\vec{r} = (r_1, r_2)$. We can approximate the hypergeometric distribution $w(n_1) = \frac{\binom{r_1}{n_1}\binom{r2}{n_2}}{\binom{N}{n}}$ by a binomial probability law since the first two moments of $w(n_1)$ are $\frac{n\,r_1}{N}$ and $\frac{n\,r_1 r_2}{N^2}\,\frac{N-n}{N-1}$ which, if $n \ll N$, coincide with those of a binomial probability law with parameters $n$ and $\frac{n\,r_1}{N}$

$$w(n_1) \simeq \binom{n}{n_1} \left(\frac{r_1}{N}\right)^{n_1} \left(\frac{r_2}{N}\right)^{n_2} \tag{4.52}$$

But for large values of $n$, the binomial distribution can itself be approximated by a Poisson distribution of the form

$$w(n_1) \simeq e^{-n\frac{r_1}{N}} \frac{\left(\frac{n\,r_1}{N}\right)^{n_1}}{n_1!} \tag{4.53}$$

so that we have approximately

$$F_2(\vec{r}; n) \simeq \sum_{n_1=0}^{n} e^{-n\frac{r_1}{N}} \frac{\left(\frac{n\,r_1}{N}\right)^{n_1}}{n_1!} \left[ c_1(n_1) + c_2(n-n_1) \right] \tag{4.54}$$

# CHAPTER V

## Discrete and Continuous Adjacency Problems

In this chapter, we examine adjacency problems and compare the results obtained in 1, 2 and 3 dimensions. This analysis is rather detailed because those clustering models have potential applications.

The first problem is an adjacency problem for discrete positions on the line with extension to adjacency on a 2 dimensional rectangular chessboard. Although the same method would work in 3 dimensions the formulas have not been explicitly written. The second problem is a continuous clustering problem in 1, 2 and 3 dimensions which reduces to an overlap problem on the line and in a circle. As we expect, analytical solutions including the distribution functions, are found in the 1 dimensional case. However, in 2 and 3 dimensions, only an approximate analysis can be performed but leads to usable estimates.

## 5.1. Discrete Clustering On the Line

This is of course the least frustrating of all cases since most results have convenient analytical forms. Let us consider a discrete clustering problem on the line.

Let $\{x\}$ be the set $\{x_1, x_2, \ldots, x_n\}$

Two elements $x_i, x_j \in \{x\}$ are adjacent if $|i-j| = 1$. In some cases we shall deal with periodic sets, i.e. where $x_1$ and $x_n$ are considered to be adjacent, $|i-j| = 1$ modulo $n-2$. A cluster $\{y\}$ is a subset of $\{x\}$ such that

$$\exists \text{ at most } 2 \ x_i \in \{y\} \ni \text{either } x_i \not\mathcal{R}\, x_{i+1} \text{ or } x_i \not\mathcal{R}\, x_{i-1}$$

$\mathcal{R}$ being the adjacency relationship.

We now perform an experiment which is the selection of k x's out of n with equal probability. What is the expected number of clusters?

Let A(n, k, c) denote the number of arrangements of k 1's and (n-k) 0's forming c clusters. Introduce an extra 0 in any of the n+1 available positions, thus raising n to n+1 for constant k; if the number of clusters is now c+1, we can write the recurrence formula

$$A(n+1, k, c+1) = (n+c-k+2)\,A(n, k, c+1) + (k-c)\,A(n, k, c) \qquad (5.1)$$

meaning that either we had already (c+1) clusters on n points and the new 0 did not fall amid any existing cluster, or we had only c clusters, one of which was broken into two. The probability of obtaining c clusters is related to the number of arrangements by

$$p(n,k,c) = \frac{A(n,k,c)}{k!\ (n-k)!}\ \frac{1}{\binom{n}{k}} = \frac{A(n,k,c)}{n!} \tag{5.2}$$

Recurrence formula (5.1) now yields

$$p(n+1,k,c+1)(n+1) = (n+c-k+2)\,p(n,k,c+1) + (k-c)\,p(n,k,c) \tag{5.3}$$

or

$$p(n,k,c) = (1 - \frac{k-c}{n})\ p(n-1,\ k,\ c) + (\frac{k-c+1}{n})\ p(n-1,k,c-1) \tag{5.4}$$

Define the generating function $g_{n,k}(z)$ by

$$g_{n,k}(z) = \sum_{1 \leq c \leq k} p(n,k,c)\ z^{c} \tag{5.5}$$

By recurrence formula (5.4), $g_{n,k}(z)$ can be expressed as

$$g_{n,k}(z)=\sum_{1\leqslant c\leqslant k}\left[p(n-1,k,c)(1-\frac{k-c}{n})+p(n-1,\ k,c-1)(\frac{k-c+1}{n})\right] z^{c}$$

$$=(1-\frac{k}{n})\sum_{1\leqslant c\leqslant k}p(n-1,k,\ c)\ z^{c}+\frac{1}{n}\sum_{1\leqslant c\leqslant k}p(n-1,k,\ c)cz^{c}$$

$$+(\frac{k+1}{n})\sum_{1\leqslant c\leqslant k}p(n-1,k,c-1)z^{c}-\frac{1}{n}\sum_{1\leqslant c\leqslant k}p(n-1,k,c-1)cz^{c} \qquad (5.6)$$

By differentiation of $g_{n,k}(z)$ we get

$$zg'_{n,k}(z)=\sum_{1\leqslant c\leqslant k}p(n,k,c)cz^{c}$$

$$\sum_{1\leqslant c\leqslant k}p(n-1,k,c-1)z^{c}=zg_{n-1,k}(z)-p(n-1,k,k)z^{k+1}$$

$$\sum_{1\leqslant c\leqslant k}p(n-1,k,c-1)cz^{c}=z\left[g_{n-1,k}(z)+zg'_{n-1,k}(z)-(k+1)p(n-1,k,k)z^{k}\right]$$

so that the generating function must satisfy the differential equation

$$g_{n,k}(z)=(1-\frac{k}{n})g_{n-1,k}(z)+\frac{1}{n}zg'_{n-1,k}(z)+$$

$$+\left(\frac{k+1}{n}\right)\left[zg_{n-1,k}(z)-p(n-1,k,k)z^{k+1}\right]$$

$$-\frac{1}{n}\left[zg_{n-1,k}(z)+z^{2}g'_{n-1,k}(z)-(k+1)p(n-1,k,k)z^{k+1}\right]$$

$$=\frac{1}{n}\left[(n-k+kz)\ g_{n-1,k}(z)+z(1-z)g'_{n-1,k}(z)\right] \qquad (5.7)$$

From that equation, we can now derive the expected value of c as

$$g'_{n,k}(1) = \sum_{1 \leqslant c \leqslant k} p(n, k, c)c \tag{5.8}$$

but

$$g'_{n,k}(z)=\frac{1}{n}\left[kg_{n-1,k}(z)+(n-k+kz+1-2z)g'_{n-1,k}(z)+z(1-z)g''_{n-1,k}(z)\right]$$

which yields

$$g'_{n,k}(1) = \frac{1}{n}\left[kg_{n-1,k}(1) + (n-1)g'_{n-1,k}(1)\right] \tag{5.9}$$

The first term is

$$g_{n-1,k}(1) = \sum_{1 \leqslant c \leqslant k} p(n-1, k, c) = 1$$

so that finally

$$g'_{n,k}(1) = (1 - \frac{1}{n})\, g'_{n-1,k}(1) + \frac{k}{n} \tag{5.10}$$

It is easy to see that the general term satisfying relation (5.10) is of the form

$$g'_{n,k}(1) = \frac{(n-k+1)k}{n}$$

since $g'_{k,k}(1) = 1$ and

$$g'_{n+1,k}(1) = \frac{n}{n+1}\,\frac{(n-k+1)k}{n} + \frac{k}{n+1} = \frac{(n-k+2)k}{n}$$

which proves that the expected number of clusters of 1's is

$$\mathcal{E}(c_1) = \frac{(n-k+1)k}{n} \tag{5.11}$$

A similar result clearly holds for the number of clusters of 0's

$$\mathcal{E}(c_o) = \frac{(n-k)(k+1)}{n} \tag{5.12}$$

and if $c=c_1+c_o$ then

$$\mathcal{E}(c) = 1 + 2k(1 - \frac{k}{n}) \tag{5.13}$$

It is tempting to derive some further results from the differential equation (5.7). For instance

$$\operatorname{var}(c_1) = \sum_{1\leqslant c\leqslant k} c^2 p(n,k,c) - \left[\sum_{1\leqslant c\leqslant k} cp(n,k,c)\right]^2$$

$$= z\left[g'_{n,k}(z) + zg''_{n,k}(z)\right] - g'_{n,k}(z)^2 \bigg|_{z=1}$$

$$= g''_{n,k}(1) + g'_{n,k}(1) - g'_{n,k}(1)^2 \qquad (5.14)$$

The second derivative of $g_{n,k}(z)$ has the form

$$g''_{n,k}(z) = \frac{1}{n}\Big[kg'_{n-1,k}(z) + (k-2)g'_{n-1,k}(z) + (n-k+kz+1-2z)g''_{n-1,k}(z) +$$

$$+ (1-z)g''_{n-1,k}(z) - zg''_{n-1,k}(z) + z(1-z)g'''_{n-1,k}(z)\Big] \qquad (5.15)$$

which evaluated at $z=1$ yields

$$g''_{n,k}(1) = \frac{1}{n}\left[(n-2)g''_{n-1,k}(1) + \frac{2k(k-1)(n-k)}{n-1}\right] \qquad (5.16)$$

The first terms produced by this recurrence relation are

$$g''_{k,k}(1) = 0 \quad ,$$

$$g''_{k+1,k}(1) = \frac{2(k-1)}{k+1} \quad ,$$

$$g''_{k+2,k}(1) = \frac{6k(k-1)}{(k+1)(k+2)} \quad , \ \ldots ,$$

and we can prove by induction that the general term is

$$g''_{n,k}(1) = \frac{k(k-1)(n-k)(n-k+1)}{n(n-1)} \tag{5.17}$$

since

$$g''_{n+1,k}(1) = \frac{1}{n+1}\left[\frac{k(k-1)(n-k)(n-k+1)(n-1)}{n(n-1)} + \frac{2k(k-1)(n+1-k)}{n}\right]$$

$$= \frac{k(k-1)(n-k+1)(n-k+2)}{(n+1)n}$$

Therefore, the variance of $c_1$ is equal to

$$\text{var}(c_1) = \frac{k(k-1)(n-k)(n-k+1)}{n(n-1)} + \frac{k(n-k+1)}{n}\left(1 - \frac{k(n-k+1)}{n}\right)$$

$$= \frac{k(k-1)(n-k)(n-k+1)}{n^2(n-1)} \tag{5.18}$$

Similarly

$$\text{var}(c_o) = \frac{k(k+1)(n-k)(n-k-1)}{n^2(n-1)} \tag{5.19}$$

## 5.2 Explicit Representation of the Generating Function

Instead of dealing with recurrence formula (5.7), we could have determined a series representation of $g_{n,k}(z)$. We first show by induction on $n$ that the series can be written in the form

$$g_{n,k}(z) = \sum_{\nu=1}^{\min(k-1,\,n-k+1)} a_{n-k+1,\nu} \frac{(k-1)!\;k!}{(k-\nu)!\;n!}\, z^{\nu} \tag{5.20}$$

with

$$g_{k,k}(z) = z$$

Replacing $g_{n,k}(z)$ in the recurrence relation (5.7) we obtain successively

$$g_{n+1,k}(z)=\frac{1}{n+1}\left[(n+1-k+kz) \sum_{\nu=1}^{\min(k-1,\,n-k+1)} a_{n-k+1,\nu} \frac{(k-1)!\,k!}{(k-\nu)!\,n!}\, z^{\nu} \right.$$

$$\left. + z(1-z) \sum_{\nu=1}^{\min(k-1,\,n-k+1)} a_{n-k+1,\nu} \frac{(k-1)!\,k!}{(k-\nu)!\,n!}\, \nu z^{\nu}\right] \tag{5.21}$$

$$= \sum_{\nu=1}^{\min(k-1,\,n-k+1)} a_{n-k+1,\nu} \frac{(k-1)!\,k!}{(k-\nu)!\,(n+1)!} \left[(n+1-k+\nu)z^{\nu} + (k-\nu)z^{\nu+1}\right]$$

$$= \sum_{\nu=1}^{\min(k-1,\,n-k+2)} a_{n-k+2,\nu} \frac{(k-1)!\,k!}{(k-\nu)!\,(n+1)!}\, z^{\nu} \tag{5.22}$$

where $a_{n-k+2,\nu}$ satisfies

$$a_{j,\nu} = (j+\nu)\, a_{j-1,\nu} + a_{j-1,\nu-1}$$

$$a_{j,0} = 0 \quad \forall j\,, \quad a_{0,1} = 1$$

Table ( 5. 1) shows the first values of a.

The first terms of $g_{n,k}(z)$ are explicitly

$$g_{k,k} = z$$

$$g_{k+1,k} = \frac{2z + (k-1)z^2}{k+1}$$

$$g_{k+2,k} = \frac{6z + 6(k-1)z^2 + (k-1)(k-2)z^3}{(k+1)(k+2)}$$

$$g_{k+3,k} = \frac{24z + 36(k-1)z^2 + 12(k-1)(k-2)z^3 + (k-1)(k-2)(k-3)z^4}{(k+1)(k+2)(k+3)}$$

These coefficients can still be expressed in another way. For convenience, let $\ell = n-k+1$. Then, if we write that $g_{\ell+i-1,i}(1) = 1$ for $i = 1, 2, \ldots, k+1$ we obtain a system of linear equations in the unknowns $a_\ell$

$$\sum_{\nu=1}^{\min(i-1,\ell)} a_{\ell,\nu} \frac{(i-1)!}{(i-\nu)!} = \frac{(\ell+i-1)!}{i!} \tag{5.23}$$

or in matrix form

$$\underbrace{\begin{pmatrix} 1 & & & & \\ 1 & 1 & & 0 & \\ 1 & 2 & 2 & & \\ 1 & 3 & 6 & 6 & \\ & \cdots & & & \\ 1 & k & k(k-1) & k(k-1)(k-2) & k! \end{pmatrix}}_{A} \begin{pmatrix} a_{\ell,1} \\ a_{\ell,2} \\ \\ \\ \\ a_{\ell,k+1} \end{pmatrix} = \begin{pmatrix} \ell!/1! \\ (\ell+1)!/2! \\ (\ell+2)!/3! \\ \\ \\ (\ell+k)!/(k+1)! \end{pmatrix}$$

The inverse of A is found without difficulty since it is also triangular

$$A^{-1} = \begin{pmatrix} \binom{0}{0}/0! & & & & \\ -\binom{1}{0}/1! & \binom{1}{1}/1! & & 0 & \\ \binom{2}{0}/2! & -\binom{2}{1}/2! & \binom{2}{2}/2! & & \\ & & & & \\ (-1)^k\binom{k}{0}k! & (-1)^{k+1}\binom{k}{1}/k! & & & \binom{k}{k}/k! \end{pmatrix} \tag{5.24}$$

with general element

$$(A^{-1})_{i,j} = (-1)^{i+j}\,\frac{\binom{i-1}{j-1}}{(i-1)!} \qquad j \leq i \;, \quad \text{otherwise } 0$$

It suffices to verify that indeed $(AA^{-1})_{i,j} = \delta_{ij}$

$$(AA^{-1})_{i,j} = \sum_{j\leq k\leq i} \frac{(i-1)!}{(i-k)!} (-1)^{k+j} \frac{\binom{k-1}{j-1}}{(k-1)!}$$

$$= \sum_{0\leq m\leq i-j} (-1)^{m} \binom{i-1}{m+j-1}\binom{m+j-1}{j-1}$$

$$= \binom{i-j}{j-1} \sum_{0\leq m\leq i-j} (-1)^{m} \binom{i-j}{m} = \delta_{ij}$$

Now, the coefficients of the series are obtained using the inverse matrix; we obtain

$$g_{n,k}(z) = \sum_{\nu=1}^{\min(k-1,\, n-k+1)} \frac{(k-1)!\, k!}{(k-\nu)!\, n!} \sum_{j=1}^{\nu} (-1)^{\nu+j} \frac{\binom{\nu-1}{j-1}}{(\nu-1)!} \frac{(n-k+j)!}{j!}$$

$$= \sum_{\nu=1}^{\min(k-1,\, n-k+1)} \sum_{j=1}^{\nu} (-1)^{\nu+j} \frac{\binom{k-1}{\nu-1}\binom{\nu-1}{j-1}\binom{n-k+j}{n-k}}{\binom{n}{k}} z^{\nu}$$

$$= \sum_{\nu=1}^{\min(k-1,\, n-k+1)} (-1)^{\nu} \frac{\binom{k-1}{\nu-1}}{\binom{n}{k}} (-1)^{\nu} \binom{n-k+1}{\nu} z^{\nu}$$

$$= \sum_{\nu=1}^{\min(k-1,\, n-k+1)} \frac{\binom{k-1}{\nu-1}\binom{n-k+1}{\nu}}{\binom{n}{k}} z^{\nu} \tag{5.25}$$

Identifying corresponding powers of $z$ in (5.5) and (5.25) we find

$$p(n,k,c) = \frac{\binom{k-1}{c-1}\binom{n-k+1}{c}}{\binom{n}{k}}$$

## 5.3. Discrete Clustering on a Rectangular Board

A normal extension of the preceding problem is to consider a similar adjacency rule for a 2 dimensional chessboard.

The particular problem that we examine is to estimate the average number of clusters formed by $k$ marks occupying $k$ distinct positions selected with equal probability from the $nm$ available ones of an $n \times m$ rectangular board $B_{nm}$. Hence $b_{ij} = 1$ or $0$ depending upon whether a mark is present or not in the $(i, j)$ position. Two simple adjacency rules are reviewed:

i) on $B_{nm}$, $(i_1, j_1)$ and $(i_2, j_2)$ are adjacent if $|i_1 - i_2| + |j_1 - j_2| = 1$

ii) on $B^*_{nm}$, they are adjacent if

$$|i_1 - i_2| = 1 \quad \text{mod} \quad n-2 \quad \text{and} \quad j_1 = j_2$$

or

$$|j_1 - j_2| = 1 \quad \text{mod} \quad m-2 \quad \text{and} \quad i_1 = i_2$$

Therefore, two marks are adjacent if they occupy positions next to each other in either the same row or the same column. $B^*_{nm}$ is assumed to be wrapped around on a torus so that opposite sides of the board become adjacent. The adjacency rule is clearly an equivalence relation and determines clusters among the $k$ marks.

The complement of a cluster of $k$ marks is a set of $nm-k$ marks $b_{ij} = 0$. On $B^*_{nm}$ we make the following definitions: a cluster is called linear if its complement on $B^*_{nm}$ consists of only one cluster, otherwise it is called cyclic; the complement of a cluster always consists of linear clusters. The number of cycles of a cluster is the number of cuts to be performed to obtain a linear cluster out of a cyclic cluster. The meaning of a cut is intuitively clear: it is the removal from $\Gamma$ of a linear cluster $\subset \Gamma$ which possesses marks adjacent to exactly two distinct linear clusters of the complement of $\Gamma$ on $B^*_{nm}$. Therefore, if the complement of $\Gamma$ is made up of $\nu$ linear clusters, $\nu - 1$ cuts will be necessary to connect them while transforming $\Gamma$ into a linear cluster.

To perform the actual counting we now make use of some nice topological properties of linear and cyclic clusters. Assume that with every mark on $B^*_{nm}$, we associate a weight 1, 0 or -1 depending upon the occupancy of three neighboring positions north, west and northwest as follows:

$w_{ij} = 1$ if N, W are both empty (type +)

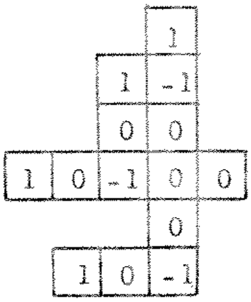
$w_{ij} = -1$ if only NW is empty (type -)

$w_{ij} = 0$ otherwise (type 0)

<u>Lemma</u>

$$\sum_i \sum_j w_{ij} b_{ij} = \# \text{ clusters} - \# \text{ cycles} \qquad (5.26)$$

proof: we decide to ignore any mark which has either an N or W neighbor, or all three N, W, NW neighbors, letting either W or N act as a possible cluster representative. A cluster representative, assigned a weight 1, is in an upper left corner position with neither an N nor a W neighbor (NW is immaterial since it is not directly adjacent). However, even a linear cluster can have several such cluster representatives as shown in the figure below. But if it does have $\nu$ + marks, then it must also have $\nu - 1$ - marks to satisfy connectivity (this can be shown by examining the perimeter curve). The - mark then carries the information that the clusters stemming from its N and W neighbors are actually parts of the same unique cluster and its weight is set equal to - 1 to compensate for the fact that they will be counted twice. However, this assumption does not hold if the N and W clusters are indeed connected elsewhere to form a cycle in which case we have inadvertently reduced the class count by 1 for each cycle.



Designate by $c(nm, k)$ and $c^*(nm, k)$ the expected number of clusters on boards $B_{nm}$ and $B^*_{nm}$, respectively. We have the following result

Theorem

$$c^*(nm, k) = \frac{k(1 - k\mathcal{E})}{(1 - \mathcal{E})(1 - 2\mathcal{E})} \left[ 1 - (k+1)\mathcal{E} - \beta(nm, k) \frac{(k-1)(k-2)\mathcal{E}^2}{(1 - 3\mathcal{E})} \right] \tag{5.27}$$

$$c(nm, k) = k\mathcal{E}\left[1 + \frac{(n+m-2)(1-k\mathcal{E})}{1-\mathcal{E}}\right] + \mathcal{E}(n-1)(m-1)\, c^{*}(nm, k) \tag{5.28}$$

where $\beta(nm, k)$ is monotone decreasing for fixed $nm$ and satisfies

$$1 \geqslant \beta(nm, k) \geqslant -\frac{1}{3}$$

proof: using the counting argument presented in the last lemma, we can write immediately

$$c^{*}(nm, k) = k \frac{\binom{nm-3}{k-1} - \beta(nm, k)\binom{nm-4}{k-3}}{\binom{nm-1}{k-1}} \qquad k \leqslant nm - 3 \tag{5.29}$$

in which $\beta(nm, k)$ is an unknown function representing the fraction of - marks actually belonging to linear clusters; therefore $\beta(nm, k) = 1$, $1 \leqslant k \leqslant 7$ since the smallest cyclic cluster is of size 8. After simplification

$$c^{*}(nm, k) = \frac{k(nm-k)}{(nm-1)(nm-2)}\left[nm - k - 1 - \beta(nm, k)\frac{(k-1)(k-2)}{nm-3}\right] \tag{5.30}$$

To derive a similar expression for $B_{nm}$, board with no wrap around, we have to distinguish between four regions:

1) $(1, 1)$ $\qquad w_{11} = 1$ if $b_{11} = 1$

2) $(1, j)$ $\quad 2 \leqslant j \leqslant m$ $\qquad w_{1j} = 1$ if $b_{1\,j-1} = 1$

3) $(i, 1)$ $\quad 2 \leqslant i \leqslant n$ $\qquad w_{i1} = 1$ if $b_{i-1\,1} = 1$

4) otherwise apply the same rule as for $B^{*}(n, m)$.

We now proceed as we did before to obtain

$$c(nm, k) = \frac{k}{nm} + k\,\frac{(n+m-2)}{nm}\,\frac{\binom{nm-2}{k-1}}{\binom{nm-1}{k-1}} + \frac{(n-1)(m-1)}{nm}\;c^{*}(nm, k)$$

$$= \frac{k}{nm}\left[1 + \frac{(n+m-2)(nm-k)}{nm-1}\right] + \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{m}\right) c^{*}(nm, k) \tag{5.31}$$

For convenience of notation, let $\mathcal{E} = \frac{1}{nm}$ . Then formulas (5.30) and (5.31) can be rewritten in the form

$$c^{*}(nm, k) = \frac{k(1 - k\mathcal{E})}{(1 - \mathcal{E})(1 - 2\mathcal{E})}\left[1 - (k+1)\mathcal{E} - \beta(nm, k)\,\frac{(k-1)(k-2)\mathcal{E}^{2}}{1 - 3\mathcal{E}}\right] \tag{5.32}$$

$$c(nm, k) = k\mathcal{E}\left[1 + \frac{(n+m-2)(1 - k\mathcal{E})}{1 - \mathcal{E}}\right] + \mathcal{E}(n-1)(m-1)c^{*}(nm, k) \tag{5.33}$$

If $\mathcal{E} << 1$ , that is for a sufficiently large board, we obtain the asymptotic formula

$$c^{*}(nm, k) = k - 2\mathcal{E}\,k(k-1) + k\mathcal{E}^{2}\left[k^{2}(1-\beta) + k(3\beta-5) + 4-2\beta\right] + \mathcal{O}(k^{4}\mathcal{E}^{3}) \tag{5.34}$$

$$c(nm, k) = k - 2\mathcal{E}\,k(k-1) + k\mathcal{E}^{2}\left[k^{2}(1-\beta) + k(m+n+3\beta-5) + 4-2\beta-m-n\right] + \mathcal{O}(k^{4}\mathcal{E}^{3}) \tag{5.35}$$

Substituting k by nm-3 in formula (5.34) gives the limiting value $\beta(nm, nm-3)$ since $c^{*}(nm, nm-3) = 1$ . We find

$$\beta(nm, nm-3) = \left[2\mathcal{E} - \frac{(1 - \mathcal{E})(1 - 2\mathcal{E})}{3(1 - 3\mathcal{E})}\right]\frac{(1 - 3\mathcal{E})}{(1 - 4\mathcal{E})(1 - 5\mathcal{E})}$$

$$= \frac{-1 + 9\mathcal{E} - 20\mathcal{E}^{2}}{3(1 - 4\mathcal{E})(1 - 5\mathcal{E})}$$

$$= -\frac{1}{3}$$

In the interval $8 \leq k \leq nm-4$ however, $\beta$ truly depends on $nm$. For instance

$$\beta(nm, nm-4) = -\frac{1}{4} + \mathcal{O}(\varepsilon) \tag{5.36}$$

This analysis can indeed be viewed in several ways:

i) if we ignore the variation of $\beta$ with $k$ and set $\beta = 1$ formula (5.35) provides a lower bound for the expected number of clusters.

ii) alternatively, an estimated variation of $\beta$ can be used to compute better estimates of $\beta$ than the lower bound, especially if $k >> 1$

iii) from a different standpoint, if we are willing to perform Monte Carlo calculations to compute the average number of classes, we obtain numerical estimates of $\beta(nm, k)$ which measure the average number of cycles in the clusters on $B^*_{nm}$, explicitly

$$\begin{aligned} \mathcal{E}(\#\,\text{cycles}) &= \left[1 - \beta(nm, k)\right] \frac{k(k-1)(k-2)\,\varepsilon^2(1-k\varepsilon)}{(1-\varepsilon)(1-2\varepsilon)(1-3\varepsilon)} \\ &= \left[1 - \beta(nm, k)\right] k^3 \varepsilon^2 \left(1 + \mathcal{O}\left(\frac{1}{k}\right)\right) \end{aligned} \tag{5.37}$$

Although being a by-product of the original estimation problem, this is definitely an interesting result in itself.

The approach we have just taken could quite easily provide other results such as the expected perimeter length for the clusters and be applied to spaces of higher dimensionality.

## 5.4. Continuous Clustering On An Interval

After considering a discrete occupancy problem, we now examine the continuous case of *overlapping segments on an interval.* Let k segments of equal length $\ell$ fall on an interval $[A, B]$ of length 1, such that the probability distribution of the middle of the segments is uniform, equal to 1 on AB, 0 outside.

Definition - A k-clump is a set of k overlapping segments such that there exists a completely covered interval $[\alpha, \beta]$ containing all k segments and there is no other interval in which $[\alpha, \beta]$ is properly included.

When scanning $[A, B]$ from A to B, a new clump begins whenever a segment of length $\ell$ void of marks extends to the left of a given mark. The probability of finding such a segment to the left of x is

$$P\left\{0 \text{ mark in } [x - \ell, x]\right\} = \left[1 - \min(x, \ell)\right]^{k-1} \tag{5.38}$$

so that the probability of starting a clump at x is

$$\left[1 - \min(x, \ell)\right]^{k-1} p(\text{mark at } x)dx = \left[1 - \min(x, \ell)\right]^{k-1} \tag{5.39}$$

The expected number of clumps is therefore

$$\mathcal{E}\left(c(k, \ell)\right) = k\int_0^1 \left[1 - \min(x, \ell)\right]^{k-1} dx$$

$$= k\int_0^{\ell} (1 - x)^{k-1} dx + k\int_{\ell}^{1} (1-\ell)^{k-1} dx = 1 + (1-\ell)^k (k-1) \tag{5.40}$$

When $\ell$ is small compared to 1, it is interesting to see the variation of $c(k, \mathcal{E})$; we have

$$\mathcal{E}\left(c(k, \ell)\right) = 1 + \left(1 - k\ell + \frac{k(k-1)}{2}\ell^2\right)(k-1) + \mathcal{O}(k^4\ell^3)$$

$$= k + k(1-k)\ell + \frac{k(k-1)^2}{2}\ell^2 + \mathcal{O}(k^4\ell^3) \qquad (5.41)$$

Several related results can of course be obtained at once. For instance, the density $\delta$ of clumps is

$$\delta = \frac{1}{k} + (1-\ell)^k \left(1 - \frac{1}{k}\right) \qquad (5.42)$$

and the expected length $\lambda$ covered by the clumps is

$$\lambda = 1 - \int_0^{\ell} (1-x)^k dx - \int_{\ell}^{1} (1-\ell)^k dx$$

$$= 1 - \frac{1-(1-\ell)^{k+1}}{k+1} - (1-\ell)^{k+1}$$

$$= \frac{k}{k+1}\left[1 - (1-\ell)^{k+1}\right] \qquad (5.43)$$

so that the average clump length within AB is

$$\lambda = \left(\frac{k}{k+1}\right)\left(\frac{1-(1-\ell)^{k+1}}{1+(1-\ell)^k(k-1)}\right) \qquad (5.44)$$

Another approach which has the advantage of leading to the actual distribution of the number of clumps is to compute the distribution of the k-1 interior intervals determined by k random points on a unit interval. If the left and right most intervals are discarded, it is clear that the probability of having c inner intervals greater than $\ell$ is the same as that of obtaining c+1 clumps.

Let the inner intervals be numbered $I_1, I_2, \ldots, I_{k-1}$ and write $P_{i_1 i_2 \ldots i_m}$ for the probability that $I_{i_1} > x$, $I_{i_2} > x, \ldots, I_{i_m} > x$. Let $S_\nu$ be the sum of $P_{i_1 i_2 \ldots i_m}$ for all subsets of $\nu$ intervals out of k-1. Let P($\nu$ intervals) be the probability that exactly $\nu$ intervals are greater than x under the constraint $\nu x \leq 1$. By applying the principle of inclusion and exclusion we obtain the relationship

$$P(\nu \text{intervals} > x) = \sum_{i=0}^{\min\left(\left\lfloor \frac{1}{x} \right\rfloor - \nu,\, k-\nu-1\right)} (-1)^i S_{\nu+i} \binom{\nu+i}{\nu} \tag{5.45}$$

and a typical $S_j$ term is

$$S_j = \sum_{\text{all j-subsets}} P_{i_1 i_2 \ldots i_j} = \binom{k-1}{j} P_{12\ldots j}$$

The probability $P_{12\ldots j}$ can itself be written as a multiple integral

$$S_j = \binom{k-1}{j} P(I_1 > x,\ I_2 > x,\ \ldots,\ I_j > x)$$

$$= \binom{k-1}{j} k! \int_x^{1-(j-1)x} dx_1 \int_{x_1+x}^{1-(j-2)x} dx_2 \int \cdots \int_{x_{j-1}+x}^{1} dx_j \int_{x_j}^{1} dx_{j+1} \int \cdots \int_{x_{k-1}}^{1} dx_k$$

Assume that

$$\int_{x_m}^{1} dx_{m+1} \int \cdots \int_{x_{k-1}}^{1} dx_k = \frac{(1 - x_m)^{k-m}}{(k - m)!}$$

This is obviously true for $m=k-1$ ; by induction on $m$ for $j \leqslant m \leqslant k-1$ , we prove that formula since

$$\int_{x_{m-1}}^{1} \frac{(1-x_m)^{k-m}}{(k-m)!} dx_m = - \frac{(1-x_m)^{k-m+1}}{(k-m+1)!} \Bigg|_{x_{m-1}}^{1} = \frac{(1-x_{m-1})^{k-m+1}}{(k-m+1)!}$$

To perform the last integration steps corresponding to $0 \leqslant m \leqslant j-1$ assume that

$$\int_{x_{m+x}}^{1-(j-m-1)x} dx_{m+1} \int \cdots \int_{x_{k-1}}^{1} dx_k = \frac{\left[1 - x_m + (m-j)x\right]^{k-m}}{(k - m)!}$$

which matches our previous result at $m=j$ . The next integral then becomes

$$\int_{x_{m-1}+x}^{1-(j-m)x} \frac{\left[1-x_m+(m-j)x\right]^{k-m}}{(k-m)!} dx_m = - \frac{\left[1-x_m+(m-j)x\right]^{k-m+1}}{(k-m+1)!} \Bigg|_{x_{m-1}+x}^{1-(j-m)x}$$

$$= \frac{\left[1-x_{m-1}+(m-j-1)x\right]^{k-m+1}}{(k-m+1)!}$$

$S_j$ is thus obtained as the value of the last expression for $m=1$ which is equal to

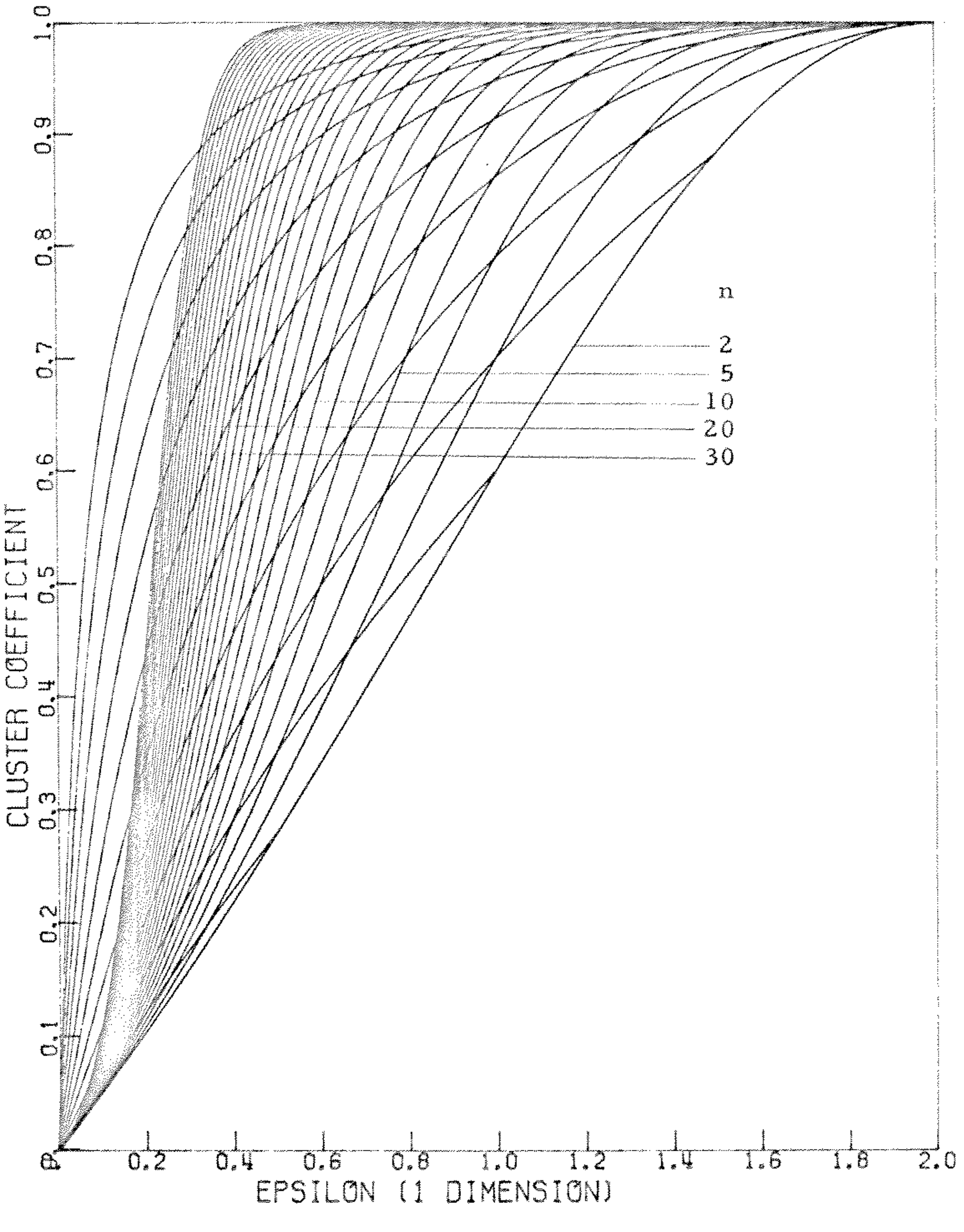

CLUSTER COEFFICIENT
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
EPSILON (1 DIMENSION)
0.
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
n
2
5
10
20
30

Figure 5.1

$$S_j = \binom{k-1}{j} k! \ \frac{(1-jx)^k}{k!} = \binom{k-1}{j}\left(1-jx\right)^k \qquad 0 \leq jx \leq 1 \qquad (5.46)$$

The probability of finding exactly $\nu$ inner intervals greater than x can therefore be represented by the summation

$$P(\nu \text{ intervals} > \ell) = \sum_{i=0}^{\min\left(\left\lfloor \frac{1}{\ell}\right\rfloor - \nu,\ k-\nu-1\right)} (-1)^i \binom{\nu+i}{\nu}\binom{k-1}{\nu+i}\left[1-(\nu+i)\ell\right]^k$$

$$= \binom{k-1}{\nu} \sum_{i=0}^{\min\left(\left\lfloor \frac{1}{\ell}\right\rfloor - \nu,\ k-\nu-1\right)} (-1)^i \binom{k-\nu-1}{i}\left[1-(\nu+i)\ell\right]^k \qquad (5.47)$$

As we have already mentioned, this result also provides the distribution of the number of clumps since having $\nu$ inner intervals greater than $\ell$ implies that there exist $\nu+1$ clumps, so that finally

$$P(\nu \text{ clumps}) = \binom{k-1}{\nu-1} \sum_{i=0}^{\min\left(\left\lfloor \frac{1}{\ell}\right\rfloor - \nu+1,\ k-\nu\right)} (-1)^i \binom{k-\nu}{i}\left[1-(\nu+i-1)\ell\right]^k \qquad (5.48)$$

Figure ( 5. 1 ) shows the variation of the cluster coefficient

$$\rho = \frac{\frac{n}{c} - 1}{n - 1} = \frac{1 - (1 - \frac{\mathcal{E}}{2})^n}{1 + (1 - \frac{\mathcal{E}}{2})^n (n-1)} \qquad (5.49)$$

as a function of $\mathcal{E}$ , length of the segments falling on an interval of length 2. Constant density trajectories $n\mathcal{E} = d$ are also graphed. They are given by

$$\rho = \frac{1 - \left(1 - \frac{\varepsilon}{2}\right)^{d/\varepsilon}}{1 + \left(1 - \frac{\varepsilon}{2}\right)^{d/\varepsilon}\left(\frac{d}{\varepsilon} - 1\right)}$$

which for $\varepsilon \to 0$ becomes

$$\rho \simeq \frac{1}{1 + \frac{d}{\varepsilon(e^{d/2} - 1)}} \tag{5.50}$$

so that the trajectories have a derivative at $\varepsilon = 0$ equal to

$$\left.\frac{d\rho}{d\varepsilon}\right|_{\varepsilon=0} = \frac{e^{d/2} - 1}{d} \tag{5.51}$$

## 5.5. Continuous Clustering in 2 and 3 Dimensions

We first examine the case of n overlapping discs of diameter $\varepsilon$ falling with uniform probability over a circle D of radius 1. The pdf of the distance of two random points in a circle can be obtained using Crofton's theorem. Following Kendall and Moran [ 17 ], suppose that n points are independently distributed in an r-dimensional domain D . The probability that a figure F formed by n points satisfies some condition only dependent upon the relative position of the points is

$$P = \frac{m^*(E)}{\left[m(D)\right]^n}$$

where $m^*(E)$ is the Lebesque measure in nr-dimensional space of the set E of points at which F has the required property.

Let $V = m(D)$ ; for $D_1 \supset D$ , $m(D_1) = V + \Delta V$ . In nr-dimensional space where $E_1 \supset E$ let

$$m^*(E) = U \qquad m^*(E_1) = U + \Delta U$$

The probability that F has the required property for random points in $D_1$ is

$$P_1 = P + \Delta P = \frac{U + \Delta U}{(V + \Delta V)^n}$$

The set $E_1$ can be divided into $(n+1)$ subsets $E_{1\nu}$ $(\nu = 0, 1, \ldots, n)$ such that for $E_{1\nu}$, $\nu$ points lie in $D_1$-$D$ and $n-\nu$ in $D$. Then

$$U + \Delta U = U + \sum_{\nu=1}^{n} \binom{n}{\nu} P_\nu V^{n-\nu}(\Delta V)^\nu$$

where $P_\nu$ is the probability that points in $E_{1j}$ form correct figures in $D_1$. From the last two equations

$$(P+\Delta P)(V+\Delta V)^n = PV^n + \sum_{\nu=1}^{n} \binom{n}{\nu} P_\nu V^{n-\nu}(\Delta V)^\nu$$

$$\Delta P (V + \Delta V)^n = \sum_{\nu=1}^{n} \binom{n}{\nu} (P_\nu - P)V^{n-\nu}(\Delta V)^\nu$$

and letting $\Delta V$ become small, we obtain Crofton's formula

$$\delta P = n (P_1 - P) \frac{\delta V}{V} \tag{5.52}$$

Using this result, let us compute the pdf of the distance of two points distributed with uniform probability over a circle $\Gamma$ of radius $R$. Let $p(x, R)dx$ be the probability that the random segment $AB$ has a length between $x$ and $x + dx$. Similarly, let $p_1(x, R)\,dx$ be the corresponding probability when $A$ is constrained to be on the circumference while $B \epsilon \Gamma$, $\Gamma$ representing the set of points at a distance less than or equal to $R$ from the center. Geometrically, we find

$$p_1(x,\ R) = \frac{2x\theta}{\pi R^2}$$

where

$$\theta = a \cos \frac{x}{2R}$$

If we hold x fixed and differentiate, we obtain

$$\tan \theta d\theta = \frac{dR}{R}$$

Using Crofton's formula, we now find

$$\frac{dp}{d\theta} = 2\left( \frac{2x\theta}{\pi R^2} - p\right) 2 \tan \theta$$

or

$$\frac{dp}{d\theta} + 4\, p \tan \theta = \frac{16\, \theta \sin 2\theta}{\pi x} \qquad (5.53)$$

A particular solution of the associated differential equation without right hand side is $p = \cos^4\theta$ ; we look for solutions of the form $\lambda \cos^4\theta$ where $\lambda$ satisfies

$$d\lambda = \frac{32\theta \sin \theta}{\pi x \cos^3 \theta} d\theta = \frac{16\theta}{\pi x}\, d\left(\frac{1}{\cos^2\theta}\right) \qquad (5.54)$$

which yields after quadrature

$$p = \frac{16}{\pi x} \cos^2\theta\, (\theta - \sin\theta\cos\theta) + c\cos^4\theta$$

but $\mu = 0$ since $p = 0$ for $\theta = 0$. Since $x = 2R\cos\theta$

$$p = \frac{8}{\pi R} \cos\theta\, (\theta - \sin\theta\cos\theta)$$

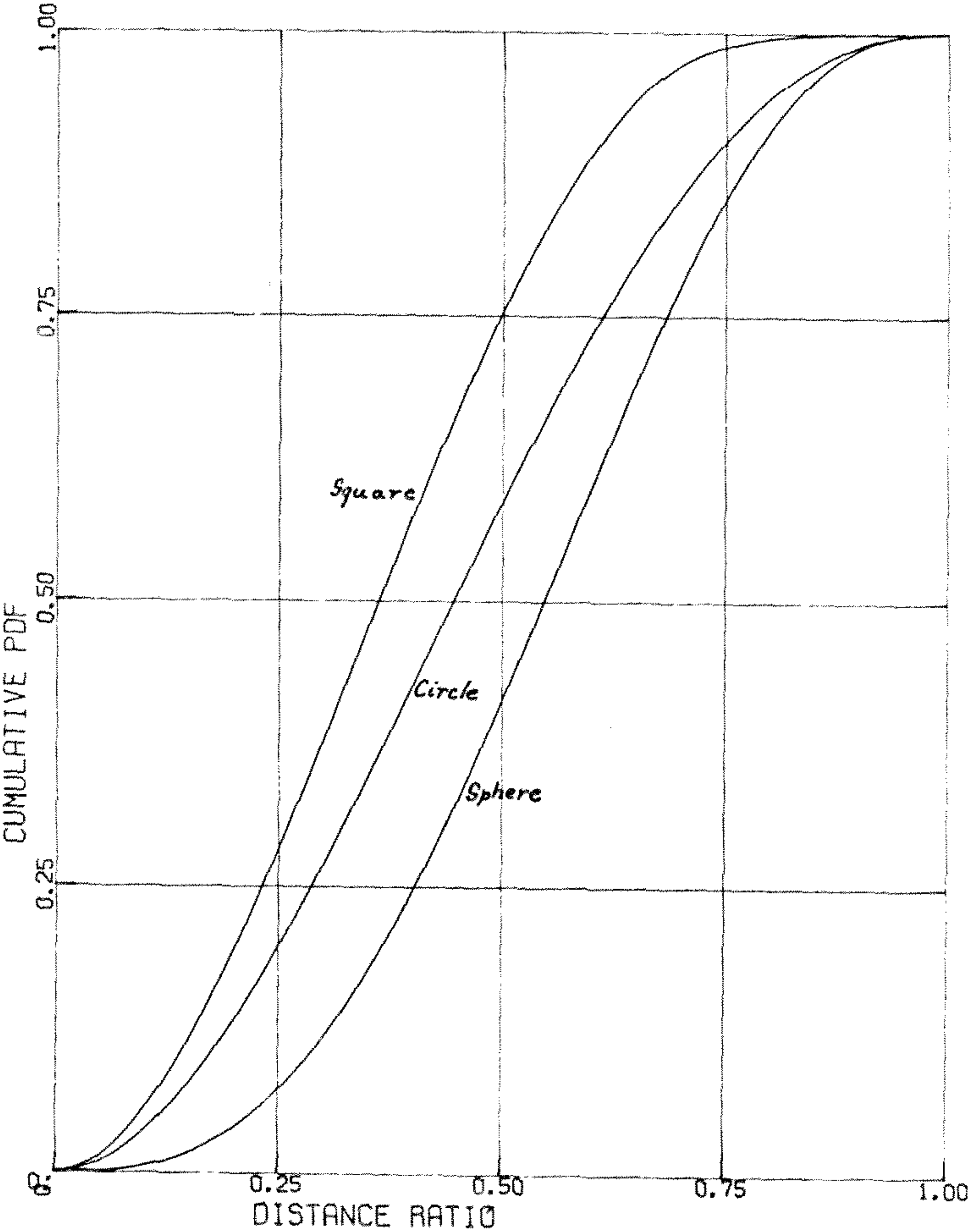

CUMULATIVE PDF
0
0.25
0.50
0.75
1.00
DISTANCE RATIO
0.25
0.50
0.75
1.00
Square
Circle
Sphere

Figure 5.2

$$p(x, R) = \frac{4x}{\pi R^2}\left[a\cos\frac{x}{2R} - \frac{x}{2R}\sqrt{1 - \frac{x^2}{4R^2}}\right] \tag{5.55}$$

Let $\alpha = \frac{x}{R}$ ; the cumulative distribution is obtained by integration with respect to x and yields

$$p(d \leqslant x,\ R) = 1 - \frac{2\theta}{\pi}(1 - \alpha^2) - \frac{\sin 2\theta}{\pi}\left(1 + \frac{\alpha^2}{2}\right) \tag{5.56}$$

The corresponding distributions for 2 random points in a unit square are simpler to obtain and give [ 16 ]

$$\begin{aligned} p(x, 1) &= 2x\left[\pi - 4x + x^2\right] && 0 \leqslant x \leqslant 1 \\ &= 2x\left[\pi - 2 - x^2 + 4\left((x^2 - 1)^{\frac{1}{2}} - a\cos\frac{1}{x}\right)\right] && 1 \leqslant x \leqslant \sqrt{2} \end{aligned} \tag{5.57}$$

$$\begin{aligned} p(d \leqslant x,\ 1) &= \pi x^2 - \frac{8x^3}{3} + \frac{x^4}{2} && x \leqslant 1 \\ &= \frac{1}{3} + (\pi-2)x^2 + 4(x^2-1)^{\frac{1}{2}} + \frac{8}{3}(x^2-1)^{\frac{3}{2}} \\ &\quad - \frac{x^4}{2} - 4x^2\, a\cos\frac{1}{x} && 1 < x \leqslant \sqrt{2} \end{aligned} \tag{5.58}$$

These results will be used in chapter ( VI ) in connection with the $d^2$ test for random number generators. We could also examine distributions in higher dimensions. For instance, let us compute the pdf of 2 random points in a sphere of radius R . We proceed as before

$$p_1(x,\ R) = \frac{2\pi x^2(1 - \cos\theta)}{\frac{4}{3}\pi R^3}$$

where

$$\theta = a\cos\frac{x}{2R}$$

Using Crofton's theorem we obtain the differential equation

$$\frac{dp}{d\theta} + 9p \tan\theta = \frac{108}{x}(1 - \cos\theta)\cos^2\theta \sin\theta$$

which has the solution

$$p(x, R) = \frac{9}{4}\frac{x^2}{R^3} - \frac{27}{20}\frac{x^3}{R^4} + \frac{9}{1280}\frac{x^8}{R^9}$$

By integration we get the cumulative distribution in terms of $\alpha = \frac{x}{R}$

$$p(d \leqslant x, R) = \frac{3}{4}\alpha^3 - \frac{27}{80}\alpha^4 + \frac{1}{1280}\alpha^9 \qquad 0 \leqslant \alpha \leqslant 2 \quad (5.59)$$

Let us now examine in detail how the number of clusters can be estimated for a random distribution of n discs over a circle. Later, we shall perform a similar computation in the case of spheres.

We consider 3 types of clusters containing 1, 2 and 3 points, respectively. We shall assume that the diameter $\mathcal{E}$ of the discs is such that $\mathcal{E} << 1$ and neglect the abnormal behavior close to the circumference in case ii).

i) the expected number of isolated points is

$$n_1 = n\left(1 - \mathcal{E}^2 + \frac{4}{3\pi}\mathcal{E}^3 - \frac{\mathcal{E}^5}{30\pi} + \mathcal{O}(\mathcal{E}^7)\right)^{n-1} \qquad (5.60)$$

the expression in parenthesis representing the probability that the distance x between 2 points be greater than $\mathcal{E}$. From (5.56), we obtain

$$P(x \geq \varepsilon, 1) = \frac{2}{\pi}(1-\varepsilon^2)\,\mathrm{acos}\,\frac{\varepsilon}{2} + \frac{\varepsilon}{\pi}\left(1 + \frac{\varepsilon^2}{2}\right)\sqrt{1 - \frac{\varepsilon^2}{4}}$$

$$= \frac{2}{\pi}(1-\varepsilon^2)\left(\frac{\pi}{2} - \frac{\varepsilon}{2} - \frac{\varepsilon^3}{48} - \frac{3\varepsilon^5}{1280}\right) + \frac{\varepsilon}{\pi}\left(1+\frac{\varepsilon^2}{2}\right)\left(1 - \frac{\varepsilon^2}{8} - \frac{\varepsilon^4}{128}\right) + \mathcal{O}(\varepsilon^7)$$

$$= 1 - \varepsilon^2 + \frac{4}{3\pi}\varepsilon^3 - \frac{\varepsilon^5}{30\pi} + \mathcal{O}(\varepsilon^7) \qquad (5.61)$$

Here, the terms of order higher than $\varepsilon^2$ represent the influence of the boundary since the relative remaining area after removing a circle of radius $\varepsilon$ is $1 - \varepsilon^2$.

ii) it is most convenient to consider all remaining cases as subcases of the event: two points $O_i$ and $O_j$ fall at a distance $x$ less than $\varepsilon$. Designate by $D_i$ and $D_j$ two circles with centers $O_i$ and $O_j$, respectively, and radius $\varepsilon$. We distinguish between three regions

$$D - (D_i \cup D_j)\,,\ (D_i \cup D_j) - (D_i \cap D_j) \text{ and } (D_i \cap D_j)$$

assuming that $(D_i \cup D_j) \subseteq D$. This assumption is false if $O_i$ or $O_j$ fall within a distance $\varepsilon$ of the circumference of $D$. This circular ring represents a fraction $2\varepsilon - \varepsilon^2$ of the domain $D$ and consequently, the next calculation is expected to have a relative accuracy $\mathcal{O}(\varepsilon)$.

If all of the other $(n-2)$ points fall into the first region, we get a 2 point cluster. However, if one point ends up in the second region we obtain a 3 point linear cluster and finally if it falls into the third one, we get a 3 point cyclic cluster. Let us now compute the conditional probabilities that these last two events occur. We need the area of overlap of $D_i$ and $D_j$ which is

$$m(D_i \cap D_j) = 2\varepsilon^2 \operatorname{acos}\frac{x}{2\varepsilon} - \frac{x}{2}\sqrt{4\varepsilon^2 - x^2} \tag{5.62}$$

and by the foregoing remark

$$m(D_i \cap D_j \cap D) = \left(2\varepsilon^2 \operatorname{acos}\frac{x}{2\varepsilon} - \frac{x}{2}\sqrt{4\varepsilon^2 - x^2}\right)\left(1 + \mathcal{O}(\varepsilon)\right) \tag{5.63}$$

If $p_i$ designate the probability that some other point falls into the $i^{th}$ region, we can write

$$p_2 = (n-2)\int_0^{\varepsilon} p(x,1)\left[2\pi\varepsilon^2 - 4\varepsilon^2 \operatorname{acos}\frac{x}{2\varepsilon} + x\sqrt{4\varepsilon^2 - x^2} + \mathcal{O}(\varepsilon^3)\right] dx$$

$$p_3 = (n-2)\int_0^{\varepsilon} p(x,1)\left[2\varepsilon^2 \operatorname{acos}\frac{x}{2\varepsilon} - \frac{x}{2}\sqrt{4\varepsilon^2 - x^2} + \mathcal{O}(\varepsilon^3)\right] dx$$

and under the assumption that clusters with more than 3 points are ignored we have

$$p_1 = P(x \leq \varepsilon, 1) - p_2 - p_3$$

Those expressions are now evaluated, replacing $p(x,1)$ by (5.55) to give

$$p_3 = (n-2)\int_o^{\varepsilon} \frac{4x}{\pi}\left[\operatorname{acos}\frac{x}{2} - \frac{x}{2}\sqrt{1-\frac{x^2}{4}}\right]\left[2\varepsilon^2 \operatorname{acos}\frac{x}{2\varepsilon} - \frac{x}{2}\sqrt{4\varepsilon^2 - x^2} + \mathcal{O}(\varepsilon^3)\right]dx$$

$$= (n-2)\int_o^{\varepsilon}\left[2x - \frac{4}{\pi}x^2 + \mathcal{O}(x^4)\right]\left[2\varepsilon^2 \operatorname{acos}\frac{x}{2\varepsilon} - \frac{x}{2}\sqrt{4\varepsilon^2 - x^2} + \mathcal{O}(\varepsilon^3)\right]dx$$

$$= (n-2)\left\{\int_o^{\varepsilon}\left[4\varepsilon^2 x \operatorname{acos}\frac{x}{2\varepsilon} - x^2\sqrt{4\varepsilon^2 - x^2}\right]dx + \mathcal{O}(\varepsilon^5)\right\}$$

$$= (n-2)\,\varepsilon^4\left[\pi - \frac{3\sqrt{3}}{4} + \mathcal{O}(\varepsilon)\right] \qquad (5.64)$$

Similarly

$$p_2 = (n-2)\;2\pi\varepsilon^2\;P(d \leq \varepsilon,\;1) - 2p_3$$

$$= (n-2)\left[2\pi\varepsilon^2\left(\varepsilon^2 + \mathcal{O}(\varepsilon^3)\right) - 2\varepsilon^4\left(\pi - \frac{3\sqrt{3}}{4}\right)\right]$$

$$= (n-2)\;\varepsilon^4\left[\frac{3\sqrt{3}}{2} + \mathcal{O}(\varepsilon)\right] \qquad (5.65)$$

and if we group these results, the probability $p_1$ that 2 discs overlap without being part of a 3-point cluster is found to be

$$p_1 = \varepsilon^2 - \frac{4}{3\pi}\varepsilon^3 - \varepsilon^4(n-2)\left(\pi + \frac{3\sqrt{3}}{4} + \mathcal{O}(\varepsilon)\right) \qquad (5.66)$$

The expected number of classes c can now be estimated using all of these particular counts. There are $\binom{n}{2}$ distinct pairs of points but in the counting process a linear 3-point cluster will be counted twice and a 3-point cyclic cluster three times. Therefore, we may write

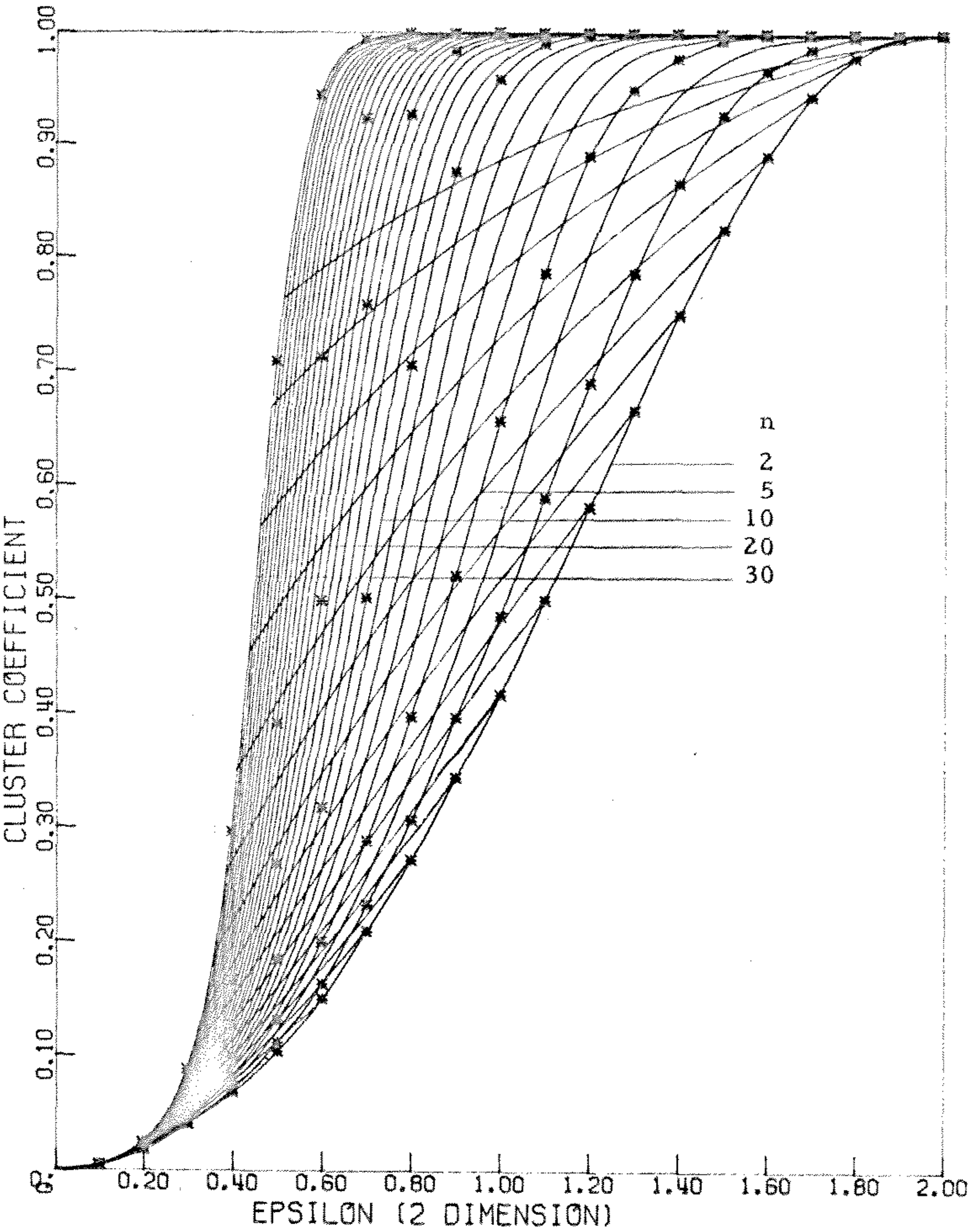

CLUSTER COEFFICIENT
EPSILON (2 DIMENSION)
0.
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
n
2
5
10
20
30

Figure 5.3

$$c = n\left[1-\varepsilon^2+\frac{4}{3\pi}\varepsilon^3+\mathcal{O}(\varepsilon^5)\right]^{n-1}+\binom{n}{2}\Big\{\varepsilon^2-\frac{4}{3\pi}\varepsilon^3+$$

$$\varepsilon^4(n-2)\left[-\left(\pi+\frac{3\sqrt{3}}{4}\right)+\frac{1}{2}\left(\frac{3\sqrt{3}}{2}\right)+\frac{1}{3}\left(\pi-\frac{3\sqrt{3}}{4}\right)+\mathcal{O}(\varepsilon)\right]\Big\}$$

$$= n\left[1-(n-1)\left(\varepsilon^2-\frac{4}{3\pi}\varepsilon^3\right)+\binom{n-1}{2}\varepsilon^4\right]+\binom{n}{2}\left[\varepsilon^2-\frac{4}{3\pi}\varepsilon^3+\varepsilon^4(n-2)\left(-\frac{2\pi}{3}-\frac{\sqrt{3}}{4}\right)\right]+\mathcal{O}(n^3\varepsilon^5)$$

$$= n-\binom{n}{2}\varepsilon^2+\binom{n}{2}\frac{4}{3\pi}\varepsilon^3-\binom{n}{3}\left(2\pi+\frac{3\sqrt{3}}{4}-3\right)\varepsilon^4+\mathcal{O}(n^3\varepsilon^5) \qquad (5.67)$$

An interesting comparison is to perform an identical computation in 3-dimensional space to determine the variation of the number of clusters of spheres when their density is small. We find successively

i) $$n_i = n\left(1-\frac{3}{4}\varepsilon^3+\frac{27}{80}\varepsilon^4+\frac{1}{1280}\varepsilon^9\right)^{n-1}$$

ii) $$m(D_i \cap D_j \cap D) = \frac{2\pi}{3}\varepsilon^2 x\left(1+\mathcal{O}(\varepsilon)\right)$$

$$p_3 = (n-2)\left\{\int_0^{\varepsilon}\left(\frac{9}{4}x^2-\frac{27}{20}x^3+\frac{9}{1280}x^8\right)\frac{2\pi}{3}\varepsilon^2 x\,dx+\mathcal{O}(\varepsilon^7)\right\}$$

$$= (n-2)\left[\frac{3\pi}{8}\varepsilon^6+\mathcal{O}(\varepsilon^7)\right]$$

$$p_2 = (n-2)\left[\frac{5\pi}{4}\varepsilon^6+\mathcal{O}(\varepsilon^7)\right]$$

$$p_1 = \frac{3}{4}\varepsilon^3-\frac{27}{80}\varepsilon^4-(n-2)\left[\frac{13\pi}{8}\varepsilon^6+\mathcal{O}(\varepsilon^7)\right]$$

so that the expected value of the number of classes is now

$$c = n-\binom{n}{2}\frac{3}{4}\varepsilon^3+\binom{n}{2}\frac{27}{80}\varepsilon^4-\binom{n}{3}3\left(\frac{9}{16}-\frac{7\pi}{8}\right)\varepsilon^6+\mathcal{O}(n^3\varepsilon^7)$$

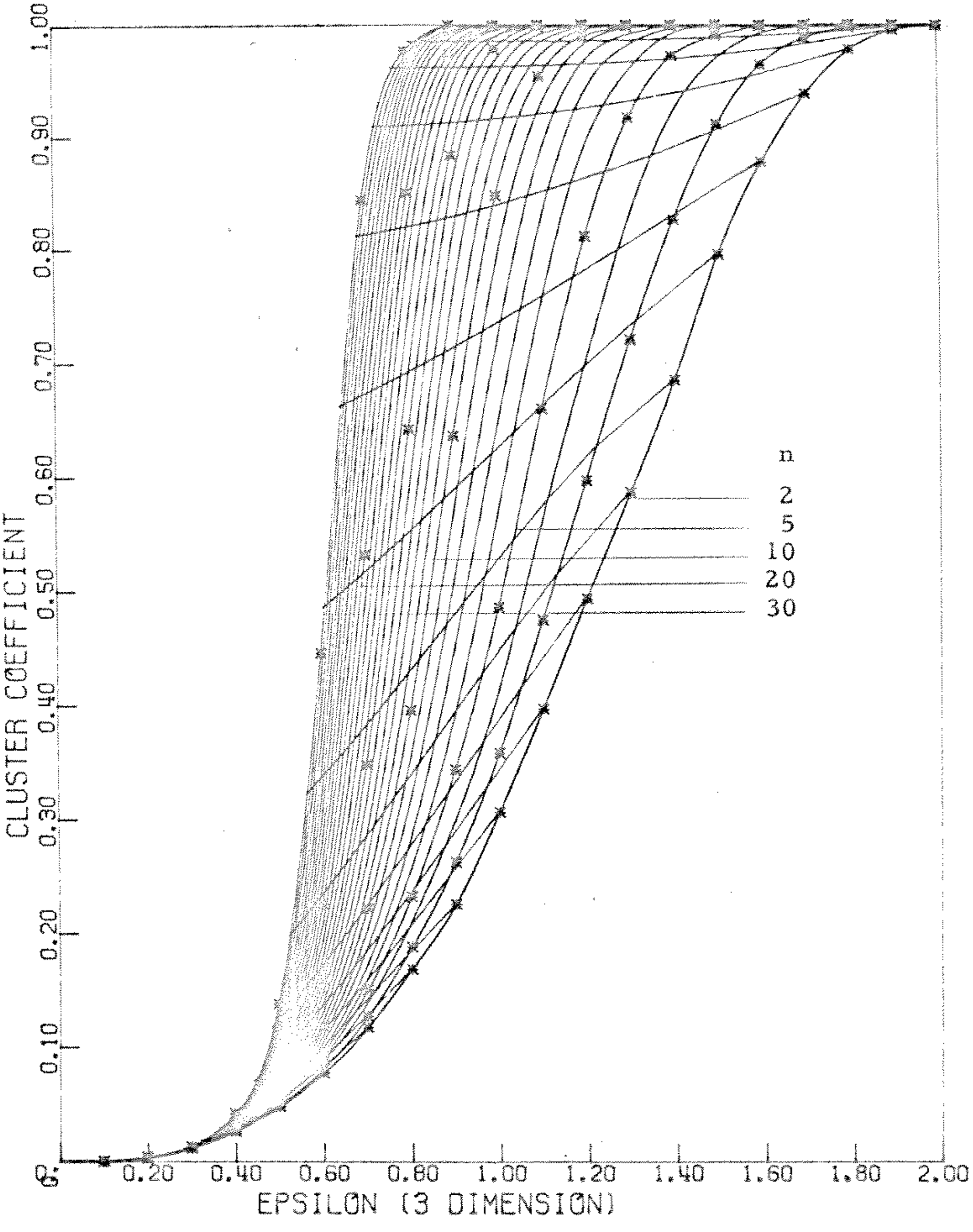

CLUSTER COEFFICIENT
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.
0.
0.20
0.40
0.60
0.80
1.00
1.20
1.40
1.60
1.80
2.00
EPSILON (3 DIMENSION)
n
2
5
10
20
30

Figure 5.4

# CHAPTER VI

## Computational Tools

In the preceding chapters, many numerical results were obtained either by recursive calculations or Monte Carlo sampling. In view of the importance of these results, caution had to be exercised throughout the course of these computations. As a consequence, several methods were developed in order to ensure efficiency and accuracy. This chapter describes three particular areas of investigation. First, common sampling methods such as sampling with and without replacement are compared and are shown to be equivalent by means of suitable transformations for the purpose of class counting. Then random storage assignment techniques used for the storage and retrieval of large families of graphs are examined and significantly improved. Finally random number generators used in the previous experimentation are described, their properties compared and a method for producing reliable pseudo-random sequences is presented.

## 6.1. Relationship Between Various Sampling Methods

Most of the graph problems that we are concerned with require generating subgraphs, the vertices of which are selected with equal probability from the vertex set of some complete graph. Selecting such a sample is equivalent to generating a random arrangement of k integers from the set $\{1, 2, \ldots, n\}$. Details on how to perform this operation efficiently $\left(\mathcal{O}(k)\right.$ operations$\left.\right)$ can be found in Reference [ 8 ].

However, when the number of vertices of the source graph increases, the requirement that the k selected vertices be distinct gradually loses its importance as the probability of finding a match decreases. This is very fortunate indeed, since the labor needed to impose that constraint increases like k itself. Depending upon whether all elements in the sample are distinct or not, we get the classical sampling without or with replacement, respectively.

In the case of large samples, the approach we adopt is to sample with replacement (which is obviously the easiest method), then perform a transformation to recover, if need be, the results that would have been obtained had we computed every time the precise number of distinct elements in the sample or selected samples of distinct elements.

Let the population $\mathcal{P}_N$ contain N distinct elements. Our experiment is the uniform selection of samples of size k designated by

$S_k$ when sampling with replacement,

$s_k$ when sampling without replacement.

For each $S_k$, let $s_{k_1}$ designate the subset of $S_k$ having $k_1$ distinct elements $1 \leq k_1 \leq k$ such that $k_1$ is maximum. Finally let $f$ be a function which is defined for every sample $\mathscr{S}$ of $\mathscr{P}$.

Theorem - *For all functions* $f$ *defined on samples of* $\mathscr{P}_N$ for which $f(S_k) = f(s_{k_1})$, $0 \leq k \leq N$, then

$$\mathscr{E}\left[f(s_k)\right] = \sum_{\nu=1}^{k} N^{\nu-1} \frac{(N-k)!}{(N-1)!} \left[ {k \atop \nu} \right] \mathscr{E}\left[f(S_\nu)\right] \qquad (6.1)$$

$\left[ {k \atop \nu} \right]$ being the Stirling number of the first kind.

proof: given k-tuples $\left\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\right\}$ such that $x_{i_j} \in \left\{x_1, x_2, \ldots, x_N\right\}$, $j = 1, 2, \ldots, k$, the probability of finding $\nu$ distinct elements among the $k$ selected is

$$p(N, k, \nu) = N^{-k} N(N-1)\ldots(N-\nu+1) \left\{ {k \atop \nu} \right\} \quad \nu=1, 2, \ldots, k \qquad (6.2)$$

since this is a coupon collector's problem with $N$ equally probable distinct coupons. In this formula $\left\{ {k \atop \nu} \right\}$ designated Stirling numbers of the second kind, i.e., the number of ways of partitioning a set of $k$ elements into $\nu$ non-empty subsets. The expected value of the function $f$ evaluated over fixed size samples depends only upon the size of the sample; for convenience let us write

$$\mathcal{E}\left[f(s_k)\right] = g(k)$$

$$\mathcal{E}\left[f(S_k)\right] = G(k)$$

which are related, using ( 6.2 ), in the following way

$$G(k) = \sum_{\nu=1}^{\min(k,N)} p(N,\ k,\ \nu)\ g(\nu)$$

$$= (N-1)!\ N^{1-k} \sum_{\nu=1}^{\min(k,N)} \frac{\begin{Bmatrix} k \\ \nu \end{Bmatrix} g(\nu)}{(N-\nu)!} \qquad (6.3)$$

If we now look at $g(k)$ as the $k^{th}$ component of a vector $\vec{g}$ and similarly for $G(k)$ we obtain the system

$$\vec{G} = \mathcal{C}\ \vec{g}$$

where the kxk lower triangular matrix $\mathcal{C}$ has the form

$$\mathcal{C} = \begin{pmatrix} 1 & & & 0 \\ \frac{1}{N} & \frac{N-1}{N} & & \\ \frac{1}{N^2} & \frac{3(N-1)}{N^2} & \frac{(N-1)(N-2)}{N^2} & \\ & \cdots & \frac{(N-1)!}{N^{i-1}(N-j)!}\begin{Bmatrix} i \\ j \end{Bmatrix} & \end{pmatrix} \qquad (6.4)$$

As $\mathcal{C}$ is lower triangular, it is not difficult to write down its inverse by inspection, namely

$$\mathcal{C}^{-1} = \begin{pmatrix} 1 & & & & \\ -\frac{1}{N-1} & \frac{N}{N-1} & & & 0 \\ \frac{2}{(N-1)(N-2)} & \frac{-3N}{(N-1)(N-2)} & \frac{N^2}{(N-1)(N-2)} & & \\ & \cdots & & \frac{N^{j-1}(N-i)!}{(N-1)!}\begin{bmatrix} i \\ j \end{bmatrix} & \end{pmatrix} \qquad (6.5)$$

We simply have to verify that indeed, the product $\mathcal{C}\mathcal{C}^{-1}$ produces the identity matrix

$$(\mathcal{C}\mathcal{C}^{-1})_{ij} = \sum_{\nu=1}^{k} \frac{(N-1)!}{N^{i-1}(N-\nu)!}\left\{\begin{matrix} i \\ \nu \end{matrix}\right\}\frac{N^{j-1}(N-\nu)!}{(N-1)!}\begin{bmatrix} \nu \\ j \end{bmatrix}$$

$$= \sum_{\nu=1}^{k} N^{j-i}\left\{\begin{matrix} i \\ \nu \end{matrix}\right\}\begin{bmatrix} \nu \\ j \end{bmatrix}$$

$$= \sum_{\nu=j}^{i} N^{j-i}\left\{\begin{matrix} i \\ \nu \end{matrix}\right\}\begin{bmatrix} \nu \\ j \end{bmatrix} = \delta_{ij}$$

It is important to stress that both transformation matrices being lower triangular, the computation of $g_k$ requires only values of $G_\nu$ up to $\nu=k$. It should be clear by now that this approach will prove advantageous for all sampling problems where the function $f$ is insensitive to the presence of duplicate elements in the sample.

For large N and sample sizes satisfying $k \ll N$ we might even operate with a straightforward sampling with replacement and never apply transformation $\mathcal{T}^{-1}$. We derive an estimate for the error.

Lemma

$$g(k) - G(k) = \frac{k(k-1)}{2N}\Big(G(k) + G(k-1)\Big) + \mathcal{O}\left(\frac{1}{N^2}\right) \quad \text{if } k \ll N \tag{6.6}$$

proof: using expression (6.3) for G(k) we write

$$g(k) - G(k) = \sum_{\nu=1}^{k-1} N^{\nu-1}\frac{(N-k)!}{(N-1)!}\begin{bmatrix} k \\ \nu \end{bmatrix} G(\nu) + \left[N^{k-1}\frac{(N-k)!}{(N-1)!} - 1\right] G(k)$$

the coefficient of G(k) now becomes

$$N^{k-1}\frac{(N-k)!}{(N-1)!} - 1 = \frac{1}{(1-\frac{1}{N})(1-\frac{2}{N})\dots(1-\frac{k-1}{N})} - 1$$

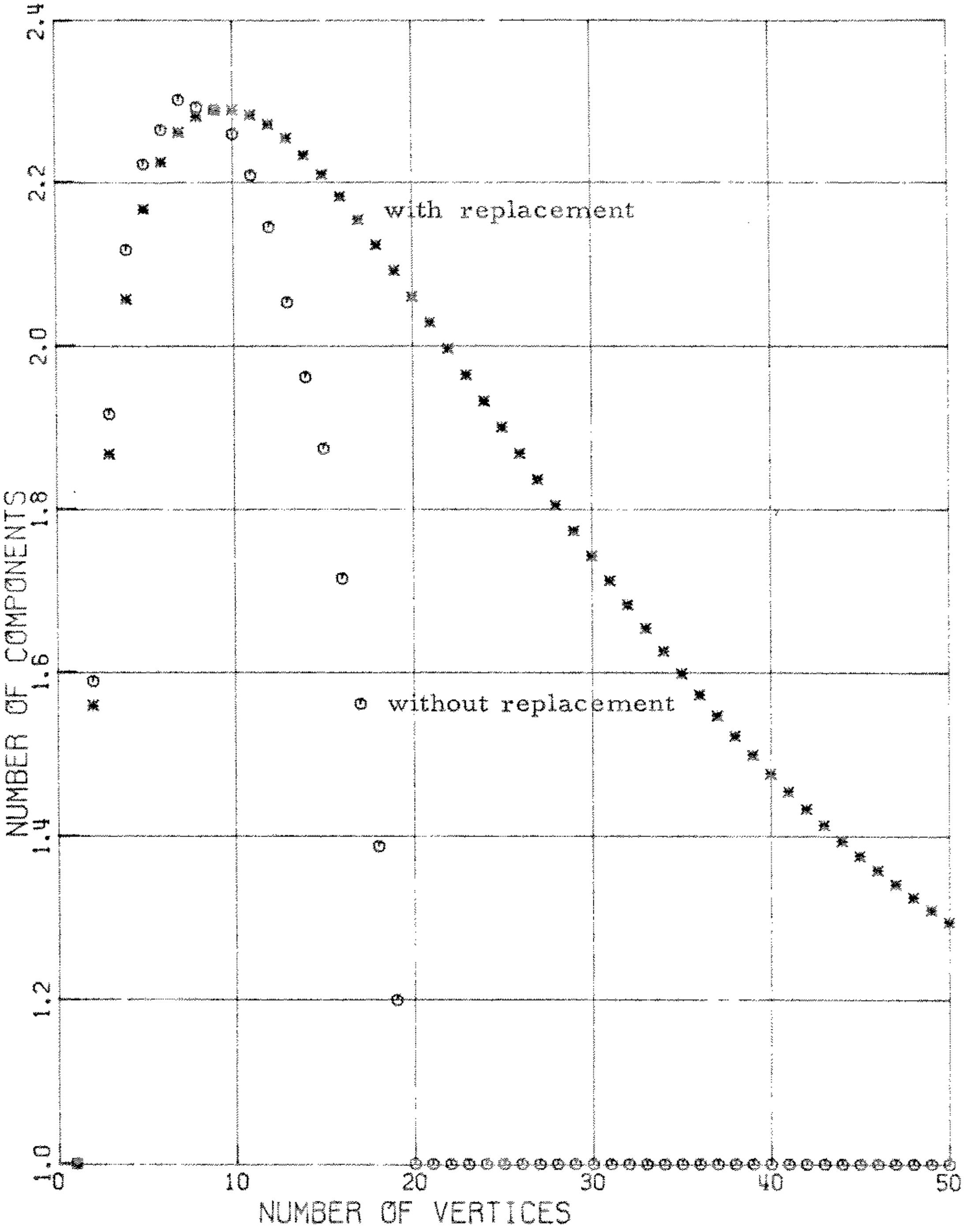
$$= \frac{1}{1 - \frac{k(k-1)}{2N} + \mathcal{O}(\frac{1}{N^2})} - 1$$

$$= \frac{k(k-1)}{2N} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{6.7}$$

which yields

$$g(k) - G(k) = \frac{k(k-1)}{2N} G(k) + \frac{\begin{bmatrix} k \\ k-1 \end{bmatrix} G(k-1)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

$$= \frac{k(k-1)}{2N}\Big(G(k) + G(k-1)\Big) + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{6.8}$$



NUMBER OF COMPONENTS
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
NUMBER OF VERTICES
0
10
20
30
40
50
with replacement
without replacement

Figure 6. 1

## 6.2. Random Storage Assignment

In section (4.7), a large number of encoded graphs, together with their associated information, were stored and retrieved using a random storage method which will now be described in some detail, since it is an improvement over existing "hash" algorithms. Because we do restrict our comparison to linear and random probing, the reader is referred to [21] and [22] for other methods of search.

Hash algorithms are primarily distinguished by the way in which they handle collisions. Elements from a set $S = \{s_1, s_2, \ldots, s_N\}$ can be mapped into a table T with n available positions $\{t_o, t_1, \ldots, t_{n-1}\}$, each $t_i$ being able to accommodate a single element of $\mathcal{P}$. Since the mapping function $\mathcal{F}$ is in general many to 1, several elements drawn from S may be initially assigned the same t position. If $t_{j_1}$ for $j_1 = \mathcal{F}(s_i)$ is already occupied, we have collision and some vacant t position elsewhere in the table must be selected. Thus, any particular hash algorithm provides a way of computing from $s_i$ a sequence $t_{j_1}, t_{j_2}, \ldots, t_{j_\nu}$ satisfying

$$j_1 = \mathcal{F}(s_i)$$

$$t_{j_1}, t_{j_2}, \ldots, t_{j_{\nu-1}} \quad \text{occupied}$$

$$t_{j_\nu} \quad \text{vacant}$$

$j_1, j_2, \ldots, j_\nu$ forming some permutation of $\nu$ distinct integers $\in \left[ 0, n-1 \right]$. The element $s_i$ is subsequently assigned to slot $t_{j_\nu}$ and can be retrieved in an identical fashion, provided that neither $t_{j_1}, t_{j_2}, \ldots$ nor $t_{j_{\nu-1}}$ have been changed to vacant in the meantime.

Such an algorithm will be optimal storagewise if given any distribution of table occupancy, the probability of assigning the next item to any of the still vacant slots is equal; optimality here means that the expected value of the length $\nu$ of the probing sequence $j_1, j_2, \ldots, j_\nu$ is minimized.

For example, the worst strategy corresponds to

$$j_\nu = j_1 + (\nu - 1) \quad \mod n \qquad \nu = 2, 3, \ldots, n$$

because if a collision has occurred at $t_{j_1}$, the probability that a collision will occur at $t_{j_1} + 1$ is higher than the average over all t's. Linear probing is therefore replaced by random probing, whereby a fixed permutation $\left\{ \ell_1, \ell_2, \ldots, \ell_{n-1} \right\}$ of $\left\{ 1, 2, \ldots, n-1 \right\}$ is used to form the probing sequence

$$j_\nu = j_1 + \ell_{\nu-1} \quad \mod n \qquad \nu = 2, 3, \ldots, n \tag{6.9}$$

The probability that the (k+1)st item entered into the table will require $\nu$ probes is

$$\left( 1 - \frac{k}{n} \right) \prod_{j=0}^{\nu-2} \left( \frac{k-j}{n-1-j} \right)$$

We can easily verify by induction that

$$1 + \frac{k}{n-1} + \frac{k(k-1)}{(n-1)(n-2)} + \ldots + \frac{k!}{(n-1)(n-2)\ldots(n-k)} = \frac{n}{n-k}$$

Thus, the expected value of $\nu$ when k items are already in the table is

$$\mathcal{E}(\nu_{k+1}) = \left(1 - \frac{k}{n}\right)\left[1 + \sum_{\nu=1}^{k} \nu \prod_{j=0}^{\nu-1} \left(\frac{k-j}{n-1-j}\right)\right]$$

$$= 1 + \sum_{\nu=1}^{k} \prod_{j=0}^{\nu-1} \left(\frac{k-j}{n-j}\right) \qquad (6.10)$$

the second form being based directly on the probability that the probing sequence is of length greater than or equal to $\nu$. We rewrite

$$\mathcal{E}(\nu_{k+1}) = \frac{1}{1 - \frac{k}{n+1}} \qquad (6.11)$$

so that if $\alpha = \frac{k}{n}$ is the occupancy factor,

$$\mathcal{E}(\nu_{k+1}) \leqslant \frac{1}{1 - \alpha} \qquad (6.12)$$

The expected value of $\nu$ when retrieving an item from the table can then be approximated by

$$\mathcal{E}(\nu) \leqslant \frac{n}{k} \int_{o}^{\alpha} \frac{dx}{1 - x} = -\frac{1}{\alpha} \log (1 - \alpha) \qquad (6.13)$$

A few values of $\mathcal{E}(\nu)$ are

| $\alpha$ | $\xi(\nu)$ | $\xi(\nu_{k+1})$ |
|---|---|---|
| .1 | 1.06 | 1.11 |
| .25 | 1.16 | 1.33 |
| .5 | 1.39 | 2.00 |
| .75 | 1.85 | 4.00 |
| .9 | 2.56 | 10.00 |
| .95 | 3.16 | 20.00 |

However, the foregoing computation makes the tacit assumption that the probability of occupancy of any table position is the same, which is likely to be false if the mapping function $\mathcal{F}$ does not satisfy that property for all samples of the set $S$. Indeed, $\mathcal{F}$ is usually chosen on intuitive grounds, hoping that it will distribute the elements of $S$ uniformly over the table, since the distribution of the data in the s-space may not even be known. Consequently, the probability of collision for the $(k+1)$st element is bound to be higher than $\alpha$ and the above algorithm will not perform as well as expected.

Alternatively, consider now the following algorithm. The table $T$ has positions $\{t_o, t_1, \ldots, t_{n-1}\}$ but $n$ is constrained to be prime. Compute:

$$\begin{aligned} j_1 &= \mathcal{F}_1(s) && \ni \quad 0 \leq j_1 \leq n-1 \\ \ell &= \mathcal{F}_2(s) + 1 && \ni \quad 1 \leq \ell \leq n-1 \end{aligned} \tag{6.14}$$

then probe at

$$j_\nu = j_1 + (\nu - 1)\ell \mod n \qquad \nu = 1, 2, \ldots, n \tag{6.15}$$

until either a vacant slot or the data item is encountered. If $\mathcal{F}_1$ and $\mathcal{F}_2$ are chosen in such a way that $j_1$ and $\ell$ are uncorrelated for all

$s \epsilon S$ , then the previous estimate of the length of the probing sequence will be valid. The probability that any two distinct elements from S have identical hashing sequences $j_1, j_2, \ldots, j_\nu$ is now $\mathcal{O}(\frac{1}{n^2})$ instead of $\mathcal{O}(\frac{1}{n})$. The requirement that n be prime ensures that no matter what value $\ell$ has, all table positions will have been visited after n probes.

$\mathcal{F}_1$ and $\mathcal{F}_2$ can be chosen as follows: consider each data item (or some transform of the data item, for instance only a fractional representation of it) as an integer x. Using the Chinese remainder theorem in its simplest form, we know that any integer m in the range $[0, \; n(n-1) - 1]$ can be uniquely represented by the pair of remainders $r_1$ , $r_2$ where

$$\begin{aligned} r_1 &= m \mod n \\ r_2 &= m \mod n-1 \end{aligned} \qquad (6.16)$$

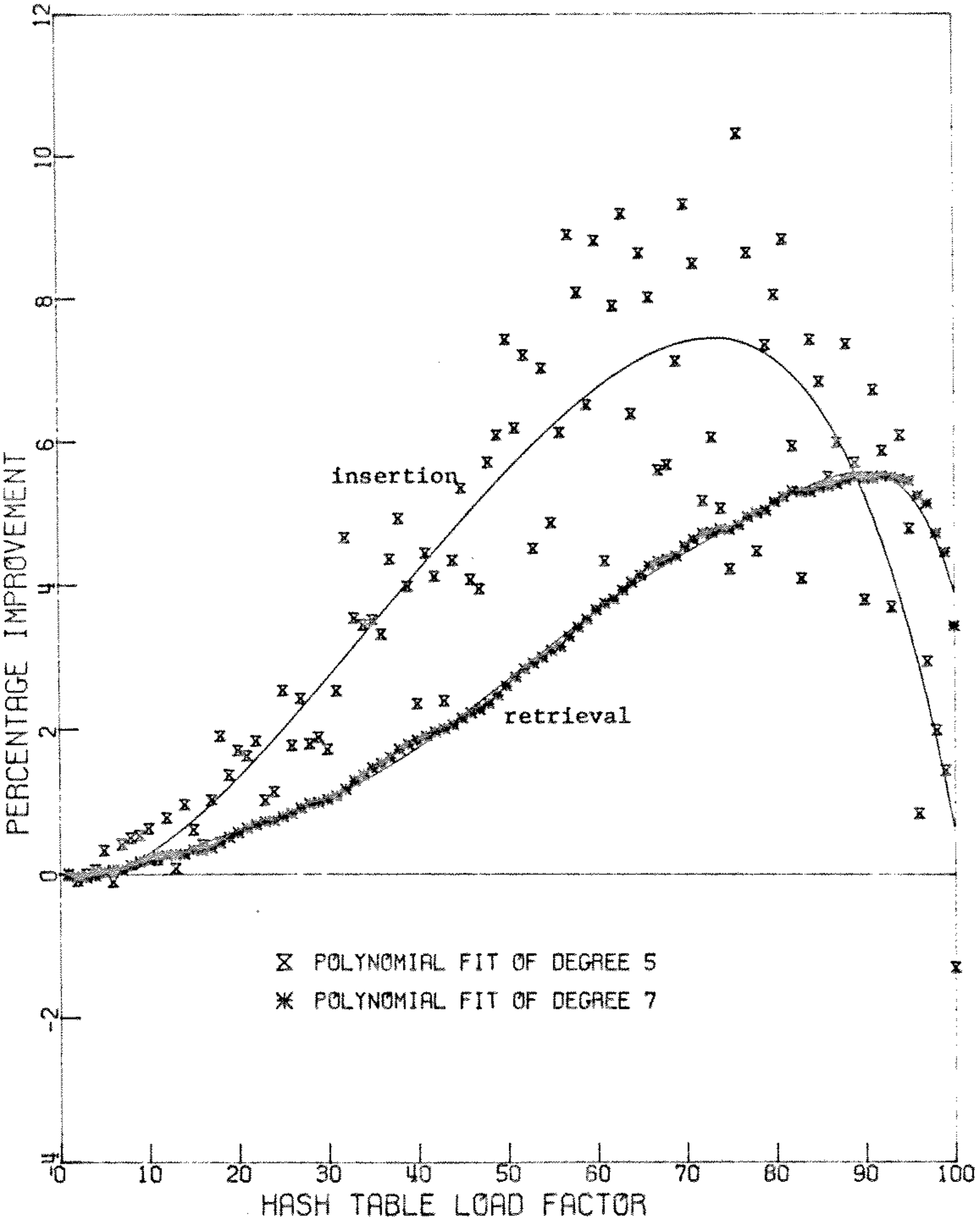
Therefore, in the present case we can simply use

$$\begin{aligned} \mathcal{F}_1 &: \quad j_1 = x \mod n \\ \mathcal{F}_2 &: \quad \ell = (x \bmod n-1) + 1 \end{aligned}$$

Clearly, for all n(n-1) distinct pairs $(j_1, \ell)$ to be feasible, the range of x should be at least equal to n(n-1). We can thus recommend to apply to the data item a transformation which will distribute x over as large a range as possible; then its remainder modulo n(n-1) should be approximately uniformly distributed. This is, of course, the typical approach to "randomization" as implemented by linear congruential random number generators.

In practice, one need not compute each time $j_1$ and $\ell$ ; most of the time, for as long as the load factor is moderate, one probe will suffice to store or retrieve an item so that only $j_1$ need be calculated. Still, an objection may be raised concerning table sizes which are primes rather than powers of two. Usually, there are two motivations for choosing $n = 2^k$ ; tables can be simply combined or broken to form similar tables and operations modulo n are easily performed by just masking the high order bits. Let us remark, however, that once a table size has been chosen to implement a hash algorithm, it cannot in general be altered. Table extensions are usually achieved by performing multiple searches; if the item is not found in the first table, a second table is searched and so forth, but this is of course significantly worse than having a unique table set up in the first place. As far as modulo n operations are concerned, the objection disappears if the algorithm is implemented in a higher level language such as FORTRAN where an honest division is actually carried out to obtain the remainder. At any rate, the cost of division should prove advantageous over the time required to generate the successive permuted increments in the classical random probing scheme as the table starts to fill up.

Finally, let us mention briefly how deletions are handled. In order to indicate which table entries are either vacant or deleted, we use two special codes which are not members of S . Although it is commonly said that lost space is reclaimed but lookup time not reduced, still some lookup time can be eliminated. When an item is retrieved from $t_{j\nu}$ using the probing sequence $t_{j_1}, t_{j_2}, \ldots, t_{j_\nu}$ , it



PERCENTAGE IMPROVEMENT
-4
-2
0
2
4
6
8
10
12
0
10
20
30
40
50
60
70
80
90
100
HASH TABLE LOAD FACTOR
insertion
retrieval
⧗ POLYNOMIAL FIT OF DEGREE 5
✳ POLYNOMIAL FIT OF DEGREE 7

Figure 6.2

may also be moved to $t_{j_\mu}$ , $1 \leq \mu < \nu$, which is the first deleted entry, if any, found in the probing sequence. If this compacting operation takes place, $t_{j_\nu}$ is subsequently changed to deleted. Thus, if $k_1$ items were removed from a table and assuming that all remaining entries have a non-zero probability of being accessed, the expected value of the length $\nu$ of the probing sequence would gradually decrease from $\mathcal{E}(\nu_k)$ to $\mathcal{E}(\nu_{k-k_1})$

$$\lim_{N \to \infty} \mathcal{E}_N(\nu) = \mathcal{E}(\nu_{k-k_1})$$

N being the number of retrievals made after the deletions occurred. Since compacting a table involves a slight additional effort, it should only be performed if any entry is accessed on the average more than twice.

Both random probing algorithms were tested. With $n = 997$ and a uniform distribution of $j_1$ between 0 and $n-1$ , figure (6.2 ) shows the percentage improvement in the length of the probing sequence to install or retrieve an item when the load factor is $\alpha$ . This percentage is computed as a percentage of $\nu$ for the classical random probing algorithm.

## 6.3. Random Number Generation

In this section we put in their proper perspective several methods used to generate uniformly distributed random numbers. The mere fact that there exists several, not to say many methods, is a clear indication that we do not have a best one in some absolute sense. On the contrary, the situation is that of a typical engineering trade-off, cost versus quality. The cost will be measured by the time required to produce a sequence of random numbers and the quality by how well the sequence passes some carefully chosen tests of randomness. The complexity of the problem is further compounded by our inability to establish computable necessary and sufficient randomness criteria valid for all situations. No wonder then, that commonly used methods are plethoric even if we exclude most "random methods", the most imaginative but probably the most deceptive ones.

There are several properties of infinite, random sequences of independent samples drawn from the uniform distribution that the finite deterministic sequence we construct ought to have. Those necessary conditions are that the sequence be equipartitioned, equidistributed and white, notions which are studied at length by Franklin [ 12 ] . Under those constraints, there are two possible candidates:

i) sequences $\{\theta^n\}$ for $\theta$ transcendental

ii) multiply sequence $x_{n+1} = \{a x_n + b\}$

The former is characterized by an inherent difficulty of generation, since $\theta^n$ has to be calculated without rounding, but has good statistical properties; the latter is intrinsically mediocre since it only has the required properties asymptotically, but is easy to generate with a period of the order of the maximum integer representable in a computer word, depending upon the specific choice of multiplier a . This multiplier can in fact be chosen a priori to yield a pseudo random sequence with predictable statistical behavior over the whole period, as was shown by Coveyou and Macpherson.

Unfortunately, typical Monte Carlo calculations demand more numbers than we can afford to generate by the first method, but would require only a very minute fraction of the period of a typical multiply-sequence.

k-product pseudo random sequence

A method for producing pseudo random sequences by combining other deterministic sequences is now described. Although specifically treated in the case of binary sequences, extension to an r-ary number representation is feasible.

Consider k sequences of uncorrelated stationary random variables $x_{ij}$ taking only the values ± 1 with probability

$$\left.\begin{aligned} p(x_{i\nu} = 1) &= p_i \\ p(x_{i\nu} = -1) &= q_i = 1 - p_i \end{aligned}\right\} \begin{aligned} i &= 1, \ldots, k \\ \nu &= 0, 1, 2, \ldots \end{aligned} \qquad (6.17)$$

and define as their k-product sequence, the sequence

$$z_\nu = \prod_{i=1}^{k} x_{i\nu} \qquad (6.18)$$

Theorem - A k-product sequence $z_\nu = \prod_{i=1}^{k} x_{i\nu}$ has even moments unity and odd (2n+1) moments in absolute value less than $\varepsilon^{k(2n+1)}$

$$\mathcal{E} = \max_{i \in \{1,2,\ldots,k\}} \left| 2p_i - 1 \right| \tag{6.19}$$

proof: the n$^{\text{th}}$ moment of the random variable $x_i$ can be written

$$\mathcal{M}_n(x_i) = j^{-n}\left[\frac{d^n}{d\omega^n}\left(p_i e^{j\omega} + q_i e^{-j\omega}\right)\right]_{\omega=0}$$

$$= j^{-n}\left[\frac{d^n}{d\omega^n}\left(\cos\omega + j\sin\omega(p_i-q_i)\right)\right]_{\omega=0} \tag{6.20}$$

Similarly for $z$

$$\mathcal{M}_n(z) = j^{-n}\left[\frac{d^n}{d\omega^n}\left(\cos\omega\prod_{i=1}^{k}(p_i+q_i) + j\sin\omega\prod_{i=1}^{k}(p_i-q_i)\right)\right]_{\omega=0}$$

$$= j^{-n}\left[\frac{d^n}{d\omega^n}\left(\cos\omega + j\sin\omega\prod_{i=1}^{k}(2p_i-1)\right)\right]_{\omega=0} \tag{6.21}$$

which becomes after separating even and odd cases

$$\begin{cases} \mathcal{M}_n(z) = 1 & \text{if } n \text{ even} \\ \mathcal{M}_n(z) = \prod_{i=1}^{k}(2p_i-1)^n & \text{if } n \text{ odd} \end{cases} \tag{6.22}$$

<u>Theorem</u> - The autocorrelation $R_z(\tau)$ of a k-product sequence is bounded by

$$\left| R_z(\tau) \right| \leqslant r^k \tag{6.23}$$

where

$$r = \max_{i\in\{1,\ldots,k\}} \quad \max_{\tau \text{integer} \neq 0} \left| R_{x_i}(\tau) \right| \tag{6.24}$$

proof: the autocorrelation of $x_i$ is

$$R_{x_i}(\tau) = \mathcal{E}\left\{ x_{ij}\, x_{ij+\tau} \right\} \qquad \tau \text{ integer}$$

For a truly white sequence we have

$$R_{x_i}(\tau) = \delta(\tau)$$

Here, the autocorrelation of the k-product sequence is similarly

$$R_z(\tau) = \mathcal{E}\left\{ z_j\, z_{j+\tau} \right\} = \mathcal{E}\left\{ \prod_{i=1}^{k} x_{ij} x_{ij+\tau} \right\} \tag{6.25}$$

but because the sequences $x_i$ have been assumed to be uncorrelated, (6.25) becomes

$$R_z(\tau) \leqslant \prod_{i=1}^{k} R_i(\tau) \tag{6.26}$$

However, whiteness is not a strong criterion of randomness so that we now examine the equipartition properties of these sequences.

<u>Theorem</u> - k-product sequences of independent ± sequences are equipartitioned and completely equidistributed asymptotically as $k \rightarrow \infty$.

proof: given $\lambda$ numbers $x_{i1}\; x_{i2} \ldots x_{i\lambda}$ taking discrete values $\pm 1$, they form $2^{\lambda}$ distinct configurations $C_j$, $j \in \{1, 2, \ldots, 2^{\lambda}\}$. A sequence $x_i$ is equipartioned by $\lambda$ if

$$p\left(\left\{x_{in}x_{in+1}\cdots x_{in+\lambda-1}\right\}\equiv C_\nu\right)=2^{-\lambda} \tag{6.27}$$

Let $p_{i\nu}=2^{-\lambda}+\varepsilon_{i\nu}$ designate the actual probability that $\left\{x_{in}\,x_{in+1}\cdots x_{in+\lambda-1}\right\}\equiv C_\nu$ in the $i^{th}$ sequence. We then form the product sequence $z=x_ix_j$ so that

$$p\left(\left\{z_nz_{n+1}\cdots z_{n+\lambda-1}\right\}\equiv C_\nu\right)=\sum_{\mu=1}^{2^\lambda} p_{i\mu}p_{j\theta(\mu)} \tag{6.28}$$

$\theta(\mu)$ being the image of $\mu$ under some permutation $\theta$ of the integers $\left\{1,\,2,\,\ldots,\,2^\lambda\right\}$. Using expression (6.28) we get

$$P_r\left(\left\{z_nz_{n+1}\cdots z_{n+\lambda-1}\right\}\equiv C_\nu\right)=2^{-\lambda}\left(1+\sum_{\mu=1}^{2^\lambda}\varepsilon_{i\mu}\varepsilon_{j\theta(\mu)}\right)$$

$$\leqslant 2^{-\lambda}+\varepsilon^2 \tag{6.29}$$

where $$\mathcal{E}=\max_{\mu\in\left\{1,2,\ldots,2^\lambda\right\}}\left[\varepsilon_{i\mu},\varepsilon_{j\mu}\right] \tag{6.30}$$

Thus if the constituent sequences are equipartitioned $\mathcal{O}(\mathcal{E})$, the k-product sequence will be equipartitioned $\mathcal{O}(\mathcal{E}^k)$.

In the particular case of the discrete $\pm$ 1 sequence, equidistribution by $\lambda$ for the $\lambda$-dimensional sequence $\left\{x_{in},x_{in+1},\ldots,x_{in+\lambda-1}\right\}$ is implied by equipartition and holds for every $\lambda$.

## Generation of k-product Pseudo Random Sequences

Implementation of the k-product operation on a binary computer with word length $\ell$ can be easily achieved by performing k-product operations on each one of the $\ell$ bits. For that purpose, the exclusive OR of two 0-1 sequences $s_1$ and $s_2$ corresponds identically to the 2-product of the ± 1 sequence $s_1$ with the complement of $s_2$, or

$$s_1(0,1) \oplus s_2(0,1) \Leftrightarrow s_1(1, -1) \times s_2(-1,1)$$

Since the exclusive OR operates on all $\ell$ bits in parallel and is associative and commutative, the k-product sequence is obtained as the outcome of $(k-1)$ $\ell$-bits exclusive OR's.

Of course, the necessity of obtaining k independent constituent sequences introduces a factor $\frac{1}{k}$ in the overall speed of the algorithm so that we must justify the use of a k-product generator. Several areas of important applications are:

i) extension of $\{\theta^n\}$ sequences: we recall from [12] this very good but quite costly method of producing pseudo random sequences with all desired properties of randomness. Rather than keep stubbornly generating $\{\theta^n\}$ for increasing n (one of course might think of starting the sequence over with different transcendental $\theta$), a particular sequence, say 20,000 numbers, can be stored permanently and used in conjunction with a multiply sequence to form a 2-product sequence which will be as good as $\{\theta^n\}$ but with a period at least equal to that of the linear congruential generator, typically of the order of $2^\ell$.

ii) generation of multiply sequences with homogeneous properties over all bits: a multiply sequence $x_{n+1} = ax_n + b \bmod m$ can be analyzed by means of the spectral test to determine its expected accuracy over the whole period; in this sense we mean that k-tuples of only the $s_k$ most significant bits of adjacent values can be considered essentially independent. Typically if the accuracy is 16 bits for pairs, it will be, say, 10 bits for triples, probably less for quadruples and quintuples may not even be independent. This can be observed quite directly by computing bit serial and cross correlations; in particular, we have subsequently compared the data obtained for $\{\pi^n\}$ and a good multiply sequence as indicated by the spectral test

$$x_{n+1} = 273673163155_8 \, x_n + \text{cst}$$

The degradation of serial correlation for lags up to 15 is quite characteristic as we move from the most to the least significant bit. The distribution of the serial correlation coefficients should be normal with mean 0 and variance N (number of samples) as a consequence of the De-Moivre-Laplace theorem on the limiting form of the binomial distribution. Next, we picked 3 distinct multipliers and combined their multiply sequences applying the transformation

$$b'_i = \sum_{\nu=1}^{k} b_{\left\{i+(\nu-1)\left\lfloor \frac{\ell}{k} \right\rfloor\right\} \bmod \ell \;+1,\nu} \qquad \text{mod } 1 \tag{6.31}$$

for $k = 3$ ; here $b_{i,\nu}$ designates the $i^{th}$ bit of the random integer just obtained from the $\nu^{th}$ sequence, $b'_i$ is the value of the $i^{th}$ bit in the 3-product sequence. This transformation simply performs a circular bit permutation equal to $\left\lfloor \frac{\ell}{3} \right\rfloor$ bits for the $2^{nd}$ sequence and $2\left\lfloor \frac{\ell}{3} \right\rfloor$ bits for the third. Several tests were subsequently performed:

- bit auto and cross correlations, bit serial correlations
- frequency, poker and coupon collector's tests
- distance of 2 random points in a square ($d^2$ test)

It is interesting to notice that the 3-product sequence performed equally well as the sequence $\{\pi^n\}$. We emphasize that the study of k-product sequences made earlier assumes all along that even though the constituent sequences may not be good pseudo random sequences, they are nevertheless independent.

Results from these tests are given in appendix.

# CHAPTER VII

## Conclusions

We have developed efficient computational methods to analyze complex problems of partitioning. Examples were given to illustrate how asymptotic expansions can be used whenever possible to relieve the actual computing task. The analysis of the general graph problem has given us some insight into the behavior of the expected number of components when sampling from an arbitrary space.

Even though we only considered uniform distributions, it is interesting to notice that all results obtained are also valid in situations where the probability of selection of any element is not equal. Indeed, from the graph standpoint, each vertex can be replaced by a p-clique whose vertices are adjacent to the same subset of vertices as before. The number $p$ is proportional to the probability of selecting that particular vertex and $p$ is finite only if these probabilities are commensurate. We can now perform a sampling with replacement of the new graph and apply the necessary transformation to recover the result of sampling without replacement. Therefore, the expected number of components will have the same pattern of variation as before.

There are of course many related questions which have been uncovered during this study. They should be the subject of further endeavor.

For instance, the method of sampling from the space of graphs with constrained local degrees should lead to some interesting estimation problems when sampling from an infinite population.

## APPENDIX

Frequency Test for Random Integers Between 1 and 10
(Chi Square with 9 Degrees of Freedom

| | Samples | Mean | Stand. Dev. | Chi Square |
|---|---|---|---|---|
| Observed | 4000 | 0.500 | 0.290 | 4.265 |
| | 4000 | 0.508 | 0.291 | 9.350 |
| | 4000 | 0.507 | 0.290 | 8.645 |
| | 4000 | 0.508 | 0.291 | 7.210 |
| | 4000 | 0.491 | 0.289 | 10.925 |
| Expected | 4000 | 0.500 | 0.289 | 9. |
| Cumulated | | | | |
| Observed | 20000 | 0.503 | 0.290 | 8.443 |
| Expected | 20000 | 0.500 | 0.289 | 9. |

Chi Square Sampling Distribution

| | | |
|---|---|---|
| Observed | mean = 8.08 | variance = 6.33 |
| Expected | mean = 9. | variance = 18. |

| | Samples | Mean | Stand. Dev. | Chi Square |
|---|---|---|---|---|
| Observed | 4000 | .503 | .290 | 3.845 |
| | 4000 | .503 | .284 | 8.960 |
| | 4000 | .504 | .291 | 17.755 |
| | 4000 | .505 | .291 | 5.925 |
| | 4000 | .501 | .290 | 7.345 |
| Expected | 4000 | .500 | .289 | |
| Cumulated | | | | |
| Observed | | .503 | .289 | 8.181 |
| Expected | | .500 | .289 | 9. |

Chi Square Sampling Distribution

| | | |
|---|---|---|
| Observed | mean = 8.77 | variance = 28.8 |
| Expected | mean = 9. | variance = 18. |

Poker Test for Random Integers Between 1 and 10
(Chi Square with 5 Degrees of Freedom)

| | Samples | Chi Square |
|---|---|---|
| Observed | 4000 | 3.753 |
| | 4000 | 7.527 |
| | 4000 | 3.698 |
| | 4000 | 14.218 |
| | 4000 | 6.617 |
| Expected | 4000 | 5. |
| Cumulated | | |
| Observed | 20000 | 13.479 |
| Expected | 20000 | 5. |
| Chi Square Sampling Distribution | | |
| Observed | mean = 7.16 | variance = 18.46 |
| Expected | mean = 5. | variance = 10. |
| | Samples | Chi Square |
| Observed | 4000 | 3.019 |
| | 4000 | 4.637 |
| | 4000 | 2.963 |
| | 4000 | 5.360 |
| | 4000 | 11.314 |
| Expected | 4000 | 5. |
| Cumulated | | |
| Observed | 20000 | 1.289 |
| Expected | 20000 | 5. |
| Chi Square Sampling Distribution | | |
| Observed | mean = 5.46 | variance = 11.79 |
| Expected | mean = 5. | variance = 10. |

$D^2$ Test for Random Points in a Square
(Chi Square with 6 Degrees of Freedom)

| | Samples | Mean | Stand. Dev. | Chi Square |
|---|---|---|---|---|
| Observed | 4000 | 0.512 | 0.249 | 4.914 |
| | 4000 | 0.536 | 0.255 | 5.995 |
| | 4000 | 0.517 | 0.251 | 1.048 |
| | 4000 | 0.540 | 0.243 | 10.730 |
| | 4000 | 0.522 | 0.253 | 2.249 |
| Expected | 4000 | 0.521 | | 6. |
| Cumulated | | | | |
| Observed | 20000 | 0.526 | 0.250 | 6.715 |
| Expected | 20000 | 0.521 | | 6. |

Chi Square Sampling Distribution

| | | |
|---|---|---|
| Observed | mean = 4.99 | variance = 14.25 |
| Expected | mean = 6. | variance = 12. |

| | Samples | Mean | Standard Dev. | Chi Square |
|---|---|---|---|---|
| Observed | 4000 | 0.523 | 0.245 | 1.692 |
| | 4000 | 0.519 | 0.237 | 6.257 |
| | 4000 | 0.529 | 0.257 | 3.457 |
| | 4000 | 0.527 | 0.247 | 6.545 |
| | 4000 | 0.537 | 0.245 | 7.902 |
| Expected | 4000 | 0.521 | | 6. |
| Cumulated | | | | |
| Observed | 20000 | 0.527 | 0.246 | 3.509 |
| Expected | 20000 | 0.521 | | 6. |

Chi Square Sampling Distribution

| | | |
|---|---|---|
| Observed | mean = 5.17 | variance = 6.39 |
| Expected | mean = 6. | variance = 12. |

Coupon Collector's Test for Random Integers Between 1 and 10
(Chi Square with 8 Degrees of Freedom)

| | Samples | Mean | Stand. Dev. | Chi Square |
|---|---|---|---|---|
| Observed | 3967 | 30.053 | 12.097 | 9.268 |
| | 3982 | 27.846 | 8.770 | 11.246 |
| | 3995 | 29.593 | 11.678 | 4.885 |
| | 3983 | 29.073 | 12.443 | 10.158 |
| | 3992 | 28.312 | 11.100 | 9.488 |
| Expected | | 29.290 | 11.211 | 8. |
| Cumulated | | | | |
| Observed | 19919 | 28.975 | 11.293 | 12.039 |
| Expected | | 29.290 | 11.211 | 8. |

Chi Square Sampling Distribution

Observed mean = 9.01 variance = 5.91

Expected mean = 8. variance = 16.

| | Samples | Mean | Stand. Dev. | Chi Square |
|---|---|---|---|---|
| Observed | 3984 | 29.511 | 10.581 | 6.163 |
| | 3987 | 30.907 | 11.029 | 11.972 |
| | 3983 | 29.287 | 10.612 | 9.919 |
| | 3988 | 30.212 | 12.178 | 5.448 |
| | 3990 | 28.633 | 10.311 | 9.387 |
| Expected | | 29.290 | 11.211 | 8. |
| Cumulated | | | | |
| Observed | 19922 | 29.710 | 10.962 | 4.438 |
| Expected | | 29.290 | 11.211 | 8. |

Chi Square Sampling Distribution

Observed mean = 8.58 variance = 7.40

Expected mean = 8. variance = 16.

# BIBLIOGRAPHY

[1] Beckenbach, Edwin F.: Applied Combinatorial Mathematics, Wiley, New York, 1964.

[2] Berg, W. F.: "Aggregates in One- and Two-Dimensional Random Distribution", Philosophical Magazine (Series 7), Vol. 36 (1945), pp. 337-346.

[3] Berge, Claude: The Theory of Graphs and Its Applications, Methuen, London, 1962.

[4] Blackett, Donald W.: Elementary Topology, A Combinatorial and Algebraic Approach, Wiley, New York, 1967.

[5] Blanc-Lapierre, A. and Fortet, R.: Theory of Random Functions, Gordon and Breach, New York, 1965.

[6] Bloemena, A. R.: Sampling from a Graph, Mathematical Centre Tracts, Amsterdam, 1964.

[7] Caine, Stephen H.: "Citran User's Guide", California Institute of Technology, Computing Center, Nov. 1966.

[8] de Balbine, Guy: "Note on Random Permutations", Mathematics of Computation, Vol. 21, No. 100, Oct. 1967, pp. 710-712.

[9] Domb, C.: "The Problem of Random Intervals on a Line", Proc. Cam. Phil. Soc., Vol. 43, 1947, pp. 329-341.

[10] Erdélyi, A.: Asymptotic Expansions, Dover, 1956.

[11] Flegg, Graham H.: Boolean Algebra and Its Applications, Wiley, New York, 1964.

[12] Franklin, Joel N.: "Deterministic Simulation of Random Processes", Mathematics of Computation, Vol. 17, No. 81, Jan. 1963, pp. 28-59.

[13] Hadwiger, Hugo and Debrunner, Hans: Combinatorial Geometry in the Plane, Holt, Rinehart and Winston, New York, 1964.

[14] Hall, Marshall, Jr.: Combinatorial Theory, Waltham: Blaisdell, 1967.

[15] Harary, Frank: A Seminar in Graph Theory, Holt, Rinehart and Winston, New York, 1967.

[16] Jansson, Birger: Random Number Generators, Victor Petterson's Bokindustri Aktiebolag, Stockholm, 1966.

[17] Kendall, M. G. and Moran, P. A. P.: Geometrical Probability, Hafner, New York, 1963.

[18] Knuth, Donald E.: The Art of Computer Programming, Vol. 1, Fundamental Algorithms, Addison-Wesley, Reading, 1968.

[19] Lomont, J. S.: Applications of Finite Groups, Academic Press, New York, 1959.

[20] Mack, C.: "The Expected Number of Aggregates in a Random Distribution of Points", Proc. Cam. Phil. Soc., Vol. 46, 1949, pp. 285-292.

[21] Maurer, W. D.: "An Improved Hash Code for Scatter Storage", Communications of the ACM, Vol. 11, No. 1, Jan. 1968, pp. 35-38.

[22] Morris, Robert: "Scatter Storage Techniques", Communications of the ACM, Vol. 11, No. 1, Jan. 1968, pp. 38-44.

[23] Ore, Oystein: Theory of Graphs , American Mathematical Society Colloquium Publications, 1962.

[24] Ore, Oystein: The Four-Color Problem, Academic Press, New York, 1967.

[25] Riordan, John: An Introduction to Combinatorial Analysis, Wiley, New York, 1958.

[26] Rogers, C. A.: Packing and Covering, University Press, Cambridge, 1964.

[27] Ryser, Herbert J.: Combinatorial Mathematics. The Mathematical Association of America, Wiley, New York, 1963.

[28] Tutte, W. T.: Connectivity in Graphs, University of Toronto Press, 1966.

[29] Varga, Richard S.: Matrix Interative Analysis, Prentice Hall, 1962.

[30] : Proceedings of the IBM Scientific Computing Symposium on Combinatorial Problems, White Plains, New York, IBM, 1964.