A FORMULA FOR THE PROBABILITY OF OBTAINING A TREE FROM A GRAPH CONSTRUCTED RANDOMLY EXCEPT FOR AN "EXOGAMOUS BIAS"[1]

By Hwa Sung Na and Anatol Rapoport

University of Michigan

1. Introduction. A general problem in the probabilistic theory of linear graphs can be stated as follows:

Given a randomly constructed linear graph $G(n, N)$ with n nodes and N links and a property of linear graphs A, what is the probability that $G(n, N)$ will have the property A as a function of n and N?

The phrase "randomly constructed" needs to be more precisely specified, for example, by describing the process of construction. One such process consists of selecting from the $\binom{n}{2}$ pairs of nodes a random sample of $N \leqq \binom{n}{2}$ pairs to be connected by links. Accordingly, the probability of having property A will then be defined as the ratio of the number of distinct labelled graphs with n nodes and N links which have this property, to the total number of such graphs, namely $C(n, N) = \binom{\binom{n}{2}}{N}$.

In particular, if A is the property of being a connected graph, it was shown by Erdös and Rényi (1960) that if

$$N = (\tfrac{1}{2})n \log_e n + an + o(n), \tag{1}$$

then, as n and N approach infinity, the probability that the randomly constructed graph is connected approaches

$$P(A) = \exp\{-e^{-2a}\}. \tag{2}$$

In other words, given N and n, both sufficiently large and connected by equation (1), the probability that $G(n, N)$ is connected is approximately

$$\exp\{-ne^{-2N/n}\}. \tag{3}$$

Many situations can be represented as linear graphs, for example, acquaintance nets in which the nodes are people and a link represents the relation of being acquainted; word association nets, where the nodes are words and a link represents the property of being associated in some sense (syntactic, semantic, etc.). One can imagine such graphs being generated by a stochastic process of some sort. However, it is clearly improbable that in this process links are formed entirely at random. Biases can certainly be expected to influence the probabilities of con-

Received 4 February 1966; revised 18 August 1966.

[1] The research work on which this paper is based was supported in part by the Office of Naval Research under Contract Nonr 1224(46), and in part by the National Science Foundation Grant GS-1027, Mental Health Research Institute, University of Michigan.

nections. Thus, in an acquaintance net, if nodes A and B are joined by a link and also nodes B and C, we can expect that a link will join A and C with probability greater than what it would be if the other connections had not taken place. Therefore biased graphs become objects of interest.

A particular type of bias is the distance bias which can be defined for a set of nodes in a metric space. For example, the probability that two nodes close together are connected by a link can be supposed to be greater than if the nodes were far apart.

A special case of the distance bias results if the set of nodes consists of a number of subsets and if the probability that two nodes are linked is greater or smaller if both nodes belong to the same subset than otherwise. If the number of connections within the subsets is greater than that expected by chance, we have an "endogamous bias." In the opposite case, we have an "exogamous bias".

In this paper we investigate the case where $N = n - 1$. Consequently the graph $G(n, n - 1)$ is connected if and only if G is a tree. We shall investigate the probability that G is connected if the set of nodes consists of subsets and the $n - 1$ links are apportioned into some which connect the nodes within the subsets and others which connect nodes from different subsets.

2. The probability of obtaining a tree with prescribed partitions of nodes and intra-links. Assume that n labelled points are given and that these are divided into k subsets, and n_i points in subset i $(i = 1, 2, \cdots, k)$, such that

$$\sum_{i=1}^{k} n_i = n. \tag{4}$$

Pairs of points are then connected by a number of links to form graphs. Links joining points of the same subset will be called *intra-links*, those connecting points of different groups will be called *inter-links*. Let l_i denote the number of intra-links among the nodes of subset i, and l_0 denote the total number of inter-links among the k subsets.

We are here specially interested in the following problem. Suppose each set of the l_i links $(i = 0, 1, \cdots, k)$ are placed randomly with equal probability; that is, within the subset i each selection of the l_i pairs of nodes is equally probable and similarly for the set of l_0 inter-links. Then, what is the probability that the resulting graph is a tree?

To answer this question, we determine how many different labeled trees can be obtained by the above described procedure. To do this, we shall use the determinant method of H. M. Trent (1954) and L. Weinberg (1958).

Consider the completely connected graph $G(\mathbf{n}, \mathbf{N})$ with $N_i = \binom{n_i}{2}$ intra-links among the nodes of the ith subset $(i = 1, 2, \cdots k)$ and N_0 inter-links among the k subsets. Our notation $\mathbf{n}$ specifies the vector $(n_1, n_2, \cdots, n_k)$. Since the completely connected graph has $\binom{n}{2}$ links, we have

$$N_0 = \binom{n}{2} - \sum_{i=1}^{k} \binom{n_i}{2}. \tag{5}$$

We shall label each of the N_i intra-links of the ith subset by the same indeterminate symbol x_i, which will serve as a marker in computing the determinant defined below. Each of the N_0 inter-links will be correspondingly labelled x_0.

Let M stand for a matrix of degree $(n - 1)$ with submatrices A_{ij}. Each matrix A_{ij} shall have $n_i - \delta_{ik}$ rows and $n_j - \delta_{jk}$ columns, where

$$(6) \qquad \delta_{ik} = 0, \quad \text{if} \quad i \neq k, \qquad = 1, \quad \text{if} \quad i = k.$$

Thus

$$(7) \qquad M = \begin{bmatrix} A_{11} & \cdots & A_{1k} \\ & \cdots & \\ A_{k1} & \cdots & A_{kk} \end{bmatrix}.$$

The entries of A_{ij} will be represented by $t_{pq}(ij)$ defined as follows:

$$(8) \qquad t_{pp}(ii) = (n_i - 1)x_i + (n - n_i)x_o; \qquad p \,\varepsilon\, [1, n_i - \delta_{ik}];$$

$$(9) \qquad t_{pq}(ii) = -x_i \quad (\text{for} \quad p \neq q), \qquad p, q \,\varepsilon\, [1, n_i - \delta_{ik}];$$

$$(10) \qquad t_{pq}(ij) = -x_0 \quad (\text{for} \quad i \neq j), \qquad p \,\varepsilon\, 1, n_i - \delta_{ik}]; \quad q \,\varepsilon\, [1, n_j - \delta_{jk}].$$

From results obtained by H. M. Trent (1954) and Lindsey Perkins (unpublished), it follows that the terms of the determinant $|M|$ represent all the trees contained in G. In particular, the coefficient of the term $x_0^{l_0}x_1^{l_1} \cdots x_k^{l_k}$ will give the number of trees with exactly l_i intra-links in the ith subset and l_0 inter-links among the k subsets. In other words, $|M|$ will be the generating function of the number of trees with prescribed partitions $\mathbf{n} \equiv (n_1, n_2, \cdots, n_k)$, and $\mathbf{l} \equiv (l_0, l_1, \cdots, l_k)$. Denote the coefficient of $x_0^{l_0}x_1^{l_1} \cdots x_k^{l_k}$ by $T(\mathbf{n}, \mathbf{l})$.

THEOREM 1. *In the expression of the determinant of M, the coefficient of the term $x_0^{l_0}x_1^{l_1} \cdots x_k^{l_k}$ will be given by*

$$(11) \qquad T(\mathbf{n}, \mathbf{l}) = n^{k-2}\prod_{i=1}^{k} [\tbinom{n_i-1}{l_i} n_i^{l_i}(n - n_i)^{n_i - l_i - 1}].$$

PROOF. Define $R_\nu = \sum_{s=1}^{\nu-1} n_s$ for $\nu \,\varepsilon\, [2, k]$; $R_1 = 0$. Consider four elementary square matrices of degree $n - 1$, namely:

$$(12) \qquad T_\tau, \text{ with entries } \tau_{ii} = 1, \quad \text{for} \quad i \,\varepsilon\, [1, n - 1]; \quad \tau_{ij} = 1, \quad \text{for} \quad i = 1, j \,\varepsilon\, [2, n - 1]; \quad \tau_{ij} = 0 \quad \text{otherwise};$$

T_τ', the transpose of T_τ;

$$T_\lambda \text{, with entries } \lambda_{ii} = 1, \quad \text{for} \quad i \varepsilon [1, n-1]$$

$$\lambda_{ij} = 1, \quad \text{for} \begin{Bmatrix} i = R_\nu + 1 \\ j \varepsilon [R_\nu + 2, R_\nu + n_\nu] \end{Bmatrix} \tag{13}$$

$$\text{with } \nu \varepsilon [2, k-1] \text{ for } k \geqq 3$$

$$\lambda_{ij} = 0 \quad \text{otherwise;}$$

$$T_\rho \text{, with entries } \rho_{ii} = 1, \quad \text{for} \quad i \varepsilon [1, n-1]$$

$$\rho_{ij} = -1, \quad \text{for} \begin{Bmatrix} i = R_\nu + 1 \\ j \varepsilon [R_\nu + 2, R_\nu + n_\nu] \end{Bmatrix} \tag{14}$$

$$\text{with } \nu \varepsilon [1, k-1]$$

$$\rho_{ij} = 0 \quad \text{otherwise.}$$

Next, consider the matrix

$$M_1 = T_\lambda T_\tau' T_\tau M T_\rho . \tag{15}$$

Since the determinants of T_τ, T_τ', T_λ, and T_ρ are all unity, we have

$$|M_1| = |M|. \tag{16}$$

Furthermore, the transformation of M by T_τ, T_τ', T_λ, and T_ρ is such that the determinant of M_1 is simply the product of its diagonal elements (the transformation matrices were chosen to insure this). Thus the determinant of $|M|$ turns out to be

$$\begin{aligned} |M| &= n^{k-2} x_0^{k-1} \textstyle\prod_{i=1}^{k} [n_i x_i + (n - n_i)x_0]^{n_i - 1} \\ &= n^{k-2} x_0^{k-1} \textstyle\prod_{i=1}^{k} \{ \sum_{l_i=0}^{n_i-1} \binom{n_i-1}{l_i} (n_i x_i)^{l_i} [(n - n_i)x_0]^{n_i - l_i - 1} \} \\ &= \textstyle\sum_{l_1=0}^{n_1-1} \sum_{l_2=0}^{n_2-1} \cdots \sum_{l_k=0}^{n_k-1} \{ n^{k-2} \prod_{i=1}^{k} \binom{n_i-1}{l_i} n_i^{l_i} (n - n_i)^{n_i - l_i - 1} \} \\ &\qquad \cdot x_0^{l_0} x_1^{l_1} \cdots x_k^{l_k} \\ &= \textstyle\sum_{l_1=0}^{n_1-1} \sum_{l_2=0}^{n_2-1} \cdots \sum_{l_k=0}^{n_k-1} T(\mathbf{n}, 1) \cdot x_0^{l_0} x_1^{l_1} \cdots x_k^{l_k}, \end{aligned} \tag{17}$$

where

$$T(\mathbf{n}, 1) = n^{k-2} \textstyle\prod_{i=1}^{k} \binom{n_i-1}{l_i} n_i^{l_i} (n - n_i)^{n_i - l_i - 1}. \tag{18}$$

By way of illustration, let

$$M = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix}, \tag{19}$$

where $n = 9$; $k = 3$; $n_1 = 4$; $n_2 = 3$; $n_3 = 2$. Consequently, according to our definitions and according to (8), (9), and (10), we shall have

(20)
$$A_{11} = \begin{bmatrix} 3x_1 + 5x_0 & -x_1 & -x_1 & -x_1 \\ -x_1 & 3x_1 + 5x_0 & -x_1 & -x_1 \\ -x_1 & -x_1 & 3x_1 + 5x_0 & -x_1 \\ -x_1 & -x_1 & -x_1 & 3x_1 + 5x_0 \end{bmatrix};$$

$$A_{12} = \begin{bmatrix} -x_0 & -x_0 & -x_0 \\ -x_0 & -x_0 & -x_0 \\ -x_0 & -x_0 & -x_0 \\ -x_0 & -x_0 & -x_0 \end{bmatrix}; \quad A_{13} = \begin{bmatrix} -x_0 \\ -x_0 \\ -x_0 \\ -x_0 \end{bmatrix};$$

$$A_{21} = \begin{bmatrix} -x_0 & -x_0 & -x_0 & -x_0 \\ -x_0 & -x_0 & -x_0 & -x_0 \\ -x_0 & -x_0 & -x_0 & -x_0 \end{bmatrix};$$

$$A_{22} = \begin{bmatrix} 2x_2 + 6x_0 & -x_2 & -x_2 \\ -x_2 & 2x_2 + 6x_0 & -x_2 \\ -x_2 & -x_2 & 2x_2 + 6x_0 \end{bmatrix}; \quad A_{23} = \begin{bmatrix} -x_0 \\ -x_0 \\ -x_0 \end{bmatrix};$$

$$A_{31} = [-x_0 \quad -x_0 \quad -x_0 \quad -x_0]; \quad A_{32} = [-x_0 \quad -x_0 \quad -x_0];$$

$$A_{33} = [x_3 + 7x_0].$$

The T matrices corresponding to M illustrated by (19) and (20) are shown below:

(21)
$$T_\tau = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

(22)
$$T_\lambda = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix};$$

$$(23) \qquad T_p = \begin{bmatrix} 1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

We shall then have

$$(24) \; M_1 = \begin{bmatrix} x_0 & 0 & 0 & 0 & x_0 & 0 & 0 & x_3 \\ x_0-x_1 & 4x_1+5x_0 & 0 & 0 & 0 & 0 & 0 & x_3-x_0 \\ x_0-x_1 & 0 & 4x_1+5x_0 & 0 & 0 & 0 & 0 & x_3-x_0 \\ x_0-x_1 & 0 & 0 & 4x_1+5x_0 & 0 & 0 & 0 & x_3-x_0 \\ 0 & 0 & 0 & 0 & 9x_0 & 0 & 0 & 3x_3-3x_0 \\ 0 & 0 & 0 & 0 & x_0-x_2 & 3x_2+6x_0 & 0 & x_3-x_0 \\ 0 & 0 & 0 & 0 & x_0-x_2 & 0 & 3x_2+6x_0 & x_3-x_0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2x_3+7x_0 \end{bmatrix}.$$

The determinant of this matrix is readily seen to be the product of its diagonal elements. Note that if there are no inter-links, we must set $x_0 = 0$, and $|M|$ vanishes, i.e., there are no trees, as, of course, should be the case. However, if $k = 1$, i.e., if we have a single population, then, as we set $x_0 = 0$ in A_{ij} and $n_1 = n$, (17) reduces to

$$(25) \qquad |M| = n^{-1}(nx_1)^{n-1} = n^{n-2}x_1^{n-1},$$

so that we obtain for the number of trees n^{n-2}, which is Cayley's number.

On the other hand, consider the special case where all the links are inter-links, i.e., $l_0 = n - 1$. This amounts to setting $x_i = 0$ for all $i \neq 0$. Equation (17) then becomes

$$(26) \qquad \begin{aligned} |M| &= n^{k-2}x_0^{k-1}\textstyle\prod_{i=1}^{k} [(n - n_i)x_0]^{n_i-1} \\ &= \{n^{k-2}\textstyle\prod_{i=1}^{k} (n - n_i)^{n_i-1}\}x_0^{n-1}, \end{aligned}$$

and the number of trees reduces to

$$(27) \qquad T(\mathbf{n}, 1) = n^{k-2}\textstyle\prod_{i=1}^{k} (n - n_i)^{n_i-1}.$$

Equation (11) gives the number of "preferred" outcomes, i.e., the number of labelled trees with the prescribed partitions of n nodes and $n - 1$ links. To obtain the probability of the occurrence of a tree in a graph constructed randomly except for the constraints mentioned, we must calculate also the number of labelled graphs with the prescribed partitions. There are $\binom{N_i}{l_i}$ different ways of selecting the l_i pairs among the n_i nodes of the ith subset, and $\binom{N_0}{l_0}$ ways of obtaining the l_0 inter-links. Hence the total number of graphs obtained under the constraint of the given partition is

$$C(\mathbf{n}, \mathbf{l}) = \prod_{i=0}^{k} \binom{N_i}{l_i}, \tag{28}$$

and consequently the probability of obtaining a tree under the constraint is

$$P(\mathbf{n}, \mathbf{l}) = T(\mathbf{n}, \mathbf{l})/C(\mathbf{n}, \mathbf{l}). \tag{29}$$

If the number of links is $n - 1$, the graph is a tree if and only if it is connected. Hence, in this special case, we can identify the probability that the graph is connected with the probability that it is a tree. We are interested in the question of how this probability is affected by partitioning the nodes into subsets and apportioning intra-links among them, in particular, which partitions increase the probability that the graph is connected and which ones decrease it. As we have said, the partitions introduce a sort of distance bias in the constructions of the graph. Therefore results of the sort we seek may shed light on the way distance bias affects the probability of connectedness in an otherwise randomly constructed graph. We shall express these results as ratios of the probability of obtaining a tree when the population of nodes is partitioned to the corresponding probability when it is not partitioned.

3. The total "Exogamous bias." In this paper, we shall confine ourselves to the case where $l_i = 0$. That is to say, all the links will be inter-links. As an illustration, consider the fictitious case of a tribe with n members divided into k clans. Occasionally two members from different clans establish a friendship pact. A person may establish any number of such pacts but only with members of other clans and never with the same person more than once. The friendship relation is transitive. We are investigating the probability that after exactly $n - 1$ such pacts, all the members of the tribe will be "friends." [Note that a somewhat more realistic situation would be one where the pacts *within* a clan predominate, and only occasionally pacts occur between members of different clans (i.e. where there are "leaks in the cliques"). This situation will be treated in later papers. We begin with the case $l_i = 0$ simply because it seemed easier.]

We first examine the case where $k = 2$. Here the two populations may be the two sexes.

In the case where one subset contains only one node, the graph will always be connected. All the $n - 1$ "men" will have met with the one "woman." This case is therefore devoid of interest. Assume, then, $n_1 > 1$, $n_2 > 1$.

The total number of labelled graphs, connected or not, with n nodes and $n - 1$ links is clearly

$$C(n, n - 1) = \binom{\binom{n}{2}}{n-1}. \tag{30}$$

Of these graphs, $K(n)$ of them are distinct labelled trees, where

$$K(n) = n^{n-2}, \tag{31}$$

which is Cayley's number [Riordan, (1958)].

The probability that an arbitrarily selected graph is a tree is, accordingly,

$$P_1(n) = n^{n-2} \Big/ \binom{\binom{n}{2}}{n-1}. \tag{32}$$

When n is large, we approximate formula (32) by Stirling's formula

$$n! \cong (2\pi)^{\frac{1}{2}} n^{n+\frac{1}{2}} e^{-n} \tag{33}$$

and obtain

$$P_1(n) \cong (2\pi)^{\frac{1}{2}} \cdot 2^{n-1} \exp \{-\tfrac{1}{2}(n^2 - 3n + 5)\} \log_e n + \tfrac{1}{2} \log_e (n - 1) + \tfrac{1}{2}(n^2 - 3n + 3) \log_e (n - 2)\}. \tag{34}$$

Suppose now the n nodes are divided into 2 subsets, containing n_1 and n_2 nodes respectively, $n_1 + n_2 = n$.

Denote by $P_2(n_1, n_2)$ the probability that the corresponding graph (with $n - 1$ inter-links) is connected. From (29) with $k = 2$, $l_1 = l_2 = 0$, we obtain

$$P_2(n_1, n_2) = n_1^{n_2-1} n_2^{n_1-1} / \binom{n_1 n_2}{n-1}. \tag{35}$$

The quantity of interest is

$$R_2(n_1, n_2) \equiv P_2(n_1, n_2)/P_1(n), \tag{36}$$

which indicates to what extent the probability of being connected is enhanced (if $R_2 > 1$) or diminished (if $R_2 < 1$) when the set of n nodes is divided into two subsets and only "exogamous" connections are allowed.

THEOREM 2. *Let n be fixed. Then for n sufficiently large $P_2(n_1, n_2)$ attains a maximum when $|n_1 - n_2|$ is minimal,* i.e., *0 or 1.*

PROOF. Let $n_1 = \alpha n$; $n_2 = (1 - \alpha)n$. Substituting into (35), we have

$$\begin{aligned} P_2(n_1, n_2) &= P_2[\alpha n, (1 - \alpha)n] \\ &= (\alpha n)^{(1-\alpha)n-1}[(1 - \alpha)n]^{\alpha n-1} / \binom{\alpha(1-\alpha)n^2}{n-1}. \end{aligned} \tag{37}$$

Applying Stirling's formula, we obtain from (37)

$$\begin{aligned} &P_2[\alpha n, (1 - \alpha)n] \\ &= n^{n-2}(n - 1)!\,[\alpha(1 - \alpha)n^2 - n + 1]!\,\alpha^{(1-\alpha)n-1}(1 - \alpha)^{n-1}/[\alpha(1 - \alpha)n^2]! \\ &\cong 1/e \cdot (2\pi/n)^{\frac{1}{2}} \cdot [1 - (n - 1)/\alpha(1 - \alpha)n^2]^{\alpha(1-\alpha)n^2-n+\frac{3}{2}} \\ &\quad [\alpha^{\alpha n}(1 - \alpha)^{(1-\alpha)n}]^{-1}. \end{aligned} \tag{38}$$

$$\begin{aligned} \text{Let}\quad L_p &= \log_e P_2 \\ &= \tfrac{1}{2} \log_e (2\pi/n) - 1 - \alpha n \log_e \alpha - (1 - \alpha)n \log_e (1 - \alpha) \\ &\quad + [\alpha(1 - \alpha)n^2 - n + \tfrac{3}{2}] \log_e [1 - (n - 1)/\alpha(1 - \alpha)n^2] \\ &= -n + \tfrac{1}{2} \log_e (2\pi/n) - \alpha n \log_e \alpha - (1 - \alpha)n \log_e (1 - \alpha) \\ &\quad + [(n - 1)(n - 2)/2\alpha(\alpha - 1)n^2] + \textstyle\sum_{k=2}^{\infty} f_k(n)/[\alpha(1 - \alpha)]^k \end{aligned} \tag{39}$$

where the $f_k(n)$'s are polynomials in n^{-1}.

$P_2[\alpha n, (1 - \alpha)n]$ will attain its maximum when L_p attains its maximum. Hence, we set

$$\partial L_p/\partial\alpha = 0 \tag{40}$$

and solve for α.

$$\begin{aligned} \partial L_p/\partial\alpha = & -n \log_e \alpha - n + n \log_e (1 - \alpha) + n \\ & + [(n - 1)(n - 2)/n^2][(2\alpha - 1)/\alpha^2(1 - \alpha)^2] \\ & + (2\alpha - 1)\textstyle\sum_{k=2}^{\infty} k\cdot f_k(n)/[\alpha(1 - \alpha)]^{k+1} = 0. \end{aligned} \tag{41}$$

Since (41) must hold for all values of n, the solution must be independent of n. Thus we have

$$\alpha = \tfrac{1}{2}. \tag{42}$$

Consequently, $P_2[\alpha n, (1 - \alpha)n]$ attains its maximum when $n_i = n_2$, or, if n is odd, when $|n_1 - n_2| = 1$.

THEOREM 3. *Let* $|n_1 - n_2| \leqq 1$. *Then for* $n > 3$,

$$1 < R_2(n_1, n_2) < 2, \tag{43}$$

$$\lim_{n\to\infty} R_2(n_1, n_2) = 2. \tag{44}$$

REMARK. If $n = 2$ or 3,

$$P_2(n_1, n_2) = P_1(n) = R_2(n_1, n_2) = 1. \tag{45}$$

PROOF. First consider the case when n is even. Let $n = 2m$, so that $n_1 = n_2 = m$. By elementary calculation we establish

$$R_2(m, m) = 2^{2(1-m)}\prod_{i=0}^{2m-2} [(2m^2 - m - i)/(m^2 - i)]. \tag{46}$$

Let

$$K_1(i) = (2m^2 - m - i)/2(m^2 - i). \tag{47}$$

Then

$$\begin{aligned} K_1(i) &= 1, \quad \text{when} \quad i = m, \\ K_1(i) &> 1, \quad \text{when} \quad i > m, \\ K_1(i) &< 1, \quad \text{when} \quad i < m, \end{aligned} \tag{48}$$

and

$$R_2(m, m) = 2\prod_{i=0}^{2m-2} K_1(i). \tag{49}$$

When $m \geqq 3$, we rewrite (49) as follows

$$R_2(m, m) = 2\cdot\{\prod_{i=0}^{m-1} K_1(i)\}\cdot K_1(m)\cdot\{\prod_{i=m+1}^{2m-2} K_1(i)\}. \tag{50}$$

Set $i + j = 2m$, then

$$j = 2m - i, \text{ or } i = 2m - j. \tag{51}$$

Substituting (51) into the last factor in (50), we obtain:

$$(52)\quad R_2(m, m) = 2\cdot\{\textstyle\prod_{i=0}^{m-1} K_1(i)\}\cdot K_1(m)\cdot\{\prod_{j=2}^{m-1} K_1(2m - j)\}$$
$$= \{2\cdot K_1(0)\cdot K_1(1)\cdot K_1(m)\}\{\textstyle\prod_{i=2}^{m-1} K_1(i)\cdot K_1(2m - i)\}.$$

Let

$$(53)\qquad F_1 = 2\cdot K_1(0)\cdot K_1(1)\cdot K_1(m),$$

and

$$(54)\qquad G_1(i) = K_1(i)\cdot K_1(2m - i).$$

Hence,

$$(55)\qquad R_2(m, m) = F_1\cdot \textstyle\prod_{i=2}^{m-1} G_1(i).$$

$$F_1 = 2\cdot[(2m^2 - m)/2m^2]\cdot[(2m^2 - m - 1)/(2m^2 - 2)]$$
$$(56)\qquad \cdot[(2m^2 - 2m)/(2m^2 - 2m)]$$
$$= 2 - [2/(m + 1))((4m + 1)/4m]$$
$$(57)\qquad > 1 \text{ for all } m \geqq 2.$$

$$G_1(i) = K_1(i)\cdot K_1(2m - i)$$
$$(58)\qquad = [(2m^2 - m - i)/(2m^2 - 2i)][(2m^2 - 3m + i)/(2m^2 - 4m + 2i)]$$
$$= 1 + 3(m - i)^2/(4m^4 - 8m^3 + 8im - 4i^2).$$

Since $3(m - i)^2 > 0$ for all i in $[2, m - 1]$ and all m, and $4m^4 - 8m^3 + 8im - 4i^2 = 4m^3(m - 2) + 4i(2m - i) > 0$ for all i in $[2, m - 1]$ and all $m > 2$, we have

$$(59)\qquad G_1(i) > 1 \quad \text{for all} \quad i \quad \text{in} \quad [2, m - 1] \quad \text{for all} \quad m > 2.$$

By combining the results of (57) and (59) we obtain

$$(60)\qquad R_2(m, m) = F_1\cdot \textstyle\prod_{i=2}^{m-1} G_1(i) > 1 \quad \text{for all} \quad m \geqq 3.$$

When $m = 2$,

$$(61)\qquad R_2(m, m) = 2\cdot K_1(0)\cdot K_1(1)\cdot K_1(2)$$
$$= 2\cdot K_1(0)\cdot K_1(1)\cdot K_1(m)$$
$$(62)\qquad = F_1\,.$$

However, according to (57), $F_1 > 1$ for all $m \geqq 2$. Thus, when $m = 2$, $R_2(m, m) > 1$. Consequently, we have:

$$(63)\qquad R_2(m, m) > 1 \quad \text{for all} \quad m \geqq 2.$$

To show $R_2(m, m) < 2$, we write

$$(64)\qquad R_2(m, m) = 2\{\textstyle\prod_{i=0}^{m-2} K_1(i)\}K_1(m - 1)\{\prod_{i=m}^{2m-2} K_1(i)\}$$
$$(65)\qquad = 2K_1(m - 1)\{\textstyle\prod_{i=0}^{m-2} K_1(i)K_1(2m - 2 - i)\}.$$

Let

$$(66)\quad H_1(i) = K_1(i)K_1(2m - 2 - i)$$

$$(67)\quad = 1 + (3i^2 + 6i - 6im - m^2 - 2m)/(4m^4 - 8m^3 + 8m^2 + 8im - 8i - 4i^2).$$

Denote by $\Delta_H(i)$ the numerator of the fraction in (67) and by $D_H(i)$ the denominator. Then

$$(68)\quad (d/di)\Delta_H(i) = 6(1 + i - m) < 0 \text{ for all } i \text{ in } [0, m - 2].$$

Hence

$$(69)\quad \Delta_H(0) > \Delta_H(i) > \Delta_H(m - 2).$$

However

$$(70)\quad \Delta_H(0) = -m^2 - 2m < 0 \quad \text{for all} \quad m \geqq 2.$$

Therefore

$$(71)\quad \Delta_H(i) < 0.$$

Also

$$(72)\quad D_H(i) > 0.$$

Thus,

$$(73)\quad H_1(i) < 1 \quad \text{for all } i \text{ in } [0, m - 2] \text{ and } m \geqq 2.$$

From (48) we see that $K_1(m - 1) < 1$. Thus,

$$(74)\quad R_2(m, m) = 2\cdot K_1(m - 1)\prod_{i=0}^{m-2} H_1(i) < 2 \quad \text{for all} \quad m \geqq 2.$$

To show that $\lim_{n\to\infty} R_2(n_1, n_2) = 2$, we note from (67) that

$$(75)\quad H_1(i) = 1 + \Delta_H(i)/D_H(i)$$

where

$$(76)\quad \Delta_H(i) < \Delta_H(0) \quad \text{and} \quad D_H(i) > D_H(0) \quad \text{for all } i \text{ in } [0, m - 2].$$

Thus,

$$(77)\quad 1 + \Delta_H(m - 2)/D_H(m - 2) < H_1(i) < 1 + \Delta_H(0)/D_H(0) \quad \text{for all } i \text{ in } [1, m - 3].$$

Hence,

$$(78)\quad \{1 + \Delta_H(m - 2)/D_H(m - 2)\}^{m-1} < \prod_{i=0}^{m-2} H_1(i) < \{1 + \Delta_H(0)/D_H(0)\}^{m-1}.$$

Now,

$$\lim_{m\to\infty}\{1+\Delta_H(m-2)/D_H(m-2)\}^{m-1}$$

$$(79)\qquad = \lim_{m\to\infty}\{1-(m^2-m+2)^{-1}\}^{m-1}$$

$$= \lim_{m\to\infty}\exp\{(1-m)/(m^2-m+2)\} = 1.$$

Similarly,

$$(80)\quad \lim\{1+\Delta_H(0)/D_H(0)\}^{m-1}$$

$$= \lim_{m\to\infty}\{1-m(m+2)/4m^2(m^2-2m+2)\}^{m-1} = 1.$$

Hence,

$$(81)\qquad \lim_{m\to\infty}\prod_{i=0}^{m-2} H_1(i) = 1.$$

And,

$$(82)\qquad \lim_{m\to\infty} K_1(m-1) = 1.$$

Therefore,

$$(83)\quad \lim_{m\to\infty} R_2(m,m) = \{\lim_{m\to\infty} 2\cdot K_1(m-1)\}\cdot\{\lim_{m\to\infty}\prod_{i=0}^{m-2} H_1(i)\}$$

$$(84)\qquad = 2\cdot 1 = 2.$$

The proof for the case when n is odd is entirely analogous.

THEOREM 4. *The ratio $R_2(m, m)$ increases monotonically with m for all $m > 2$.*

The proof is obtained by showing by straightforward but laborious calculations that

$$(85)\qquad R_2(m+1, m+1)/R_2(m,m) > 1 \quad \text{for all} \quad m > 2.$$

By methods similar to those used in proving Theorem 4, it can be shown also that $R_2(m+1, m)$ is monotone increasing with m.

We have thus shown that when the n nodes are divided into two equal or nearly equal subsets and only inter-links are allowed (i.e., under "exogamous constraint"), the probability of connectedness (i.e., that the graph with $n - 1$ links is a tree) is always increased for $n > 2$ by a factor (greater than unity) which does not exceed 2 and approaches 2 in the limit as n becomes infinitely large.

By a method entirely analogous to that used in Theorem 2, it can be shown that for any k, the probability of connectedness (i.e., of a tree) is greatest when the distribution of nodes is as nearly equal as possible. The only exception is the case where $k = 2$ and one of the populations has a single node, in which case $P_2 = 1$, and so R_2 increases without bound with n.

THEOREM 5. *Let n nodes be equally distributed over k subsets, so that $n = mk$, $n_i = m$. Then if the $n - 1$ links are all inter-links,*

$$(86)\quad R_k = \lim_{n\to\infty} P_k(\mathbf{n})/P_1(n) = (k/(k-1))^{k-1}\cdot\exp(k/(k-1)-2).$$

PROOF. From (11), we obtain

$$(87)\qquad T(\mathbf{n}, 1) = n^{k-2}(n - n/k)^{n-k} = n^{n-2}((k - 1)/k)^{n-k}.$$

Further,

$$(88)\quad N_0 = \binom{n}{2} - \sum_{i=1}^{k} \binom{n_i}{2} = \tfrac{1}{2}[n^2 - \sum_{i=1}^{k} n_i^2] = ([k - 1)/2k]n^2.$$

Substituting (87) and (88) into (29), we obtain with the help of Stirling's formula for large n,

$$(89)\quad P_k(\mathbf{n}) = (2\pi)^{\frac{1}{2}}((k - 1)/k)^{1-k}2^{n-1}(n - 1)^{n-\frac{1}{2}}n^{-n} \cdot \{1 - (n - 1)[(k - 1)n^2/2k]^{-1}\}^{((k-1)/2k)n^2-n+\frac{3}{2}}.$$

Then, by (34) and (89),

$$\begin{aligned} R_k &= \lim_{n\to\infty} R_k(\mathbf{n}) = \lim_{n\to\infty} P_k(\mathbf{n})/P_1(n) \\ (90)\qquad &= \lim_{n\to\infty} (2\pi)^{\frac{1}{2}}((k - 1)/k)^{i-k}2^{n-1}(n - 1)^{n-\frac{1}{2}}n^{-n} \\ &\quad \cdot[1 - (n - 1)((k - 1)/2k)n^2)^{-1}]^{((k-1/2k)n^2-n+\frac{3}{2}} \\ &\quad \cdot[2\pi 2^{n-1}n^{-\frac{1}{2}(n^2-3n+5)}(n - 1)^{\frac{1}{2}}(n - 2)^{\frac{1}{2}(n^2-3n+3)}]^{-1}. \end{aligned}$$

Now

$$\begin{aligned} (91)\quad ((n - 1)/n)^{n-1} &= \exp\{(n - 1)\log_e(1 - n^{-1})\} \\ &= \exp\{-1 + O(n^{-1});\end{aligned}$$

$$\begin{aligned} (92)\quad ((n - 2)/n)^{\frac{1}{2}(n^2-3n+3)} &= \exp\{\tfrac{1}{2}(n^2 - 3n + 3)\log_e(1 - 2/n)\} \\ &= \exp\{-n + 2 + O(n^{-1})\}; \end{aligned}$$

and

$$\begin{aligned} &[1 - (n - 1)((k - 1)/2k)n^2)^{-1}] \\ (93)\quad &= \exp\{((k - 1)n^2/2k - n + \tfrac{3}{2})\log_e(1 - (2k/(k - 1))\cdot(n - 1)/n^2)\} \\ &= \exp\{-n + 1 + k/(k - 1) + O(n^{-1})\}. \end{aligned}$$

Thus (9) becomes (94):

$$\begin{aligned} R_k &= \lim_{n\to\infty}((k - 1)/k)^{1-k} \\ (94)\qquad &\quad \cdot\exp\{-1 - (2 - n) - n + 1 + k/(k - 1) + O(n^{-1})\} \\ &= \lim_{n\to\infty}((k - 1)/k)^{1-k}\exp\{k/(k - 1) - 2 + O(n^{-1})\} \\ &= (k/(k - 1))^{k-1}\exp(k/(k - 1) - 2). \end{aligned}$$

QED

It can further be shown that the same limit obtains if k does not divide n provided the n nodes are distributed "as equally as possible" among the k subsets.

REMARKS. If the number of subsets is large, the probability of obtaining a tree from a graph with $n - 1$ inter-links is not enhanced.

TABLE 1

P_1: Probability of obtaining a tree from a randomly constructed graph with n nodes and n − 1 links; P_2 : the corresponding probability when the n nodes are divided into two subsets, n_1 and n_2 , with exogamous constraint; $R_2 = P_2/P_1$.

TABLE 1(a)

n is even

n	n_1	n_2	P_2	P_1	R_2
4	2	2	1.000	0.800	1.250
6	4	2	0.571	0.431	1.324
6	3	3	0.643	0.431	1.489
8	6	2	0.242	0.221	1.095
8	5	3	0.314	0.221	1.421
8	4	4	0.358	0.221	1.617
10	8	2	8.95×10^{-2}	1.128×10^{-1}	0.793
10	7	3	1.21×10^{-1}	1.128×10^{-1}	1.077
10	6	4	1.69×10^{-1}	1.128×10^{-1}	1.499
10	5	5	1.91×10^{-1}	1.128×10^{-1}	1.694
12	10	2	3.05×10^{-2}	5.76×10^{-2}	0.528
12	9	3	4.07×10^{-2}	5.76×10^{-2}	0.707
12	8	4	6.50×10^{-2}	5.76×10^{-2}	1.128
12	7	5	8.99×10^{-2}	5.76×10^{-2}	1.559
12	6	6	1.00×10^{-1}	5.76×10^{-2}	1.745
14	12	2	9.84×10^{-3}	2.95×10^{-2}	0.332
14	11	3	1.24×10^{-2}	2.95×10^{-2}	0.421
14	10	4	2.17×10^{-2}	2.95×10^{-2}	0.737
14	9	5	3.51×10^{-2}	2.95×10^{-2}	1.187
14	8	6	4.75×10^{-2}	2.95×10^{-2}	1.608
14	7	7	5.27×10^{-2}	2.95×10^{-2}	1.782
16	14	2	3.06×10^{-3}	1.52×10^{-2}	0.201
16	13	3	3.57×10^{-3}	1.52×10^{-2}	0.234
16	12	4	6.63×10^{-3}	1.52×10^{-2}	0.435
16	11	5	1.20×10^{-2}	1.52×10^{-2}	0.788
16	10	6	1.89×10^{-2}	1.52×10^{-2}	1.243
16	9	7	2.50×10^{-2}	1.52×10^{-2}	1.646
16	8	8	2.75×10^{-2}	1.52×10^{-2}	1.810
18	16	2	9.26×10^{-4}	7.87×10^{-3}	0.117
18	15	3	9.75×10^{-4}	7.87×10^{-3}	0.123
18	14	4	1.88×10^{-3}	7.87×10^{-3}	0.238
18	13	5	3.73×10^{-3}	7.87×10^{-3}	0.473
18	12	6	6.65×10^{-3}	7.87×10^{-3}	0.845
18	11	7	1.02×10^{-2}	7.87×10^{-3}	1.295
18	10	8	1.32×10^{-2}	7.87×10^{-3}	1.678
18	9	9	1.44×10^{-2}	7.87×10^{-3}	1.831
20	10	10	7.55×10^{-3}	4.08×10^{-3}	1.848
22	11	11	3.96×10^{-3}	2.12×10^{-3}	1.862
24	12	12	2.08×10^{-3}	1.11×10^{-3}	1.873
26	13	13	1.09×10^{-3}	5.81×10^{-4}	1.883
28	14	14	5.76×10^{-4}	3.04×10^{-4}	1.892
30	15	15	3.03×10^{-4}	1.60×10^{-4}	1.899
32	16	16	1.60×10^{-4}	8.42×10^{-5}	1.905
34	17	17	8.48×10^{-5}	4.43×10^{-5}	1.911
36	18	18	4.48×10^{-5}	2.34×10^{-5}	1.916
38	19	19	2.37×10^{-5}	1.23×10^{-5}	1.920
40	20	20	1.25×10^{-5}	6.54×10^{-6}	1.924
42	21	21	6.67×10^{-6}	3.46×10^{-6}	1.928
44	22	22	3.54×10^{-6}	1.83×10^{-6}	1.931
46	23	23	1.88×10^{-6}	9.73×10^{-7}	1.934
48	24	24	1.00×10^{-6}	5.16×10^{-7}	1.937

TABLE 1(b)
n is odd

n	n_1	n_2	P_2	P_1	R_2
5	3	2	0.800	0.595	1.344
7	5	2	0.381	0.309	1.23
7	4	3	0.467	0.309	1.509
9	7	2	0.149	0.158	0.944
9	6	3	0.199	0.158	1.264
9	5	4	0.254	0.158	1.607
11	9	2	5.26×10^{-2}	0.80×10^{-1}	0.653
11	8	3	7.13×10^{-2}	0.80×10^{-1}	0.885
11	7	4	1.07×10^{-1}	0.80×10^{-1}	1.328
11	6	5	$1,34 \times 10^{-1}$	0.80×10^{-1}	1.672
13	11	2	1.74×10^{-2}	4.12×10^{-2}	0.422
13	10	3	2.27×10^{-2}	4.12×10^{-2}	0.551
13	9	4	3.81×10^{-2}	4.12×10^{-2}	0.925
13	8	5	5.72×10^{-2}	4.12×10^{-2}	1.388
13	7	6	7.09×10^{-2}	4.12×10^{-2}	1.718
15	13	2	5.51×10^{-3}	2.12×10^{-2}	0.259
15	12	3	6.72×10^{-3}	2.12×10^{-2}	0.316
15	11	4	1.21×10^{-2}	2.12×10^{-2}	0.572
15	10	5	2.08×10^{-2}	2.12×10^{-2}	0.981
15	9	6	3.05×10^{-2}	2.12×10^{-2}	1.440
15	8	7	3.72×10^{-2}	2.12×10^{-2}	1.753
17	15	2	1.69×10^{-3}	1.09×10^{-2}	0.154
17	14	3	1.87×10^{-3}	1.09×10^{-2}	0.171
17	13	4	3.55×10^{-3}	1.09×10^{-2}	0.324
17	12	5	6.76×10^{-3}	1.09×10^{-2}	0.618
17	11	6	1.13×10^{-2}	1.09×10^{-2}	1.039
17	10	7	1.62×10^{-2}	1.09×10^{-2}	1.486
17	9	8	1.949×10^{-2}	1.09×10^{-2}	1.780
19	10	9	1.02×10^{-2}	5.67×10^{-3}	1.802
21	11	10	5.36×10^{-3}	2.94×10^{-3}	1.820
23	12	11	2.82×10^{-3}	1.53×10^{-3}	1.834
25	13	12	1.48×10^{-3}	8.03×10^{-4}	1.847
27	14	13	7.81×10^{-4}	4.20×10^{-4}	1.858
29	15	14	4.12×10^{-4}	2.20×10^{-4}	1.867
31	16	15	2.17×10^{-4}	1.16×10^{-4}	1.876
33	17	16	1.15×10^{-5}	6.11×10^{-5}	1.883
35	18	17	6.09×10^{-5}	3.22×10^{-5}	1.889
37	19	18	3.22×10^{-5}	1.70×10^{-5}	1.895
39	20	19	1.70×10^{-5}	8.99×10^{-6}	1.900
41	21	20	9.06×10^{-6}	4.75×10^{-6}	1.905
43	22	21	4.81×10^{-6}	2.52×10^{-6}	1.909
45	23	22	2.55×10^{-6}	1.33×10^{-6}	1.913
47	24	23	1.36×10^{-6}	7.09×10^{-7}	1.917

$$\lim_{k\to\infty} R_k = e \cdot e^{1-2} = 1. \tag{95}$$

Note, however, that in our context, k cannot exceed n. Note also that when $k = n$, the situation reduces to the case of a single population, hence $R_n = 1$.

When $k = 2$, R_k in (94) properly reduces to 2, which is also given in Theorem 3.

Furthermore, note that if $k = n/2$, (87) is then reduced to equation (10) of Weinberg (1954).

In summary, the "exogamous constraint" (i.e., when only inter-links are allowed) in the construction of a random graph with n nodes and $n - 1$ links enhances the probability of obtaining a connected graph (i.e., a tree) by a factor which ranges from one to two if n is large and the nodes are distributed approximately equally over k subsets. Moreover for $k = 2$, the probability of connectedness is greatest (when n is sufficiently large) the more equally the nodes are divided among the two subsets. The case when one subset has only one node is a notable exception, since in that case a tree is obtained with certainty.

Table 1 gives numerical values of P_1, P_2, and R_2 for all values of n_1 and n_2 where n ranges from 4 to 18 and for the equal or nearly equal values of n_1 and n_2 as n ranges from 19 to 48. From the table, one sees that also for small values of n for each fixed n, R_2 increases as $|n_1 - n_2|$ decreases. Also R_2 increases with n through both the even and the odd values when $|n_1 - n_2|$ is minimal. However, as n increases through successive values and $|n_1 - n_2|$ is minimal, R_2 oscillates when $8 \leqq n \leqq 18$.

REFERENCES

ERDÖS, P. and RÉNYI, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5** 17–61.

RIORDAN, J. (1958). *An Introduction to Combinatorial Analysis.* Wiley, New York.

RUBIN, H. and SITGREAVES, R. (1954). Probability distributions related to random transformations of a finite set. Technical Report No. 19A, *Applied Mathematics and Statistics Laboratory*, Stanford University.

TRENT, H. M. (1954). A note on the enumeration and listing of all possible trees in a connected linear graph. *Proc. Nat. Acad. Sci., U.S.A.* **40** 1004–1007.

WEINBERG, L. (1958). Number of trees in a graph. *Proc. I.R.E.* **46** 1954–55.