Copy #3

CONNECTIVITY IN PROBABILISTIC GRAPHS

IRWIN MARK JACOBS

TECHNICAL REPORT 356

SEPTEMBER 15, 1959

LOAN COPY ONLY

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
RESEARCH LABORATORY OF ELECTRONICS
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory of the Department of Electrical Engineering and the Department of Physics.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the U. S. Army (Signal Corps), the U.S. Navy (Office of Naval Research), and the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA36-039-sc-78108, Department of the Army Task 3-99-20-001 and Project 3-99-00-000.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 356 September 15, 1959

CONNECTIVITY IN PROBABILISTIC GRAPHS

Irwin Mark Jacobs

Submitted to the Department of Electrical Engineering, August 21, 1959, in partial fulfillment of the requirements for the degree of Doctor of Science.

Abstract

A probabilistic graph is a linear graph in which both nodes and links are subject to random erasure. Such a graph may be thought of as an idealized model of a communication network in which switching centers (nodes) and information channels (links) either operate perfectly or fail entirely. This report deals with the reliability of the communication network. Two reliability criteria are established. The first is the probability that a path exists between all pairs of nodes that remain in the associated probabilistic graph after erasure, and the second is the probability that a path exists between one pair of nodes selected at random.

For small communication networks, the interesting questions concern analysis – the calculation of the reliability of a given network – and synthesis – the construction, under certain constraints, of graphs with maximum reliability. For large communication networks, the emphasis is on the link-to-node densities necessary and sufficient for attaining a desired reliability. The sufficient densities are determined by approximate analysis of several graph configurations. It is shown that reliability (under the first reliability criterion) approaches 1 exponentially with the decrease in the difference between the link-to-node density and a constant times the logarithm of the number of nodes. This constant is a function of the graph that is being analyzed and of the link and node reliabilities. The average reliability of graphs chosen randomly by two different procedures is also determined.

TABLE OF CONTENTS

GLOSSARY

Symbol	Definition
n	Number of nodes in each base graph
L	Number of links in each base graph
d	Number of links divided by number of nodes ($d=L/n$)
p	Probability that a link is reliable ($p=1-q$)
t	Probability that a node is reliable
$\overline{S}$	Probability of strong connectivity averaged over base-graph ensemble
$\overline{W}$	Probability of weak connectivity averaged over base-graph ensemble
Node I	Intact node chosen in each member as the first node in a bush
$P_1(2k)$	Probability that at least 2k different, existing nodes are attached to node I by an existing link
$B(b, 2d)$	Probability of obtaining a base graph in which the number of links attached to node I exceeds a fraction b, $0 < b < 1$, of the mean number, 2d, of attached links
$E(e, 2bd)$	Probability of obtaining members in which 2ebd, or more, of the 2bd links attached to node I exist, if $0 < e < p$
$G(g, 2ebd)$	Probability that 2gebd different nodes are attached to node I by the 2ebd existing links, before node destruction, if $0 < g < 1$
$H(h, 2gebd)$	Probability that at least 2hgebd nodes attached to node I, out of the 2gebd different nodes attached by existing links, exists, if $0 < h < t$
Superscript i	When placed on any of the probabilities $B(b, x)$, $E(e, x)$, $G(g, x)$, or $H(h, x)$, this symbol indicates that that probability refers to nodes connected to level-i nodes rather than to node I.
k	Stands for the product hgebd ($k=hgebd$)
$M(z)$	Probability that all of the existing nodes in a graph are attached to the z, or more, existing links originating in the j^{th} level
$NR(\alpha, P)$	Exponent in the bound on the tail of a binomial distribution, where P is the probability of success, N is the number of trials, and αN is the number of successes in the first sample included in the tail (see Eq. 119)
$Q(z)$	Probability that node 2 is attached to at least one of the z, or more, existing links originating in the j^{th} level
$T(z)$	Probability that two bushes built out from nodes 1 and 2 in a purged ensemble have a common existing node attached to at least one of the z existing links originating in their j^{th} levels

I. INTRODUCTION

The study of the connectivity properties of probabilistic graphs is an interesting mathematical discipline. In common with other mathematical disciplines, it also has important physical applications. This research is principally concerned with one particular application, the use of the probabilistic graph as an abstract model of a communication system containing unreliable components. In this application, the mathematical theory allows one to define and calculate the reliability of the communication system, to determine required densities (ratios) of communication channels to message centers, and, in certain cases, to specify optimum configurations of the communication system.

1.1 THE MATHEMATICAL MODEL

The concept of the linear graph is of central importance in this research. For our purpose, the linear graph is defined as a finite collection of primitive objects, which we refer to as the "nodes" of the graph, together with a set of unordered pairs of nodes, which we refer to as the "links" of the graph. Physically, a linear graph might be an electrical network, in which the links are elements (resistors, capacitors, inductors) and the nodes are connection points; or the graph might be a chemical molecule, in which nodes are atoms and links are bonds; or, as we shall discuss in section 1.2, the graph might be a communication network, in which nodes are stations and links are channels. For example, a 6-node, 6-link graph might have nodes (a, b, c, d, e, f) and links ($\overline{ab}$, $\overline{ab}$, $\overline{ac}$, $\overline{bc}$, $\overline{de}$, $\overline{ee}$). Note that $\overline{ac}$ and $\overline{ca}$ both denote the same link. A standard representation for a graph is the line drawing in which nodes are represented by points, and links by lines (not necessarily straight) connecting the points. Figure 1 shows this 6-node, 6-link graph. Notice that a graph may contain isolated nodes such as f, links in parallel such as ($\overline{ab}$, $\overline{ab}$), and links with both ends terminating on the same node such as $\overline{ee}$. This last type of link is called a "sling."

The characteristic of a linear graph that most interests us is its connectivity. Before defining connectivity, we must introduce several concepts. We feel it unwise to introduce topological terminology and definitions into this report, since no (nontrivial) topological concepts or theorems are required. Only definitions familiar to the circuit theorist will be used.

A "path" between two nodes, say, between the nodes n_1 and n_k, is a subset of the links of the graph of the form ($\overline{n_1 n_2}$, $\overline{n_2 n_3}, \ldots, \overline{n_{k-1} n_k}$), where all the nodes $n_1, \ldots n_k$ are different. A "cycle" in a graph is a path with one additional link joining the two terminal nodes of the path. Thus there are three paths between nodes a and b in Fig. 1, the two $\overline{ab}$'s and the set ($\overline{ac}$, $\overline{cb}$). The set ($\overline{ac}$, $\overline{cb}$, $\overline{ba}$) is a

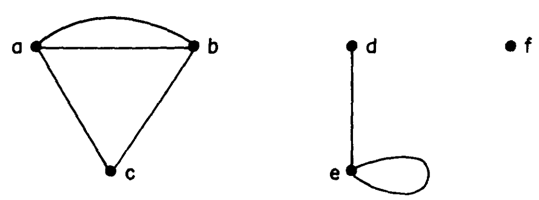


Fig. 1. A 6-node, 6-link linear graph.

cycle of the graph.

A "tree" in an n-node graph is a set of n − 1 links that contains a path between every pair of nodes in the graph. It can easily be shown that any set of n − 1 links that contains no slings or cycles is a tree. Finally, a graph is said to be "strongly connected" if it contains a path between every pair of nodes. Alternatively, we might say that a graph is strongly connected if it contains at least one tree. The graph of Fig. 1 is not strongly connected. The graph of Fig. 2 is strongly connected, and contains three trees: ($\overline{ca}$, $\overline{ab}$, $\overline{bd}$), ($\overline{cb}$, $\overline{ab}$, $\overline{bd}$), and ($\overline{ac}$, $\overline{cb}$, $\overline{bd}$).

The "probabilistic graph" is an ensemble of linear graphs with an associated probability measure. The ensemble is generated from a specified linear graph, referred to as the "base graph" of the probabilistic graph, by randomly erasing both the links and the nodes of the base graph in such a manner that every link of the base graph appears with a probability p, every node of the base graph appears with a probability t, and all appearances of links and nodes are statistically independent. (The words "destroy," "remove," and "fail" will be used synonymously with "erase" to denote the process of removing a link or node from the base graph. Nodes or links not erased are said to "appear," or "exist," or "be reliable.") When a node is erased, all links terminating on that node are left dangling – that is, these links can no longer enter into any paths. To prevent confusion, the dangling links are also erased, but this erasure is for convenience and is not part of the random-link erasure. From another viewpoint, we can say that a link appears in the ensemble with probability p conditional on the nonerasure of its two terminating nodes. Figure 3 shows all members of the ensemble that are generated from the base graph by link and node erasures. The probability associated with each ensemble member appears directly beneath the member.

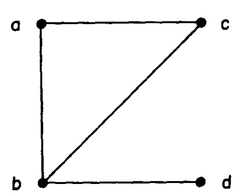


Fig. 2. A strongly connected linear graph.

Three probabilities defined on the probabilistic graph are of immediate interest. The first is the probability of strong connectivity, S; it is equal to the total measure of all graphs in the ensemble that are strongly connected. Note that this definition allows graphs with erased nodes to contribute to the probability of strong connectivity if and only if all remaining nodes are strongly connected. For the sake of completeness, the graphs consisting of one node (plus slings) and no nodes (the vacuous graph) are considered to be strongly connected.

The second probability is the path probability, P_{ij}; it is equal to the measure of all members of the ensemble in which both nodes i and j appear, and in which at least one path exists between nodes i and j. The third probability is the probability of weak

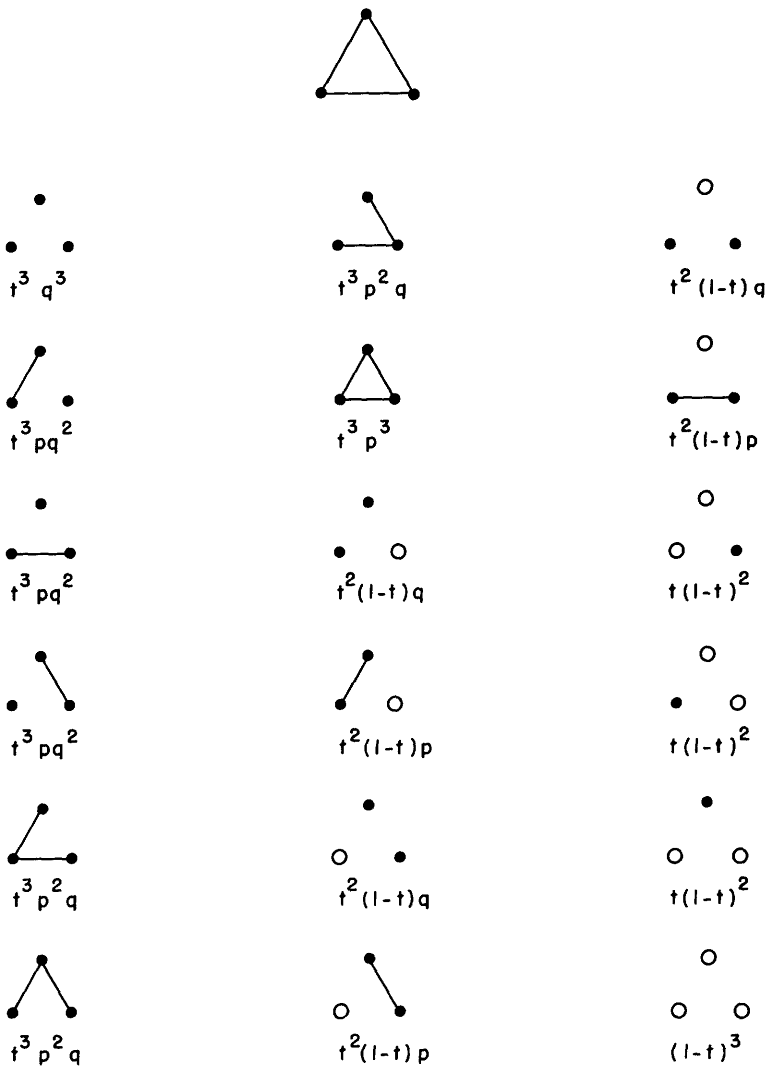


Fig. 3. Probabilistic graph (○ denotes erased node). Top central figure is the base graph.

connectivity, W; it is the path probability, P_{ij}, averaged over all pairs of nodes, when the pairs of nodes are assumed to be equally likely.

Many other probabilities might be introduced into the study of probabilistic graphs. However, the three that have been mentioned, the probability of strong connectivity S, the probability of weak connectivity W, and the path probability between nodes i and j, P_{ij}, are sufficient for our study of reliability in large communication systems.

1.2 APPLICATION OF PROBABILISTIC GRAPHS TO COMMUNICATION NETWORKS

A communication network is a collection of message (switching) centers that attempt to transfer information to one another over a variety of connecting channels. However, neither the centers nor the channels are necessarily reliable at any given time. For example, a center might lose its power supply, or be destroyed by fire or bomb, or be captured by an enemy. Likewise, a communication channel might be busy, or it might be inoperative because of an amplifier failure, a broken or cut telephone wire, or a jammed radio link. In spite of these possibilities, it is highly desirable that the remaining switching centers be able to communicate with each other.

To permit theoretical study of the reliability of a communication network, the following idealizations are made. A switching center is assumed to function with complete reliability with a probability t, or to fail entirely with a probability $1 - t$. Thus, with probability t, a switching center can accept messages for transmission, reroute messages from an incoming channel to an outgoing channel, and deliver messages. With probability $1 - t$, it can perform none of these functions. Likewise, a communication channel is assumed to have an infinite capacity with probability p, or zero capacity with probability $q = 1 - p$. Thus, with probability p, a channel is reliable and an unlimited amount of information can be sent over it without error. With probability $q = 1 - p$, no information can be sent. The failures of all centers and channels are assumed to be statistically independent.

These idealizations provide a rather black-and-white picture of network performance. In practice, a channel, for example, may have an available capacity that is neither zero nor infinite. Indeed, a realistic picture of the channel might require the assignment of capacity as a random process with rather complete statistical information. But the idealization of channel and switching center does provide a useful starting point for a study of the reliability of communication networks, and — what is of primary importance for this research — it permits the immediate abstraction of a communication network to a probabilistic graph. Stations are identified with nodes, channels with links, and the ability to transmit a message between two centers is identified with the existence of a path between the corresponding nodes.

The significance of our three probabilities can now be considered. The first, the probability of strong connectivity, S, is an indication of the reliability of the network as a whole. It equals the probability that all operative message centers are able to communicate with one another. The second, the path probability relative to two nodes, P_{ij}, is the probability that a message can be sent between two specified centers. The third, the probability of weak connectivity, W, is the probability that a message can be sent between two centers not specified in advance. Thus, W is the average probability of success of a network subscriber who attempts to send a message between two centers when his choice of sending and receiving stations is equally likely over all stations. Note that if one center has a low probability of being connected to the remainder of the stations, S may be very low, while W is quite large.

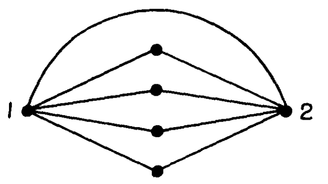
The physical application of our model to reliable communication networks increases the importance of an additional quantity. This quantity is an upper bound on the longest paths necessary for passing from one node to another. (The length of a path is defined as the number of links that it contains.) Since every passage from one link in a path to the next is equivalent to the rerouting of a message by a switching center, and because this rerouting would normally introduce a certain delay, the path length cannot be allowed to grow too long. Thus, even intact paths will not be allowed to contribute to the several probabilities if the path lengths are too long. This question will be considered later.

1.3 SUMMARY OF THIS RESEARCH

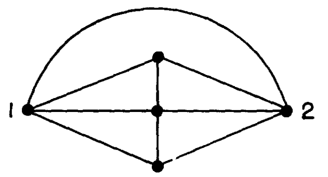
The results obtained from this research can be conveniently separated into two groups on the basis of their applicability to small or large graphs. (Whether a graph is small or large depends on the number of nodes and links it has. A graph with less than 10 nodes is usually considered small, and a graph with more than 50 nodes, large.) The distinction between small and large graphs is quite natural, for the techniques that can be used and the emphasis that must be placed are different in the two cases. For example, in small graphs, exact analysis (the calculation of the probabilities S, W, or P_{ij}) is feasible, and the results can be expressed in workable analytical form. As a result, synthesis (the determination of the n-node, L-link graph with the largest S, W, or P_{ij} of all n-node, L-link graphs) becomes a meaningful pursuit.

In large graphs, however, the exact probabilities cannot be expressed in workable analytical form. Upper and lower bounds are usually too coarse to reflect the slight changes in probability caused by moving a link around in a large graph (in an attempt to optimize). The emphasis in this synthesis must therefore be changed. Because of our interest in applying our results to reliable communication networks, the question of meaningful analysis and synthesis in large graphs can be stated: What density of links to nodes is necessary and sufficient to provide desired values of the probabilities of strong and weak connectivity? This question will be answered in the course of this research.

Before continuing the discussion of the large-graph problem, let us summarize the results obtained for small graphs. The first problem that is considered is the exact analysis of probabilistic graphs and, in particular, the exact analysis of the



(a)



(b)

Fig. 4. Optimum and non-optimum 9-link graphs.

complete-base probabilistic graph (a complete graph is a graph with one link between every pair of nodes). The complete graph was chosen for analysis for two reasons: it is the n-node, $\binom{n}{2}$-link configuration that maximizes the probabilities of strong and weak connectivity, and it forms a useful subgraph for the construction of large graphs with variable link densities, as we shall show.

The second problem concerns a synthesis for maximizing P_{12}, subject to certain constraints, and it can be stated as follows: If node reliability t is equal to 1, link reliability p is greater than $1/\sqrt{2}$, and no links are permitted in parallel, then the optimum manner of connecting two nodes with an odd number of links (optimum in the sense of maximizing the probability of a path between the two nodes) is to place one link directly between the two nodes, and all other links in distinct parallel paths of length 2 between the nodes. This arrangement is illustrated in Fig. 4a. A graph for which P_{12} is not maximized is shown in Fig. 4b. The optimum arrangement for $p < 1/\sqrt{2}$ is not known, although for very small p, the arrangement shown in Fig. 4a is optimum.

The third problem is that of obtaining an upper bound on the probability of strong connectivity in any n-node, L-link graph. This bound is obtained in two forms by two different techniques, and the equality of these forms constitutes an interesting identity on the tail of the binomial distribution. However, the bound can be attained by a graph if and only if every set of $n - 1$ links forms a tree, and this condition cannot be met if the number of links exceeds the number of nodes.

The rest of this study is concerned with large graphs. We define the link density, d, of a graph as the ratio of the number of links, L, in the graph to the number of nodes, n, and hence

$$d = L/n \tag{1}$$

First, we determine the density, d, that is necessary if a graph is to have a specified value of connectivity probability (either S or W). Next, we consider several base-graph configurations and determine the density, d, that is sufficient in these graphs to yield the specified values of S or W. A base graph for which the sufficient density is within, say, a factor of 2 of the necessary density is called "efficient."

The necessary density is obtained by developing upper bounds to S and W that apply to all n-node graphs with link density d. These bounds are:

$$S \leqslant 1 - \frac{1}{2}\, e^{-(2d \log 1/q - \log tn)} \qquad (2)$$

and

$$W \leqslant t^2(1-q^{2d}) \qquad (3)$$

where $q = 1 - p$ is the probability that a link fail, and t is the probability that a node not fail. The necessary density is just that density for which the right-hand side of Eq. 2 or Eq. 3 is equal to the specified value of S or W. A smaller density can result only in a smaller value of S or W. The arrangement of the right-hand side of Eq. 2 was chosen to exhibit the linear relationship between the necessary link density and the logarithm of the number of nodes, for a fixed value of S. This dependence of d on $\log n$ constitutes one of the more interesting results of this research. In contrast, the bound on W is independent of the number of nodes.

Equations 2 and 3 specify the necessary link densities. An examination of various base-graph configurations specifies sufficient link densities. Several of these configurations are efficient, particularly when t is close to 1. (Thus the bounds in Eqs. 2 and 3 are reasonable.) In all but one of the efficient graphs (the exception being the complete graph) the link density can be varied independently of the number of nodes, and thus any desired reliability can be achieved.

An intriguing possibility arises. Is size alone sufficient to guarantee an efficient graph in most cases; that is, is it necessary to intelligently choose a base graph or can the choice be made randomly? A partial answer to this question is obtained from the study of two different ensembles of base graphs. We find that, on the average, the base-graph ensembles are efficient when p is small and t is close to 1. For one of these ensembles, the maximum path length of paths that contribute to the reliability is determined, and a trading curve established among path length, reliability, and link density.

II. BACKGROUND OF OUR STUDY

The literature pertinent to our study of probabilistic graphs and reliable communication networks fits into four classes. The first consists of basic material on discrete probability theory and combinatorial analysis, and is well represented by the first few chapters of books by Feller (1) and Riordan (22). It is assumed that the reader is familiar with this material, which includes the concept of a sample space, the calculation of the probability of the union of nondisjoint events and of the joint occurrence of nonindependent events, and the use in combinatorial analysis of ordinary and exponential generating functions. The other classes, which will be reviewed separately, consist of material on combinatorial graph theory, applications of graph theory to reliability, and general applications of graph theory to electrical engineering.

2.1 COMBINATORIAL GRAPH THEORY LITERATURE

There is a large number of publications on counting problems in graph theory. This material is centered around a very powerful theorem, Pólya's Theorem (20), which relates the generating functions for a "store of objects" to the generating function for distinct, ordered, independent selections of the objects in the store. The distinctness of the selections is specified by a permutation group; that is, two ordered selections containing the same objects are distinct unless there exists a permutation in the permutation group which takes one of the selections into the other. Examples of the use of Pólya's theorem have been given by Riordan (22), and include the enumeration of series-parallel networks and unlabeled and labeled, as well as unconnected and connected, graphs. More examples are given by Harary (8), who enumerates unlabeled and unconnected graphs, and by Ford and Uhlenbeck (4), who count a variety of labeled and unlabeled graphs.

Of greatest significance for our work, however, is the enumeration of connected graphs in which all nodes are labeled. (A labeled graph is one in which some, or all, of the nodes are labeled, that is, given some distinguishing characteristic. Thus, two graphs are counted separately, even though they are topologically equivalent, if the labeled nodes are not in equivalent positions.) Such an enumeration does not require the power of Pólya's theorem. This enumeration problem is solved in a direct manner by Gilbert (5), who obtains a general expression for the number of labeled-node connected graphs, with and without parallel links, and with and without slings. In another paper (see Section III for more lengthy discussion) Gilbert (6) extends the enumeration of connected graphs with no parallel paths and no slings to the exact analysis of the complete graph, a result that was obtained independently by the author (11). In this same paper, Gilbert derives upper and lower bounds, and the asymptotic behavior of the probabilities of strong and weak connectivity, for the complete graph. These large-graph results are of great value in the present research (see Section IV).

2.2 THEORETICAL STUDIES OF RELIABILITY

The classical paper in this field is that of von Neumann (18), who considered methods for increasing the reliability of automata (computers). These computers were assumed to be constructed of identical elements that had finite probabilities of failure. The von Neumann elements are either majority organs or Sheffer stroke organs. A majority organ has three binary inputs and a single binary output. The output is "1" if either 2 or 3 inputs are "1", and is otherwise "0". The Sheffer stroke organ is a two-input, single-output box that realizes the Boolian operation "not A and not B."

Von Neumann suggested the idea of improving the reliability of a system by substituting for each element an array of elements. This array would have the terminal characteristics of an ideal element and a reliability as close to 1 as desired. He presented a synthesis scheme for such arrays and obtained an estimate of the number of elements used in the synthesis.

This work stimulated Moore and Shannon to make a similar study for computers composed of "crummy relays." (A "crummy relay," as defined in our terminology, is a link in a probabilistic graph that has two possible link reliabilities, p_1 and p_2 ($p_1 > p_2$). The link reliability is p_1 when the relay is excited, and is p_2 when the relay is unexcited. For the ideal relay, $p_1 = 1$ and $p_2 = 0$.) The results of their study are presented in a clear, complete, and elegant paper (15). They give a three-step synthesis procedure for obtaining a base graph in which p_{12} is arbitrarily close to 1 for $p = p_1$, and arbitrarily close to 0 for $p = p_2$, regardless of the initial values of p_1 and p_2 ($p_1 > p_2$). This base graph can then be substituted for a "crummy relay" in the computer. An upper bound is obtained on the number of links used in the synthesis of the base graph, and it is shown that, if the initial relays are reasonably good ($p_1 > 3/4$, $p_2 < 1/4$), then the number of links used in this synthesis is only slightly larger than the number of links required by any synthesis to achieve the desired reliability.

In the development of their synthesis procedure, Moore and Shannon show some properties of probabilistic graphs that are of interest in this research and will therefore be repeated. First, they note that it is always possible to write the probabilities P_{ij} and $1 - P_{ij}$ as polynomials in p (node destruction is not considered, that is, $t = 1$). Thus, if A_k is used to denote the number of distinct k-link subsets that contain at least one path between nodes i and j in a given L-link graph, P_{ij} can be written

$$P_{ij} = \sum_{k=1}^{L} A_k p^k (1-p)^{L-k} \tag{4}$$

where $p^k(1-p)^{L-k}$ is the probability of the existence of one of the k-link subsets, but of no other links. Equation 4 is thus the sum of the probabilities of a disjoint set of events.

Alternatively, if B_k is the number of distinct k-link subsets which when erased from the base graph leave no paths between nodes i and j, then $1 - P_{ij}$ can be written

$$1 - P_{ij} = \sum_{k=1}^{L} B_k(1-p)^k p^{L-k} \qquad (5)$$

Such a subset is called a "cutset" of the graph. If the subset also has the property that no proper subset of it is a cutset, then it is referred to as a "minimal cutset." Moore and Shannon also note that if s is the smallest value of k for which A_k does not vanish in Eq. 4, then s is the length of the shortest path between nodes i and j, and A_s is the number of such paths. Likewise, if m is the smallest value of k for which B_k does not vanish in Eq. 5, then m is the number of links in the cutsets with the least number of links, and B_m is the number of such cutsets.

The same reasoning can be used to express the probability of strong connectivity as a polynomial in p. In this case, the first nonvanishing coefficient, C_{n-1}, where n is the number of nodes, is equal to the number of trees in the graph. Since W is the sum of P_{ij}'s taken over all possible pairs (i, j) and divided by the number of pairs, it also has the form of a polynomial.

Another interesting property pointed out by Moore and Shannon is the factoring technique. Let P stand for either the S, W, or P_{ij} of a graph. Choose a particular link in the graph, and let P_1 be the probability (S, W, or P_{ij}) in that graph which results from opening (removing) this link from the original graph. Likewise, let P_2 be the probability in that graph which results from shorting the link (superimposing the nodes at either end of the link). Then $P = qP_1 + pP_2$, and $P_1 \leq P_2$. This factoring technique will be exploited in Section III.

Finally, a very interesting theorem on the slope of P_{ij} as a function of p is presented by Moore and Shannon. Since it applies to P_{ij}, it also must apply to W. Furthermore, the proof holds without change for S. Therefore, we can state the theorem in terms of P. Let P' denote the derivative of P with respect to p.

THEOREM: If P is neither identically zero, identically one, nor identically p, then, for $0 < p < 1$,

$$\frac{P'}{P(1-P)} > \frac{1}{p(1-p)} \qquad (6)$$

This theorem implies that P(p) can equal p for, at most, one value, p_1, and that $P'(p_1) > 1$.

Another paper on the design of reliable computers is worth noting. Kochen (13) considers particular subsystems, "and-circuits," "or-circuits," and "exclusive-or-circuits," constructed of crummy relays, and attempts to increase reliability by introducing redundancy with respect to the whole subsystem. As a criterion, he considers the probability that the system, averaged over all possible inputs, operates correctly. With this criterion, he succeeds in constructing subsystems that use less links than would be necessary if redundancy were employed only on an element basis, as in Moore and Shannon. The saving is, perhaps, a factor of 3.

Another interesting paper on reliability is that of Mine (14). After discussing certain aspects of probabilistic logic, he defines a reliability function, of which our S, W, and P_{ij} are special cases, and determines some of its properties. He then considers minimizing the cost of a system that must operate with a specified reliability. He determines the point for each element in the system at which it is better to increase the reliability of the element (the cost curve is assumed to be known) than to increase the redundancy of the element. His equations are of use in the particular case in which redundancy is achieved by paralleling like elements. Mine also considers some algebraic properties of graphs, which he derives from the incidence matrix.

Another writer who uses the algebraic approach to study probabilistic graphs is Wing (28). His intention, like ours, is to apply the results to the reliability of communication networks. Wing defines a path matrix and determines an inequality on its rank, as well as the conditions under which this inequality becomes an equality. He then introduces a failure matrix and relates its determinant to the probability of failure. Yet it is difficult to see the advantages of this approach, even for the analysis of small graphs.

Moskowitz and McLean (16, 17) attack the problem of reliability in general systems. Parallel and series combinations of links are considered, and the 5-link bridge graph is analyzed. Notation and an algebra for handling probabilistic graphs are suggested, and the factoring theorem is proved. Their papers are rounded out by many simple numerical examples and figures.

2.3 GENERAL APPLICATIONS OF GRAPH THEORY

Recently, three papers that review general applications of graph theory with particular emphasis on electrical network analysis and synthesis have appeared. A lucid account of the use of graph theory in analyzing networks is given by Weinberg (27). He presents a clear proof that a necessary and sufficient condition for a determinant of an incidence matrix, of order n, to correspond to a tree is that it be nonzero.

Reza (21) covers essentially the same ground as Weinberg, but includes material on switching circuits, information networks, and system reliability. However, this material constitutes mainly a partial bibliography.

Finally, Harary (9) discusses graph theory with applications to network theory, flow (capacity) problems, the enumeration and synthesis of Boolian functions, series-parallel networks, probabilistic problems (very briefly), and sequential machines. The problems that he considers are largely combinatorial. His paper is of interest because of the introduction of topological definitions and theorems, and on account of its scope.

III. ANALYSIS AND SYNTHESIS OF SMALL PROBABILISTIC GRAPHS

The exact analysis of a probabilistic graph is, in general, very difficult. For example, the exact probabilities S, W, and P_{ij} must usually be stated as polynomials in p or q, or both, and these polynomials contain nearly L terms, where L is the number of links in the graph. Thus, a graph with 20 links, which is not a very large graph, requires a polynomial with approximately 20 terms to describe it exactly. Such expressions are difficult to obtain, and unwieldy to handle once they are obtained. Techniques for exact analysis are discussed, however, in section 3.1, and explicit formulas are given for the complete graph.

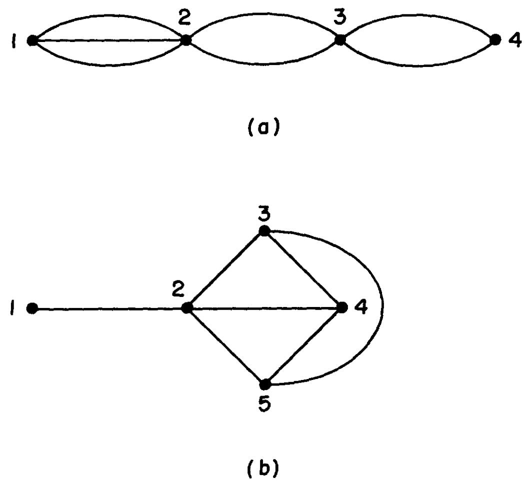


Fig. 5. Two graphs illustrating an incorrect synthesis assumption: (a) 12 trees and 7 links; (b) 16 trees and 7 links.

Likewise, exact synthesis results are difficult to obtain, and one must be exceedingly careful about assumptions. For example, it is not true that if two n-node graphs have the same number of links, the graph with the greater number of trees will have the larger probability of strong connectivity for all p (although it is true for sufficiently small p). Thus, for $t = 1$, the graph in Fig. 5a has a better reliability for p close to 1 than the graph in Fig. 5b, since it remains connected until at least two links, rather than just one, are broken. A synthesis that is valid for a limited range of p is discussed in section 3.2, and an interesting bound is discussed in section 3.3.

3.1 EXACT ANALYSIS OF PROBABILISTIC GRAPHS

The exact analysis of probabilistic graphs, that is, the calculation of S, W, or P_{ij} for a given base graph, can be accomplished in several ways. For example, we might use the method, discussed in Section II (Eqs. 4 and 5), of summing the probabilities of those disjoint link and node sets that have the desired property. Gilbert (6) has applied this technique to obtain the probability of strong connectivity in a complete-base graph directly from an enumeration by number of links of connected, labeled, n-node graphs with no parallel links or slings. The author (11) has also used this technique in one stage of his analysis of the complete graph. In general, this method is difficult to apply, because the problem of counting all possible disjoint events with the desired property (for example, the number of distinct sets of links and nodes for which a given base graph is strongly connected) is rather formidable. If there are m links and nodes, then 2^m distinct sets must be considered (each element is either in or not in a given set).

Rather than perform analysis with disjoint link and node sets, we might consider only the nondisjoint events whose union is the desired event (for example, consider all of the trees of the graph when calculating the probability of strong connectivity, and all of the paths between a desired pair of nodes when calculating the probability of path connectivity). The probability of the union of these nondisjoint events can then be obtained by the inclusion-exclusion technique, in which we add the probabilities of the nondisjoint events occurring separately, subtract the probabilities of the events occurring in all possible pairs, add the probability of events occurring three at a time, and so on. This method also involves a great deal of bookkeeping, since, for k events, there are N probabilities to be added or subtracted, where N is given by

$$N = k + \binom{k}{2} + \binom{k}{3} + \ldots + \binom{k}{k} = 2^k - 1 \tag{7}$$

The question of whether to use disjoint events or inclusion-exclusion thus appears to hinge on whether the number, k, of nondisjoint events that contribute to the desired event is greater or less than the number, m, of links and nodes in the graph. However, since the use of the method of inclusion and exclusion first requires the determination of all of the nondisjoint events, it is often easier to apply the disjoint-event method.

Another approach to analysis is graph factoring, as discussed in section 2.2, for $t = 1$. The extension for $t \neq 1$ is obvious, although it becomes necessary to distinguish in the factored graphs between nodes that have a probability of failure, t, and nodes that are perfectly reliable. In the rest of this discussion, however, we shall use factoring only when $t = 1$, and introduce node destruction separately. Note that if factoring is applied to every link in the graph, the analysis becomes identical with analysis by the summation of disjoint events. But partial factoring of links may require less effort for analysis than the disjoint-event method if some ingenuity is used. The factoring equation is particularly important for proving theorems, and will be so used in sections 3.2 and 3.3.

As we have mentioned, the complete graph has been analyzed exactly for $t = 1$ by the author (11) and by Gilbert (6). The method used by the author involves factoring the complete graph on a subset of its links, and is more involved than the direct derivation of Gilbert. Since both derivations are available, neither will be repeated here. Explicit results for S(n) and W(n), the probabilities of strong and weak connectivity, in the n-node complete graph are:

$$\left.\begin{aligned}
S(0) &= S(1) = 1 \qquad \text{(convention)} \\
S(2) &= 1-q \\
S(3) &= 1-3q^2 + 2q^3 \\
S(4) &= 1-4q^3 - 3q^4 + 12q^5 - 6q^6 \\
S(5) &= 1-5q^4 - 10q^6 + 20q^7 + 30q^8 - 60q^9 + 24q^{10}
\end{aligned}\right\} \tag{8a}$$

$$\left.\begin{aligned}
W(2) &= 1-q \\
W(3) &= 1-2q^2 + q^3 \\
W(4) &= 1-2q^3 - 2q^4 + 5q^5 - 2q^6 \\
W(5) &= 1-2q^4 - 6q^6 + 7q^7 + 12q^8 - 18q^9 + 6q^{10}
\end{aligned}\right\} \tag{8b}$$

Numerical values of $S(n)$ are included in Appendix I for comparison with upper bound expressions.

The results expressed in Eqs. 8a and 8b are for perfect nodes ($t=1$). Because of the symmetry of the complete graph, node failures are simply introduced. Let $S(t, n)$ and $W(t, n)$ be the probabilities of strong and weak connectivity in the n-node graph for $t \neq 1$. Then

$$S(t, n) = \sum_{k=0}^{n} \binom{n}{k} t^k (1-t)^{n-k} \, S(k) \tag{9}$$

$$W(t, n) = \sum_{k=2}^{n} \binom{n-2}{k-2} t^k (1-t)^{n-k} \, W(k) \tag{10}$$

Equations 9 and 10 are obtained by the disjoint-event method. For example, there are $\binom{n}{k}$ possible sets in which k nodes are present and $n - k$ nodes are destroyed. The k nodes that are present form a complete graph that is strongly connected, with a probability $S(k)$. Likewise, there are $\binom{n-2}{k-2}$ possible sets of nodes in which k nodes, including the terminal nodes of the desired path, are present. A path exists between the terminal nodes with probability $W(k) = P_{12}$.

3.2 A SYNTHESIS PROBLEM

An interesting question in probabilistic-graph synthesis is the following (assume node reliability $t = 1$): What arrangement of L-links maximizes the probability of a path between two (terminal) nodes, if no parallel links are permitted? The constraint against parallel links prevents the trivial solution of placing all L-links directly between the terminal nodes. Clearly, the optimum solutions for $L = 1, 2, 3, 4, 5, 6,$ and 7 are

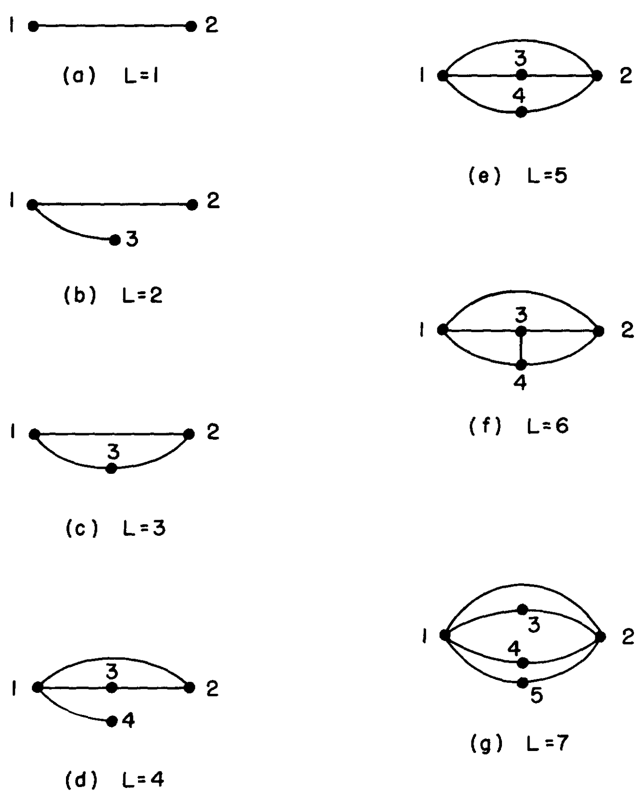


Fig. 6. Optimum L-link graphs for maximizing P_{12}.

those shown in Fig. 6 (a question might arise for $L = 7$, but a simple calculation shows that P_{12} is reduced by removing the link between nodes 5 and 2 and placing it between nodes 3 and 4, or between nodes 4 and 5).

The problem becomes more involved when we consider the 9-link graph. Should an additional parallel path of length 2 (including node 6) be added to the graph in Fig. 6g, or should the two additional links be placed between nodes 3 and 4, and nodes 4 and 5? The calculation of P_{12} for this problem already becomes difficult, but after it is made, it shows the parallel-path arrangement best for all values of link reliability, p. It may now appear plausible that when L is odd, the use of parallel paths of length 2 is optimum. Let us consider this possibility in detail. It can be seen from Eqs. 4 and 5 that for p very small, P_{12} behaves as

$$P_{12} = p + A_s p^s \qquad p \ll 1 \tag{11a}$$

and for p very close to 1, P_{12} behaves as

$$P_{12} = 1 - B_m q^m \qquad q \ll 1 \tag{11b}$$

where s is the length of the shortest path between nodes 1 and 2 (not counting the direct link); A_s is the number of such paths; m is the size of the smallest minimal cutset; and

B_m is the number of such minimal cutsets. Thus, for p close to 0, P_{12} is maximized by making s as small, and A_s as large, as possible. Both conditions are met for L odd by setting s = 2 (only one path of length 1 is allowed) and $A_s = (L-1)/2$, that is, by using the parallel-path graph. Likewise, for p close to 1, P_{12} is maximized by making m as large, and B_m as small, as possible. A little thought will show that the largest m that can be made is $1+(L-1)/2$, and this is attained, for L odd, by using the parallel-path graph. Any other arrangement that satisfies the constraint against parallel paths will have a smaller m, and thus a smaller P_{12} for sufficiently small q, irrespective of the relative sizes of the B_m's. Thus, for very small p and very large p, the parallel-path graph is optimum when L is odd.

The range of p for which the parallel-path graph is optimum is extended by the following theorem.

THEOREM: The parallel-path graph is optimum for $p \geq 1/\sqrt{2}$ when L is odd.

Before proving this theorem, let us consider the possible alternatives to the parallel-path graph. One arrangement would be that illustrated in Fig. 7a, in which every node has a direct connection to both terminal nodes, but in which there are additional "center" links spanning nonterminal nodes. It is obvious that graphs of the form of Fig. 7a are better than those of Fig. 7b. Nevertheless, a general proof of this obvious statement – as evidence of the difficulty of working with probabilistic graphs – is still not available.

The proof of the theorem involves obtaining an upper bound on P_{12} that applies to all possible L-link graphs with h center links, and then showing that this upper bound is less than the P_{12} for the L-link parallel-path graph, when $p \geq 1/\sqrt{2}$. The upper bound is obtained by applying the factoring technique to all h center links of the graph. Thus, P_{12} is expanded in the form

$$P_{12} = \sum_{m=0}^{h} \sum_{z=1}^{\binom{h}{m}} p^m q^{h-m} W_z(m) \qquad (12)$$

where each term $p^m q^{h-m} W_z(m)$ represents the contribution to P_{12} of the disjoint event obtained by shorting a particular set of m center links and opening the remaining h – m center links. (A link is shorted by superimposing its two terminal nodes. A link is opened by removing it from the base graph. See section 2.2.) There are $\binom{h}{m}$ possible ways of selecting these m links from the set of h, and, for convenience, we assign to each way a number 1, 2, ..., $\binom{h}{m}$. In the term considered in Eq. 12, $W_z(m)$ is the probability of a path in the graph obtained by shorting the z^{th} set of m links. Note that the factored graph has the form of Fig. 8a or 8b, if the unfactored graph had the form of Fig. 7a or 7b, respectively.

Let us obtain an upper bound on $W_z(m)$ for the factored graph of the type shown in Fig. 8b (which includes the graph of Fig. 8a as a special case). Consider each of the j paths joining left and right terminal nodes, other than the single link, and number them

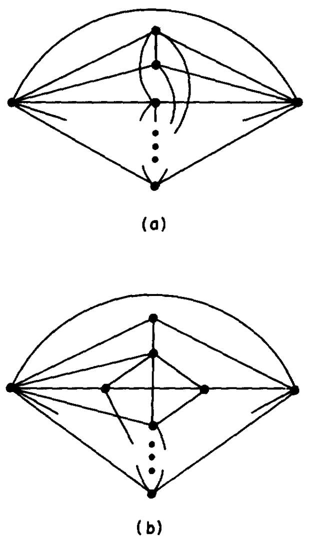


Fig. 7. Possible graph configurations.

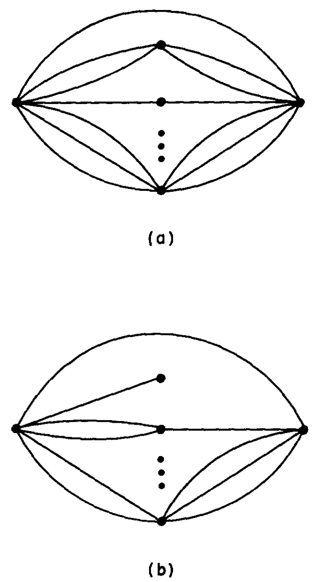


Fig. 8. Graph configurations after factoring.

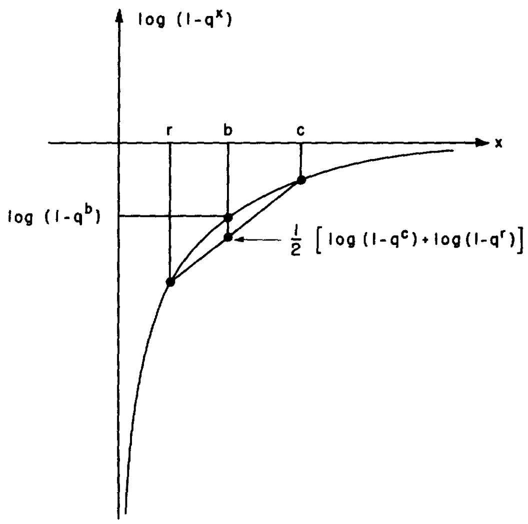


Fig. 9. Inequality resulting from concaveness of log $(1-q^x)$.

from top to bottom 1, 2, ..., j. Let c_1 and r_1 denote the number of links between left and center nodes, and between center and right nodes, respectively, in path $i (1 \leq i \leq j)$. Note that $c_1 + c_2 + \ldots + c_j = C$ and $r_1 + r_2 + \ldots + r_j = R$, where both C and R are independent of the particular factoring considered, and depend only on the total number of links leaving the left and right terminal nodes, respectively.

Now consider the probability, B(c, r), that a path with c left links and r right links is not broken. For convenience, call such a path a (c, r)-path. Clearly, B(c, r) is the product of the probabilities that the left and right terminal nodes are both connected to the center. Thus

$$B(c, r) = (1-q^c)(1-q^r) \tag{13}$$

Now, let $b = (c+r)/2$. We claim that

$$B(c, r) \leq (1-q^b)^2 \tag{14}$$

To prove Eq. 14, it is sufficient to show that the following inequality is true, since the logarithm is a monotonic increasing function of its argument. Therefore

$$\log(1-q^b) \leq \frac{1}{2}[\log(1-q^c) + \log(1-q^r)] \tag{15}$$

However, $\log(1-q^x)$ is a concave downward function of x. This property is sufficient to prove Eq. 15, as can be seen from Fig. 9. Note that the equality holds if and only if $r = c = b$, a case that would exist if the factored graph were of the type shown in Fig. 8a.

Now, let us write the probability, $W_z(m)$, for the factored graph of Fig. 8b. Since a path exists between nodes 1 and 2, unless all connecting paths are broken, it is clear that

$$W_z(m) = 1 - \{q[1-B(c_1, r_1)][1-B(c_2, r_2)] \ldots [1-B(c_j, r_j)]\} \tag{16}$$

Therefore, by substituting from Eq. 14 in Eq. 16, we obtain

$$W_z(m) \leq 1 - \left\{ q \left[1 - \left(1-q^{b_1}\right)^2 \right] \ldots \left[1 - \left(1-q^{b_j}\right)^2 \right] \right\} \tag{17}$$

Performing the squaring operation, and introducing our definitions of R and C, we obtain, from Eq. 17,

$$W_z(m) \leq 1 - \left\{ q^{\frac{R+C}{2}+1} \left[2-q^{b_1}\right]\left[2-q^{b_2}\right] \ldots \left[2-q^{b_j}\right] \right\} \tag{18}$$

Thus, the probability of a path in a factored graph obtained from a particular set of m shorted center links and h − m opened center links is bounded above by the right-hand side of Eq. 18.

Let us obtain a weaker, but more general, upper bound on $W_z(m)$ that depends only on m, and not on z. First, consider the factored graph of Fig. 8b. Note that to obtain a (c, r)-path, it is necessary to use at least $d = [\max(c,r) - 1]$ shorted links. For example, suppose $c \geq r$. The c links from the left terminal are required to connect with c different center nodes in the unfactored graph (on account of the constraint against parallel links), and for these c-nodes to be superimposed in the factored graph, a minimum of c−1 links must be shorted. Thus, $d = c-1$.

We are now in a position to generalize the upper bound on $W_z(m)$. Consider the logarithm of the bracketed terms in Eq. 18:

$$\log\left(2-q^{b_1}\right) + \log\left(2-q^{b_2}\right) + \dots + \log\left(2-q^{b_j}\right) \qquad (19)$$

Again, $\log(2-q^x)$ is a concave downward function. Since the sum of the b_i's is constrained to be B, an argument similar to that for Eq. 15 shows that the sum in Eq. 19 is minimized by making one b_i, say b_1, as large as possible at the expense of all other b_i's. That is, if b_1 can be increased by decreasing, say b_2, then the change should be made, because it decreases the sum in Eq. 19. This change is accomplished by moving a shorted link from the group of nodes forming the b_2-path to the group of nodes forming the b_1-path, and using this additional shorted link to superimpose an additional path on the b_1-group. This process is illustrated in Fig. 10; the graph in Fig. 10b has a better path probability than that in Fig. 10a.

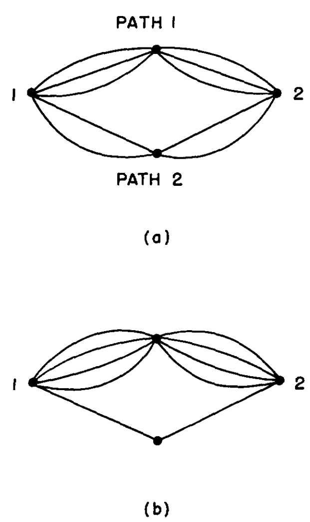


Fig. 10. Increase of P_{12} by exchange of shorted link, $P_{12}(b) > P_{12}(a)$.

The end of this improvement process is reached when all m shorted links have been switched to path 1. From our earlier discussion, the largest that b_1 can then be is $m + 1$, and this is achieved in an $(m+1, m+1)$-path for which $d = (m+1) - 1 = m$. For, if b_1 were any larger, then the minimum required number of shorted links, d, would be greater than the available number m, which, of course, is not possible. Now, $W_z(m)$

is maximized by having all remaining b_i's equal to 1; that is, by having all remaining paths be parallel paths of length 2 [(1, 1)-paths] for which $d = 0$. The number of such paths is obtained by using the constraint equation on the sum of the b_i's.

$$b_1 + b_2 + \dots b_j = B = (R+C)/2 \tag{20a}$$

Thus, if B is an integer, the number of (1, 1)-paths is $B-b_1 = B-(m+1)$. However, if B is not an integer, all b_i's ($i\neq 1$) can not be set equal to 1, because an even number of links is not available. One b_1 must equal 1/2. In either case, the sum $b_2 + b_3 + \dots + b_j$ is equal to $B-(m+1)$. Hence

$$b_2 + b_3 + \dots + b_j = B-(m+1) \tag{20b}$$

Thus, with the use of Eqs. 18 and 20b, and of the fact that $b_1 = m+1$, the upper bound on $W_z(m)$ becomes

$$W_z(m) \leq 1-q^{B+1}[2-q^{m+1}][2-q]^{B-m-1} \tag{21}$$

and this upper bound is independent of z. The upper bound in Eq. 21 can be achieved by a factored graph if the original graph is of the type shown in Fig. 7a, and $m \leq B-1$. This optimal graph has 1 (m+1, m+1)-path and B-(m+1) parallel paths of length 2 [(1, 1)-paths].

We can now substitute Eq. 21 in Eq. 12 and obtain an upper bound on P_{12}:

$$P_{12} \leq \sum_{m=0}^{h} \sum_{z=1}^{\binom{h}{m}} p^m q^{h-m}\{1-q^{B+1}[2-q^{m+1}][2-q]^{B-m-1}\} \tag{22}$$

However, since the terms in the sum are independent of z, Eq. 22 becomes

$$P_{12} \leq \sum_{m=0}^{h} \binom{h}{m} p^m q^{h-m}\{1-q^{B+1}[2-q^{m+1}][2-q]^{B-m-1}\} \tag{23}$$

The summation on m can be performed, and yields as an upper bound

$$P_{12} \leq 1-q^{B+1}(2-q)^{B-h-1}[2(1+pq)^h-q(q+2pq)^h] \tag{24}$$

We can now complete the proof of the theorem by demonstrating that the upper bound given by Eq. 24 is less than the path probability, P'_{12}, in a parallel-path graph with the same number of links, L, when L is odd and $p \geq 1/\sqrt{2}$. An exact expression for P'_{12} is obtained by noticing that a path does not exist only if all parallel paths of length 2 are broken, as well as the single direct link. Therefore

$$P'_{12} = 1-q(1-p^2)^{(L-1)/2} \tag{25}$$

Now, the number of links in the graph with center links is equal to the sum of the direct link, the links touching the right and left nodes, and the h center links. That is,

$$L = 1 + C + R + h = 2B + h + 1 \tag{26}$$

Equation 25 can now be written

$$P'_{12} = 1-q(1-p^2)^{B+h/2} = 1-q^{B+1+h/2}(2-q)^{B+h/2} \tag{27}$$

From Eqs. 24 and 27, we see that P'_{12} is greater than the upper bound on P_{12} if

$$q^{h/2}(2-q)^{h/2} \leqslant (2-q)^{-h-1}\left[2(1+pq)^h - q(q+2pq)^h\right] \tag{28}$$

or

$$(2-q)\left[q(2-q)^3\right]^{h/2} \leqslant 2\left[(1+pq)^2\right]^{h/2} - q\left[(q+2pq)^2\right]^{h/2} \tag{29}$$

Now, inequality 29 is satisfied if the two following inequalities are satisfied:

$$q(2-q)^3 \leqslant (1+pq)^2 \tag{30}$$

$$q(2-q)^3 \geqslant (q+2pq)^2 \tag{31}$$

But inequality 31 is always true, because it can be rewritten $[\text{with } (2-q) = (1+p)]$ as

$$q(1+p)(1+p)^2 \geqslant q(1+p-2p^2)(1+2p) \tag{32}$$

and each of the factors on the left-hand side is greater than or equal to the corresponding factor on the right-hand side. Furthermore, multiplying both sides of Eq. 30, and simplifying, we obtain

$$-(1+2p+p^2) \leqslant -(2+2p-p^2) \tag{33a}$$

or

$$2p^2 - 1 \geqslant 0 \tag{33b}$$

which is satisfied if $p \geqslant 1/\sqrt{2}$. Hence this condition, $p \geqslant 1/\sqrt{2}$, is sufficient to guarantee the inequality in Eq. 29, and thereby the validity of the theorem.

The question might arise as to whether or not it is necessary that Eqs. 30 and 31 both be satisfied, in order that Eq. 29 be true. The answer, of course, is no. However, if the inequality in Eq. 30 is not true (i.e., $p < 1/\sqrt{2}$), then there is a sufficiently large value of h for which inequality 29 is not true. The proof of this statement follows from the fact that if $a > b$, then there is some sufficiently large value of h, so that for any numbers A and B (even $A \ll B$) we can have

$$A\,a^h > B\,b^h \tag{34}$$

Thus, either our upper bound, Eq. 24, is too weak, or better configurations exist than the parallel-path graph, when the number of nodes and links becomes large.

One might hopefully attempt an argument of this kind. Since the parallel-path graph is best for very small p, and also for $p \geqslant 1/\sqrt{2}$, then it must be best for all p. That is, we could argue that the path probabilities exhibited in Fig. 11 are not both possible.

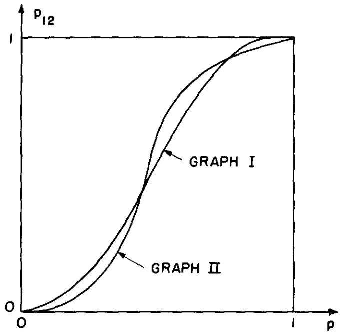


Fig. 11. P_{12} as a function of p for two graphs.

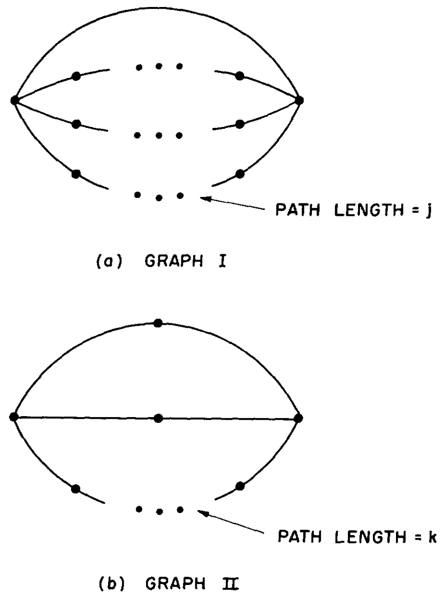


Fig. 12. Possible pair of graphs for obtaining the curves of Fig. 11.

However, this argument is not valid. Two graphs can be constructed whose path probabilities cross as often as desired, and at any desired places. For example, consider two graphs made up of disjoint parallel paths (see Fig. 12). Graph I has one path of length 1 and three paths of length j. Graph II has two paths of length 2 and one path of length $k > 1$. Thus Graph I has a higher value of P_{12} for small p, since the path of length 1 dominates, and has a higher value of P_{12} for large p, because the smallest cutset contains a number of links equal to the number of paths, and the number of paths is greater (by 1) for Graph I. But Graph II can be made to have a higher value of P_{12} for a certain range of p by proper choice of j. Consider the following inequality.

$$(1-p^2)^2 < (1-p)(1-p^j)^3 \tag{35}$$

The left-hand side of the inequality is an upper bound on $(1-P_{12})$ in Graph II, obtained by ignoring the contribution of the k-length path; the right-hand side of the inequality is $(1-P_{12})$ in Graph I. Since $(1-p^2)^2 < (1-p)$ (for $p = 3/4$, for example), it is obvious that j can be made sufficiently large that Eq. 35 is satisfied. Thus, the graphs of Fig. 12 can yield path probabilities P_{12} that behave with p as shown in Fig. 11. Also, the relative number of links between the two graphs has no influence because, if $k = 2$, Graph I has more links than Graph II, whereas if $k = 3j$, Graph II has more links than Graph I, and the previous results are not disturbed in either case. From this discussion, it should be clear that as many crossings can be obtained as desired, at values of p arbitrarily close to specified values, by building graphs with disjoint paths of various lengths containing various numbers of links in parallel.

A final result can now be shown concerning our synthesis problem. We can derive a ratio of the number of center links to the total number of links that must be satisfied if the center-link graph is to have any possibility of being better than the parallel-path graph. This ratio is obtained by using the fact that the probability of a path in a graph with C-links touching the left node, and R links touching the right node, is bounded above by the probability of a path in the graph shown in Fig. 13 (the direct link between the terminal nodes is disregarded in this discussion). The probability of a path in the graph of Fig. 13 is

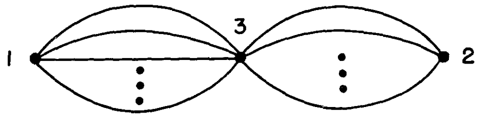


Fig. 13. Graph configuration for upper bound on P_{12}

$$P_{12} = (1-q^C)(1-q^R) \leqslant (1-q^B)^2 \leqslant (1-q^B) \tag{36}$$

in which the inequality of Eq. 14 has been used. A condition under which the upper bound in Eq. 36 is less than the exact P'_{12} for the parallel-path graph is obtained as follows:

$$(1-q^B) \leqslant 1 - (1-p^2)^{(L-1)/2} = P'_{12}$$

if

$$B \log q \geqslant \frac{L-1}{2} \log (1-p^2)$$

or

$$\frac{R+C}{L-1} \geqslant \frac{\log (1-p^2)}{\log q} = 1 - \frac{\log (1+p)}{\log 1/q} \tag{37}$$

Equation 37 can be written in terms of the number of center links, h, with h = (L-1) - (R+C). Then

$$h \leqslant \frac{\log (1+p)}{\log (1/1-p)} (L-1) \tag{38}$$

Thus, for any value of p, there is a definite upper bound on the fraction of links that may be used as center links in an optimum graph, although for p close to zero the fraction approaches 1.

3.3 AN UPPER BOUND ON THE PROBABILITY OF STRONG CONNECTIVITY

The form of the graph-factoring equation, stated in section 2.2 and repeated in Eq. 39, suggests the possibility of obtaining an upper bound on the various connectivity probabilities in terms of a difference equation (when t = 1). In this section the difference equation is solved with boundary conditions appropriate to S, the probability of strong connectivity [cf. (12)]. Properties of the solutions are determined, and an alternative derivation is made.

The graph-factoring equation is

$$S = qS_1 + pS_2 \tag{39}$$

where S_1 and S_2 are the probabilities of strong connectivity in graphs resulting from opening and shorting a particular link in the original base graph. Now, consider all possible n-node, L-link graphs. Since there are only a finite number, S must take on a maximum value among them. Call this maximum, $P_1(n, L)$. We shall obtain an upper bound, $P_2(n, L)$, so that for all n, L,

$$P_1(n, L) \leq P_2(n, L) \tag{40}$$

Let us apply Eq. 39 to an optimum graph, which has $S = P_1(n, L)$, to obtain an upper bound on P_1.

$$P_1(n, L) \leq q\, P_1(n, L-1) + p\, P_1(n-1, L-1) \tag{41}$$

The right-hand side of Eq. 41 may be larger than the left-hand side for two reasons. First, the opening of a link in an optimum n-node, L-link graph may not result in an optimum n-node, (L-1)-link graph, and so the use of $P_1(n, L-1)$ is optimistic. Second, the shorting of a link in the n-node L-link graph may result in a graph with less than L-1 links if and only if there are links in parallel with the shorted link. Thus, the assumptions that L-1 links remain, and that the arrangement of the L-1 links among the remaining n-1 nodes (two nodes have been superimposed) is optimum, are both optimistic.

The upper bound on $P_1(n, L)$ in Eq. 41 is further enhanced if we substitute P_2 for P_1 on the right-hand side. We then have

$$P_1(n, L) \leq q\, P_2(n, L-1) + p\, P_2(n-1, L-1) \tag{42}$$

From Eq. 42, we see that Eq. 40 is satisfied if we define $P_2(n, L)$ as the right-hand side of Eq. 42. Thus

$$P_2(n, L) = q\, P_2(n, L-1) + p\, P_2(n-1, L-1) \tag{43}$$

Equation 43 thus defines $P_2(n, L)$ as a difference equation. In order to solve explicitly for $P_2(n, L)$, it is necessary to introduce boundary conditions. The boundary conditions that we shall use are:

$$P_2(2, L) = P_1(2, L) = 1 - q^L \tag{44}$$

$$P_2(n, n-1) = P_1(n, n-1) = p^{n-1} \tag{45}$$

In Eq. 44, $P_1(2, L)$ is obtained from the fact that the optimum arrangement of L-links among two nodes is achieved by placing all links in parallel (clearly, slings do not help). Then the graph is strongly connected if at least one link is present. Equation 45 is obtained from the fact that if only n-1 links are present in an n-node graph, then the graph can be strongly connected if and only if all links are present.

Equation 43 can now be solved with the help of the boundary conditions, Eqs. 44 and 45. The solution is

$$P_2(n, L) = p^{n-1} \sum_{k=0}^{L-n+1} \binom{n+k-2}{k} q^k; \quad n \geq 2, \; L \geq n-1 \tag{46}$$

First, note that Eq. 46 satisfies both boundary conditions. Then, to show that Eq. 46 satisfies Eq. 43, substitute it in Eq. 43 and simplify. This procedure yields

$$p^{n-1} \sum_{k=0}^{L-n+1} \binom{n+k-2}{k} q^k = p^{n-1} \sum_{k=0}^{L-n+1} \left[\binom{n+k-3}{k-1} + \binom{n+k-3}{k}\right] q^k \tag{47}$$

where, by convention, $\binom{n}{-1} = 0$. Equation 47 is an identity if

$$\binom{n+k-2}{k} = \binom{n+k-3}{k-1} + \binom{n+k-3}{k} \tag{48}$$

which is a well-known recursion for binomial coefficients [see, for example, Riordan (23)].

Several properties of $P_2(n, L)$, as defined explicitly by Eq. 46, are of interest. First, we should note that as L tends to infinity, for any fixed n, the sum on the right-hand side of Eq. 46 approaches $(1-q)^{-(n-1)}$, and hence $P_2(n, L)$ tends to 1 from below.

Next, note that the coefficients, $a_k(n)$, of q^k can be obtained rather simply from Eq. 48. Table 1 is constructed by writing for each entry the sum of the entry directly above it and the entry directly to the left of it.

We shall now determine the conditions under which $P_2(n, L)$ is an attainable upper bound; that is, the conditions for which $P_2(n, L) = P_1(n, L)$. First, recall that if $P_2(n, L)$ is actually the probability of connectivity in an n-node graph, then the coefficient of p^{n-1} in a Taylor-series expansion of $P_2(n, L)$ about $p = 0$ must equal the number of trees

Table 1. Coefficients of q^k in Eq. 46.

n	$a_0(n)$	$a_1(n)$	$a_2(n)$	$a_3(n)$	$a_4(n)$	$a_5(n)$	$a_6(n)$
2	1	1	1	1	1	1	1
3	1	2	3	4	5	6	7
4	1	3	6	10	15	21	28
5	1	4	10	20	35	56	84
6	1	5	15	35	70	126	210
7	1	6	21	56	126	252	462

in the graph. From Eq. 46, we find that this coefficient actually equals $\binom{L}{n-1}$. In other words, if $P_2(n, L)$ is the probability of strong connectivity of a graph, then, in that graph, every possible set of n-1 links must be a tree. This is not possible if the number of links is greater than the number of nodes. That is,

$$P_1(n, L) < P_2(n, L); \qquad L > n > 2$$

Another bound that requires that every set of n-1 links be a tree is obtained as follows. First, S can be written as a sum of disjoint events, as in Section II.

$$S = \sum_{j=n-1}^{L} C_j\, p^j\, q^{L-j} \tag{49}$$

when C_j is the number of distinct sets of j-links in the graph that contains at least 1 tree. Clearly, $C_j \leqslant \binom{L}{j}$, since $\binom{L}{j}$ is the total number of distinct sets of j-links. Furthermore, $C_j = \binom{L}{j}$ if and only if every set of n-1 links forms a tree. Thus

$$S \leqslant \sum_{j=n-1}^{L} \binom{L}{j} p^j q^{L-j} = A(n, L) \tag{50}$$

We shall now show that the bound given in Eq. 50, A(n, L), is identical with $P_2(n, L)$. It is sufficient to show that A(n,L) satisfies the boundary conditions, Eqs. 44 and 45, and also the difference equation, Eq. 43, since this amounts to a proof by induction (that is, the difference-equation solution is unique). Actually, A(n, L) might have been given as the explicit form of $P_2(n, L)$, rather than Eq. 46. However, the actual order of discovery is followed here, because it demonstrates how the properties of $P_2(n, L)$ lead directly to A(n, L).

First, let us show that A(n, L) satisfies the boundary conditions. We have

$$A(2, L) = \sum_{j=1}^{L} \binom{L}{j} p^j q^{L-j} = 1 - q^L \tag{51a}$$

$$A(n, n-1) = \binom{n-1}{n-1} p^{n-1} q^o = p^{n-1} \tag{51b}$$

Next, consider A(n, L+1). We have

$$A(n, L+1) = \sum_{j=n-1}^{L+1} \binom{L+1}{j} p^j q^{L+1-j} = \sum_{j=n-1}^{L+1} \left[\binom{L}{j}+\binom{L}{j-1}\right] p^j q^{L+1-j} \tag{52}$$

in which the second equality results from the recursion of Eq. 48. Breaking the right-hand side of Eq. 52 into two terms, we have

$$A(n, L+1) = \sum_{j=n-1}^{L} \binom{L}{j} p^j q^{L+1-j} + \sum_{i=n-2}^{L} \binom{L}{i} p^{i+1} q^{L-i} \tag{53}$$

In the first term on the right-hand side of Eq. 53 we have used $\binom{L}{L+1} = 0$, and in the second term, we have substituted i+1 for j. Equation 53 thus states that

$$A(n, L+1) = q\,A(n, L) + p\,A(n-1, L) \tag{54}$$

which is the desired difference equation.

We have now proved the interesting identity for the tail of a binomial distribution. Note that

$$\sum_{j=n-1}^{L} \binom{L}{j} p^j q^{L-j} = p^{n-1} \sum_{j=0}^{L-n+1} \binom{n-2+j}{n-2} q^j \tag{55}$$

The effect on the tail of a change in n is easily obtained from the left-hand side of Eq. 55, but not from the right-hand side, while the effect of a change in L is evident from the right-hand side but not from the left-hand side. Thus, the two forms complement each other.

The usefulness of the bound $P_2(n, L)$ is greatly increased because of the equality

$$P_2(n, L) = I_p(n-1, L-n+2) \tag{56}$$

The function $I_x(p, q)$ has been tabulated by Pearson (19). The function $P_2(n, L)$ is plotted in Fig. 14 for a 6-node graph when p = 0.7. For comparison, an upper bound on S, which will be derived in Section IV,

$$S \leq 1 - nq^{2L/n}\left[1 - \frac{n-1}{2} q^{2L(1-(1/n-1))/n}\right] \tag{57}$$

is also plotted in Fig. 14. In Appendix I, several values of both bounds are tabulated. From these tabulations, it is clear that $P_2(n, L)$ is a tighter bound for small graphs. It will now be shown that $P_2(n, L)$ is a poorer bound for sufficiently large graphs. From Eq. 50,

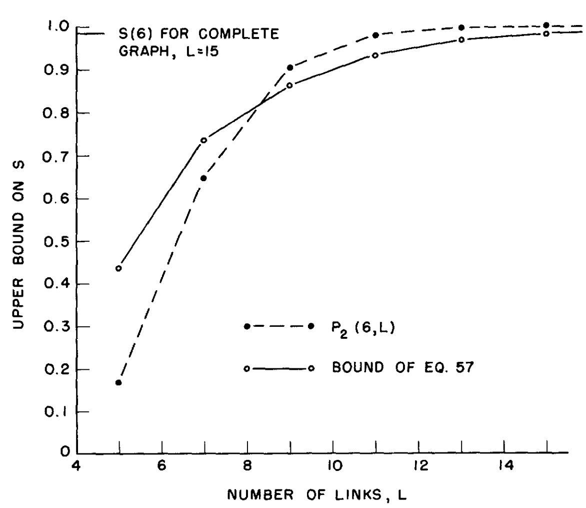


Fig. 14. Upper bounds on S in a 6-node graph (p=0.7).

$$P_2(n, L) = 1 - \sum_{j=0}^{n-2} \binom{L}{j} p^j q^{L-j} \geq 1 - \binom{L}{n-2} p^{n-2} q^{L-n+2} \tag{58}$$

where the inequality is obtained by throwing away all of the (positive) terms in the sum for $j < n - 2$. It can be shown, by using methods that will be introduced in Section IV, that

$$\binom{L}{n-2} \geq e^{n-2} \left(\frac{L}{n}\right)^{n-2}$$

Thus, Eq. 58 can be written

$$P_2(n, L) \geq 1 - \left(\frac{eLp}{q}\right)^{n-2} q^L \tag{59}$$

It is clear that for large L, the lower bound on $P_2(n, L)$ in Eq. 59 becomes larger than the right-hand side of Eq. 57, and thus it is established that $P_2(n, L)$ is a poorer bound for large graphs.

IV. RELIABILITY AND LINK DENSITY IN LARGE PROBABILISTIC GRAPHS

This study of large probabilistic graphs is directed toward a solution of the twofold problem: What link density is necessary and what link density is sufficient to provide a desired reliability in a probabilistic graph suffering from both link and node failures, when either S or W is used as the measure of reliability? In solving this problem, we are greatly interested in determining base-graph configurations that utilize link densities that are only slightly larger than the necessary density for achieving the specified reliability. Such base graphs are called "efficient." Note that the existence of an efficient base graph provides a demonstration that the necessary condition is realistic.

The link density, d, defined as $d = L/n$, is introduced as the controlled parameter, rather than the gross number of links, L, to compensate for the inherent linear growth of L with n. For example, the addition of one node to a graph requires the addition of one link if the node is not to be left a separate (disconnected) part.

This study first concerns itself with determining the necessary density. Then, various base graphs are examined and the efficient ones noted. The link densities of these efficient graphs are the desired sufficient densities. Finally, in Section V, random procedures are used to construct two base-graph ensembles, and the average sufficient densities are derived.

4.1 NECESSARY LINK DENSITY

To determine the link density that is necessary for achieving a desired reliability when the probability of strong connectivity, S, is the criterion, we must first obtain an upper bound on S that is applicable to large graphs. The following idea is basic to the derivation of this bound. A graph can be strongly connected if and only if every existing node has at least one existing link attached to it, or, stated more tersely, if "every node has a link." This idea is used to obtain a bound on S in two alternative forms, the first being weaker than the second but having application to a type of graph that has not previously been considered, the directed-link probabilistic graph.

The two forms of the bound arise because of the nonindependence of the events, "node i has a link," for $i = 1, 2, \dots, n$. For example, if node 1 and node 2 share a common link (the link $\overline{12}$), then the probability that node 1 has a link is not independent of the probability that node 2 has a link. In fact, the conditional probability that node 2 has a link, given that node 1 has a link, is strictly greater than the unconditional probability that node 2 has a link. Therefore, ignoring this condition when calculating the probability of the joint event is not acceptable, because it weakens the upper bound.

One solution to this problem is the introduction of directed links (10). A directed link is an ordered pair of nodes, $\overline{n_1n_2}\ (\neq \overline{n_2n_1})$, and is represented pictorially as a line with an arrow pointing from node n_1 to node n_2. Physically, a directed link might be a one-way radio or telephone channel. A path from node n_1 to node n_2 in a directed-link graph is defined as a set of links $(\overline{n_1n_3}, \overline{n_3n_4}, \dots, \overline{n_jn_2})$ in which all n_i's are different.

A directed-link graph is strongly connected if a path exists between every ordered pair of nodes. All directed links are assumed to have the same reliability, and failures are assumed to be statistically independent.

The (directed link) upper bound on the probability of strong connectivity of any L-link n-node undirected-link graph is obtained as follows. Replace each of the L undirected links (of reliability p) by a pair of oppositely oriented directed links of reliability $\sqrt{p}$, and obtain thereby a directed-link graph with 2L links. We claim that all of the connectivity probabilities are larger in the directed-link graph than in the undirected-link graph which it replaces. The proof follows from the observation that the two oppositely oriented directed links form a two-way path with probability $(\sqrt{p})^2 = p$, and, in addition, form one-way paths with probability $2\sqrt{p}(1-\sqrt{p})$. If the one-way paths are disregarded, the connectivity probabilities become identical in the two graphs.

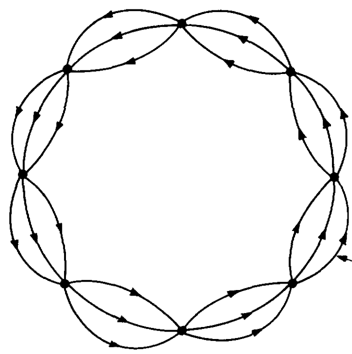


Fig. 15. Directed-link ring graph.

It is now a simple matter to calculate the upper bound on S for the directed-link, and hence the undirected-link, graph. We require that every intact node have at least one intact "outgoing" link (an "outgoing" link is a link directed away from the node). Note that no conditional probabilities need now be considered, since an outgoing link from one node cannot be outgoing from any other node. The calculation proceeds as follows. Label the n-nodes with integers 1, 2, ..., n. If L_i denotes that number of links outgoing from the i^{th} node, the probability that node i is intact while all L_i links are destroyed is tg^{L_i}, where $g = 1 - \sqrt{p}$. Thus the upper bound can be written

$$S \leq \left(1-tg^{L_1}\right)\left(1-tg^{L_2}\right)\ldots\left(1-tg^{L_n}\right) \qquad (60)$$

In accordance with Eq. 14, the right-hand side of Eq. 60 is maximized by setting $L_1 = L_2 = \ldots = L_n = 2L/n = 2d$, where d is the link density in the original (undirected-) link graph. Thus,

$$S \leq \left(1-tg^{2d}\right)^n \qquad (61)$$

When t = 1 the bound on S in Eq. 61 can actually be achieved for directed-link graphs. Consider the configuration (called a "ring graph") illustrated in Fig. 15. A necessary and sufficient condition for this directed-ring graph to be strongly connected is that every node have an intact outgoing link, which is precisely the condition used to obtain Eq. 61. However, this configuration cannot be achieved by the (previously discussed) substitution of directed links in an undirected-link graph. Also, if node destruction is allowed ($t \neq 1$), the ring graph becomes highly undesirable, since failure of one node is sufficient to prevent the graph from being strongly connected.

Equation 61 can be put into more convenient form by using the inequality $(1-x)^n \leq 1 - \frac{nx}{2}$, if $nx \leq 1$ (a direct application of Taylor's theorem). Since we are interested in the bound only when $S > 1/2$,

$$S \leq 1 - \frac{nt}{2} g^{2d} \doteq 1 - \frac{nt}{2}(1-\sqrt{p})^{2d} \qquad (62)$$

The alternative bound on Eq. 62 is obtained, without introducing directed links, by use of a Bonferroni inequality [see Feller (2)]. This method is a generalization of the technique used by Gilbert (6) to obtain an upper bound on the probability of strong connectivity of the complete-base graph. Consider the case $t = 1$. Let E_i be the event that the i^{th} node has no intact links attached to it. The union of the events E_i, $i = 1, 2, \ldots, n$, is the probability that at least one node have no attached link, and hence is 1 minus the desired upper bound on S. Although, as we have mentioned, the events are not disjoint, the Bonferroni inequality gives us our desired bound. If $\Pr\{x\}$ is the probability of event x, and $\bigcup$ and $\bigcap$ stand for union and intersection, respectively, the bound can be written

$$1 - S = \Pr\left\{\bigcup_{i=1}^{n} E_i\right\} \geq \sum_{i} \Pr\{E_i\} - \sum_{\substack{i,j \\ i<j}} \Pr\{E_i \cap E_j\} \qquad (63)$$

Now, if L_i denotes the number of links terminating on the i^{th} node, and q is the probability of link failure, then,

$$\Pr\{E_i\} = q^{L_i} \qquad (64)$$

Furthermore, if a_{ij} denotes the number of parallel links placed directly between nodes i and j, $\Pr\{E_i \cap E_j\}$ can be written as

$$\Pr\{E_i \cap E_j\} = q^{L_i + L_j - a_{ij}} \qquad (65)$$

Introducing Eqs. 64 and 65 in Eq. 63, we have

$$S \leq 1 - \sum_{i=1}^{n} q^{L_i} + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} q^{L_i + L_j - a_{ij}} \qquad (66)$$

We now desire to maximize the right-hand side of Eq. 66 by varying the L_i's and a_{ij}'s, subject to the two following constraints:

$$\sum_{i=1}^{n} L_i = 2L \qquad (67a)$$

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij} = L \qquad (67b)$$

The first constraint has already been discussed, while the second constraint results from counting each link exactly once when we sum the a_{ij}'s. There is one further constraint between the a_{ij}'s and the L_i's. The number of links in parallel between two nodes cannot exceed the number terminating on either of the nodes. Hence

$$a_{ij} \leq \min\{L_i, L_j\} \tag{68}$$

The constraint in Eq. 68, as well as the constraints that the L_i's and a_{ij}'s be positive, will not be used in performing the maximization. It will be found, however, that the absolute maximum does satisfy these constraints. By using the Lagrange method of undetermined multipliers, we want to obtain a stationary point of the quantity G. We have

$$G = 1 - \sum_{i=1}^{n} q^{L_i} + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} q^{L_i+L_j-a_{ij}} + \lambda\left[2L - \sum_{i=1}^{n} L_i\right]$$

$$- \mu\left[L - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} a_{ij}\right] \tag{69}$$

Taking the partials of G with respect to L_i and a_{ij}, and setting them equal to zero, we obtain the system of equations:

$$q^{L_i}\left[1 - \sum_{\substack{j=1 \\ j\neq i}}^{n} q^{L_j-a_{ij}}\right] = \lambda/\log q \qquad i = 1, \ldots, n \tag{70}$$

$$q^{L_i}\left[q^{L_j-a_{ij}}\right] = \mu/\log q \qquad \begin{matrix} i, j = 1, \ldots, n \\ i \neq j \end{matrix} \tag{71}$$

Substituting Eq. 71 in Eq. 70, we obtain

$$q^{L_i} - (n-1)\,\mu/\log q = \lambda/\log q \qquad i = 1, \ldots, n \tag{72}$$

Clearly, for $n > 2$, the only solution of the set of equations (Eqs. 72) is L_i equals a constant. Therefore, by the constraint of Eq. 65,

$$L_i = \frac{2L}{n} = 2d \tag{73}$$

Likewise, if L_i is constant, the set of equations (Eqs. 71) demands that a_{ij} be a constant, or, from Eq. 67b, we have

$$a_{ij} = \frac{2L}{n(n-1)} = \frac{2d}{n-1} \tag{74}$$

Since L_i and a_{ij}, as specified in Eqs. 73 and 74, are positive and satisfy Eq. 68, they form an acceptable stationary point. It can be shown that this stationary point is a maximum.

Thus, if we substitute Eqs. 73 and 74 in Eq. 66, we obtain

$$S \leq 1 - n\, q^{2d}\left[1 - \frac{n-1}{2}\, q^{2d}\left(1 - \frac{1}{n-1}\right)\right] \tag{75}$$

In all cases of interest $S > \frac{3}{4}, n \gg d$, the term in square brackets is greater than 1/2, and

$$S \leq 1 - \frac{n}{2}\, q^{2d} \tag{76}$$

Equation 76 is the desired upper bound for $t = 1$. If $t \neq 1$, the bound on $1 - S$ given by Eq. 63 must be modified. It is necessary to consider all possible distinct sets of intact and destroyed nodes. For each set, a term similar to the right-hand side of Eq. 63, but which applies only to intact nodes, is formed and then weighted by the probability that the set occurs. The sum of the weighted terms over all possible sets of intact nodes is the new bound on $1 - S$. The maximization is carried out in a manner identical with the maximization of Eq. 66, and yields the following conditions on the stationary point:

$$t\, q^{L_i}\left[1 - t \sum_{\substack{j=1 \\ j \neq i}}^{n} q^{L_j - a_{ij}}\right] = \lambda/\log q \qquad i = 1, \ldots, n \tag{77}$$

$$t^2\, q^{L_i}\left[q^{L_j - a_{ij}}\right] = \mu/\log q \qquad i, j = 1, \ldots, n \quad i \neq j \tag{78}$$

It is clear from a comparison of Eqs. 77 and 78 with Eqs. 70 and 71 that the maximum is obtained by setting L_i and a_{ij} constant. Now, consider the sum over all possible sets of intact nodes. Each set with j intact nodes makes an identical contribution to the bound because of the maximization. The probability that the j nodes each have at least one link intact is given by Eq. 76, with n replaced by j. Thus,

$$S \leq \sum_{j=0}^{n} \binom{n}{j} t^j (1-t)^{n-j}\left[1 - \frac{j}{2}\, q^{2d}\right] \tag{79}$$

Equation 79 can be rewritten as

$$S \leq 1 - \frac{q^{2d}}{2} \sum_{j=0}^{n} \binom{n}{j} t^j (1-t)^{n-j}\, j \tag{80}$$

and the summation performed by use of the identity

$$\sum_{j=0}^{n} \binom{n}{j} t^j (1-t)^{n-j}\, j = \frac{d}{dz}\left[\sum_{j=0}^{n} \binom{n}{j} (tz)^j (1-t)^{n-j}\right]_{z=1} \tag{81}$$

Hence

$$S \leq 1 - \frac{tn}{2}\, q^{2d} \tag{82}$$

It is clear that Eq. 82 is a tighter bound than Eq. 64 because $q > 1 - \sqrt{p}(q=(1-\sqrt{p})(1+\sqrt{p}))$. Equation 82 can be put into a form that exhibits the interplay between n and d more clearly. This is

$$S \leq 1 - e^{-[d(2\log 1/q) - \log tn + \log 2]} \tag{83}$$

The bound in Eq. 83 specifies the density, d, that is necessary to obtain a desired probability of strong connectivity. It is clear that the necessary density is a linear function of the logarithm of the number of nodes.

A necessary condition on the density will now be obtained, with the use of W, the probability of weak connectivity, as a criterion of reliability. A minimum condition for a path to exist between nodes i and j is that the two nodes be intact, and that each have at least one intact link. Since W is an average of the P_{ij}'s over all pairs of nodes, an exact use of the principle of inclusion-exclusion gives

$$W = \frac{1}{\binom{n}{2}} \sum_{\substack{i,j \\ i \neq j}} P_{ij} \leq \frac{t^2}{\binom{n}{2}} \sum_{\substack{i,j \\ i \neq j}} \left(1-q^{L_i}-q^{L_j}+q^{L_i+L_j-a_{ij}}\right) \tag{84}$$

where the definitions and constraints on L_j and a_{ij} are the same as before. The maximization of the right-hand side of Eq. 84 is similar to the previous maximization, and, again, results in L_i and a_{ij} being constants. Thus

$$W \leq t^2[1-2q^{2d}+q^{2d[2-(1/n)]}] \tag{85}$$

For n and d large, the third term on the right-hand side of Eq. 85 is small compared with the second, and

$$W \leq t^2(1-q^{2d}) \tag{86}$$

It is clear that the density that is necessary to achieve a desired value of W, as given by Eq. 86, is independent of the number of nodes in the graph. We are now in a position to examine various graph configurations and determine those that are efficient.

4.2 SUFFICIENT LINK DENSITIES

We stated in section 3.1 that the complete graph is an optimum configuration of n nodes and $\binom{n}{2}$ links. Let us consider its efficiency. Gilbert (6) shows that the asymptotic behaviors of S and W in the complete graph, for t = 1, are:

$$S \sim 1 - nq^{n-1} \tag{87a}$$

$$W \sim 1 - 2q^{n-1} \tag{87b}$$

The symbol ~ is used hereafter with a special meaning: $P \sim A$ indicates P is either asymptotically equal to A, or P is asymptotically equal to $B > A$. Such a bound on P (equal S or W) is sufficient to give us the link density for which P equals or surpasses a desired reliability, A, when the graph is large.

The behavior of S and W for $t \neq 1$ can be obtained by using Eqs. 9 and 10. Thus

$$S \sim \sum_{i=0}^{n} \binom{n}{i} t^i (1-t)^{n-i} \left[1 - iq^{i-1}\right] \tag{88}$$

$$W \sim \sum_{i=2}^{n} \binom{n-2}{i-2} t^i (1-t)^{n-i} \left[1 - 2q^{i-1}\right] \tag{89}$$

Performing the summations in Eqs. 88 and 89, and making use of the fact that in a complete graph the link density is $d = \frac{n-1}{2}$, we obtain

$$S \sim 1 - e^{-(2d \log 1/q - \log tn)} \tag{90}$$

$$W \sim t^2 \left[1 - 2q(1-tp)^{2d-1}\right] \tag{91}$$

Therefore, if S is the criterion of reliability, the complete graph is efficient for all values of t, whereas if W is the criterion, the complete graph is efficient only for $t = 1$. However, the lack of efficiency of the complete graph for $t \neq 1$ is the fault of the bound rather than the configuration. (Attempts to improve the bound in Eq. 86 by requiring that both terminal nodes have an intact link with an intact node at the far end have so far been unsuccessful. Note that for this bound, paths of length 1, 2, and greater than 2 must be distinguished.) The complete graph is found wanting in one important respect. Once the number of nodes is chosen, the reliability is fixed. It is not possible to decrease the link density and accept the smaller reliability specified by Eqs. 90 and 91.

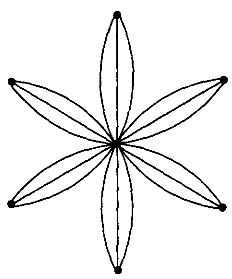


Fig. 16. Star graph.

A graph in which link density can be varied is the ring graph, which is obtained from the directed-link ring graph of Fig. 15 by replacing every directed link with one undirected link. The resulting graph is strongly connected unless two or more of the n sets of d parallel links are completely destroyed. Thus, if $t = 1$,

$$S = (1-q^d)^n + n(1-q^d)^{n-1} q^d \tag{92}$$

For S close to 1, Eq. 92 behaves as

$$S \sim 1 - e^{-(2d \log 1/q - 2 \log n)} \tag{93}$$

Therefore the density required by this (ring) graph is approximately 50 per cent greater than the necessary density if the exponent is of the order of log n. Note, however, that the ring graph has several disadvantages. The upper bound on path length necessary to join two nodes is n/2, which is rather large (half the stations in the equivalent communication networks may have to handle a given message). The configuration is extremely vulnerable to node destruction. Finally, if the reliability criterion is W, it is clear that the necessary density will be a function of n, and hence the graph is inefficient. (It can be shown, for example, that in a ring graph $W \sim 1 - 1/6 \exp[-(2d \log 1/q - 2 \log n)]$.)

Another obvious configuration is the "star graph" illustrated in Fig. 16. Such a configuration might be appealing in a physical situation in which one ultrareliable node is available. Note that the maximum path length is 2. However, the graph is not particularly efficient for either reliability criterion, even when t = 1, since, for n and d large,

$$S = (1-q^d)^n \sim 1 - e^{-(d \log 1/q - \log n)} \tag{94}$$

$$W = (1-q^d)^2 \sim 1 - 2q^d \tag{95}$$

Thus the star graph requires a density twice that of the necessary density to achieve a given reliability with either criterion.

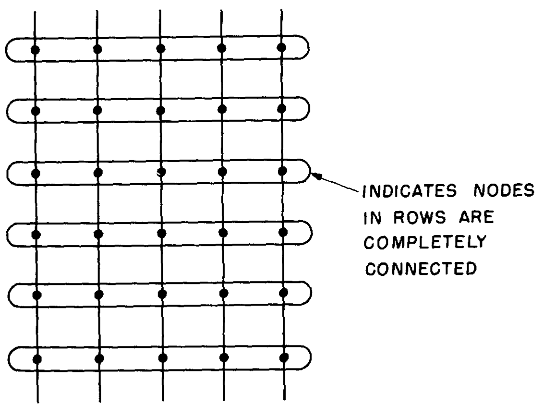


Fig. 17. Rectangular graph.

A more interesting structure, which has application when $t \neq 1$, is the rectangular graph illustrated in Fig. 17. Notice that each row forms a complete graph. If c is the number of columns (and thus the number of nodes in each complete graph), the link density d is $1 + (c-1)/2$. Therefore, for d large, the number of columns is 2d. Now, a lower bound on the probability of strong connectivity is obtained if one considers contributions from only those graphs in the ensemble for which every row is strongly connected (this is a convenient, but not a necessary, condition). Since there are n/2d rows, the probability, R, that every row will be strongly connected is, from Eq. 90,

$$R \sim [1-t(2d)q^{2d}]^{n/2d} \tag{96}$$

For R close to 1, this becomes

$$R \sim 1 - \left(\frac{n}{2d}\right) (t)(2d)q^{2d} = 1 - e^{-(2d \log 1/q - \log tn)} \tag{97}$$

If the rectangular graph is to be strongly connected, the various strongly connected rows must also be connected. Note that the probability of connecting the rows is identical with the probability of strong connectivity in a ring graph (think of each row as a node) in which the link reliability equals t^2p (a link can connect two rows only if it exists and the nodes at either end of it exist). Thus, if C is the probability that the n/2d rows are connected, we have from Eq. 93,

$$C \sim 1 - e^{[4d \log (1-pt^2) - 2 \log n + 2 \log 2d]} \tag{98}$$

where n/2d has been substituted for n, $1 - pt^2$ for q, and 2d for d. Finally, the probability of strong connectivity in a rectangular graph for S close to 1 and n large, is

$$S \geqslant R \times C \sim 1 - t\, n\, q^{2d} - \left(\frac{n}{2d}\right)^2 (1-pt^2)^{4d} \tag{99}$$

in which the nonexponential forms of Eqs. 97 and 98 have been used. For $t = 1$, the third term on the right-hand side of Eq. 99 becomes small compared with the second (it is $1/4d^2$ multiplied by the second term squared), and we have

$$S \sim 1-n\, q^{2d} = 1 - e^{-(2d \log 1/q - \log n)} \tag{100}$$

Thus, for $t = 1$, the rectangular graph is efficient. The sufficient density is just the necessary density. Furthermore, the density can be varied at will, provided, of course, that n is large and S is close to 1 (so that Eq. 100 is valid).

Let us now consider the behavior of the rectangular graph when $t \neq 1$. By straightforward algebraic procedures, it can be shown that $(1-pt^2)^2$ is greater than q if t is less than a critical value, $T = 1/(1-\sqrt{q})^{1/2}$. If t is greater than T, the second term in Eq. 99 still predominates and the graph is efficient. However, if t is less than T, then for sufficiently large densities, the third term of Eq. 99 dominates the second, and S becomes

$$S \sim 1 - \left(\frac{n}{2d}\right)^2 (1-pt^2)^{4d} = 1 - e^{-[2d\alpha(p,t) - 2\log n + 2\log 2d]} \tag{101}$$

where

$$\alpha(p,t) = -2 \log (1-pt^2) < \log 1/q \tag{102}$$

Equation 101 shows that for $t < T$ and d large, the graph is no longer as efficient as it was previously. The graph may still be considered efficient if the inequality in Eq. 102 is close to equality.

The rectangular graph is not an efficient graph to use if W is the criterion. Furthermore, path lengths grow in a direct proportion to the number of nodes in the graph.

One more graph will now be considered, the j^{th}-order hierarchical graph. See

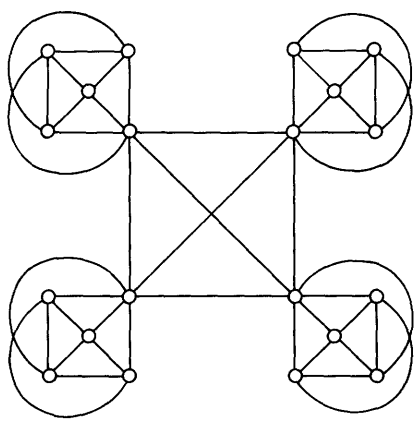


Fig. 18. Hierarchical graph (○ denotes (j-2)-order graph).

Fig. 18. Just as the rectangular graph might be thought of as a ring graph in which nodes are replaced by complete graphs, so can the hierarchical graph be thought of as a complete graph in which nodes are replaced by lower-order hierarchical graphs. A definition, by induction, of the j^{th}-order hierarchical graph is obtained as follows: Let a 0^{th}-order hierarchical graph be a single node. Then, a j^{th}-order hierarchical graph (j being any positive integer) is defined as a complete graph with N_j "supernodes," where a supernode is a $(j-1)^{th}$-order hierarchy. For $t = 1$, all external connections to a supernode are assumed to be made to just one of its (true) nodes. For $t \neq 1$, external connections must be distributed over all nodes of a supernode. This second case has not been investigated, and we shall consider only the case for $t = 1$. The hierarchical graph may be desirable in a situation in which communication is organized at a local level, regional level, national level, and so forth.

The values $N_1, N_2, \ldots, N_j$ for the j^{th}-order hierarchy have not been specified. Before doing so, let us give the formulas that apply to a j^{th}-order hierarchy.

$$n = N_1 N_2 \ldots N_j \tag{103a}$$

$$2d = N_1 + \frac{N_2}{N_1} + \frac{N_3}{N_1 N_2} + \ldots + \frac{N_j}{N_1 \ldots N_{j-1}} \tag{103b}$$

The derivation of Eq. 103a is obvious. Equation 103b is obtained as follows: First, assume that $N_1, N_2, \ldots, N_j$ are each sufficiently large that the density of links in the complete graphs can be taken as $2d = N$, rather than $2d = N-1$. Next, consider any node. This node has a density, N_1, of links that connect it only to the other nodes in the first-order supernode to which it belongs. Now, this first-order supernode is connected as part of a second-order supernode that has a density N_2. This density, N_2, must be distributed among the N_1 nodes in the first-order supernode; this gives an additional increment of density N_2/N_1. This process is continued and yields Eq. 103b.

Now, if $N_1 = N_2 = \ldots = N_j = N$, then the density would be

$$2d = N + 1 + \frac{1}{N} + \ldots + \frac{1}{N^{j-2}} = \frac{N - N^{-j+1}}{1 - 1/N} \sim N$$

Thus the density would vary as $2d = n^{1/j}$. Another possible choice of N_i's is the set that minimizes the density in a j^{th}-order hierarchy with n nodes. By use of induction, this set is obtained from the following equation:

$$N_i = \frac{1}{2}(N_{i-1})^2 \tag{104}$$

In this case,

$$d = \frac{1}{2^j - 1} n^{1/(2^j-1)}$$

and

$$d = N_1 (1-2^{-j})$$

Now, if S(n) is the probability of strong connectivity in a n-node complete graph, the probability of strong connectivity, S_j, in a j^{th}-order hierarchy is

$$S_j = [S(N_1)]^{n/N_1} [S(N_2)]^{n/(N_1N_2)} \dots [S(N_j)]^{n/(N_1\dots N_j)} \tag{105}$$

If the equal-N_i node distribution is used, and S(n) is obtained from Eq. 87a, S_j becomes

$$S_j \sim 1 - q^N \left[\frac{N^j - 1}{1 - 1/N}\right] \sim 1 - e^{-(2d \log 1/q - \log n)} \tag{106}$$

If the N_i's of Eq. 104 are used, S_j becomes (only the contribution of the first-order hierarchies is significant):

$$S_j \sim [S(N_1)]^{n/N_1} \sim 1 - e^{-[d\alpha(q,j) - \log n]}$$

where

$$\alpha(q, j) = \frac{\log 1/q}{1 - 2^{-j}}$$

Thus, for $t = 1$ and S the criterion of reliability, the j^{th}-order hierarchy with the equal-N_i node distribution is efficient, but the j^{th}-order hierarchy with the node distribution of Eq. 104 requires a density twice that of the necessary density. The second distribution has the advantage, however, that it requires a smaller value of j for a given density, and hence a smaller path length.

When W is the criterion of reliability, a lower bound on W_j of a j^{th}-order hierarchy is obtained by calculating the probability of a path between two nodes separated by the maximum possible distance. To connect these nodes, it is necessary to have a path between one of them and the node in its first-order supernode that has external connections, and then a path between this node and the node in the second-order supernode that has external connections, and so on up to the j^{th}-order supernode, and then back to the first-order supernode in which the other terminal node is situated. Therefore if W(n) is the probability of a path in an n-node complete graph,

$$W_j \geq W(N_1)\, W(N_2) \dots W(N_j) \dots W(N_1) \tag{107}$$

If we use the value of W(n) given by Eq. 87b, W_j, with $N_1 = N_2 = \dots = N_j$, becomes

$$W_j \sim 1 - (4j-2)\, q^{2d} \tag{108}$$

or with the node distribution of Eq. 104, it becomes

$$W_j \sim 1 - 2q^d \tag{109}$$

However, since $2d = n^{1/j}$ for the equal-N_i case,

$$j = \frac{\log n}{\log 2d} \tag{110}$$

and Eq. 108 becomes

$$W_j \sim 1 - 4 \frac{\log n}{\log 2d} q^{2d} \tag{111}$$

Thus, the j^{th}-order hierarchical graph is efficient in the equal-N_i case only when log log n is small compared with 2d log 1/q. Otherwise it is not efficient. The j^{th}-order hierarchical graph, with the node distribution of Eq. 104, requires a density that is twice the necessary density, independent of n, as is clear from Eq. 109.

We have now considered the efficiencies of several different base-graph configurations; the complete graph, the ring graph, the star graph, the rectangular graph, and the hierarchical graph. We have seen that some are efficient when either S or W is used as a criterion of reliability, and when node destruction is either allowed or not allowed. These configurations by no means exhaust all of the varieties that might be constructed and analyzed. However, the fact that in most of the cases studied, particularly for $t = 1$ and criterion S, the sufficient densities are within a factor of 2 of the necessary densities, makes one wonder whether or not this is true of graphs in general. This question can be answered by studying the average behavior of randomly chosen base graphs.

V. THE RANDOM-BASE PROBABILISTIC GRAPH

We shall now explore probabilistic graphs whose base graphs are chosen in a random fashion. This study is intended to provide information about the care with which a base graph must be chosen in order to provide an efficient probabilistic graph. Our approach is to specify a procedure for choosing an ensemble of base graphs, and then to determine the average values, $\overline{S}$ and $\overline{W}$, of the probabilities of strong and weak connectivity as a function of the link density d and number of nodes n. We shall find that, on the average, the base-graph ensembles that were studied are efficient only for small p.

The reader will find it convenient to turn to the Glossary for definitions of some of the symbols used in the following sections.

5.1 A RANDOM-BASE PROBABILISTIC GRAPH WITH A RANDOM NUMBER OF LINKS

Let us consider an ensemble of n-node base graphs chosen in such a way that the probability of a link being present between each of the $\binom{n}{2}$ different node pairs is β, the probability of no link being present is $1-\beta$, and all appearances are independent. No slings or parallel links are allowed. Thus, the number of links, L, present in a base graph is a random variable. The mean of this random variable is $\beta\binom{n}{2}$. If β is set equal to $n\,\overline{d}/\binom{n}{2}$, the mean number of links is n multiplied by $\overline{d}$, and the link density is $\overline{d}$.

The average reliability of the probabilistic graphs generated from such an ensemble of base graphs, if the links have a reliability p and the nodes a reliability t, is derived as follows. Consider the <u>average</u> probabilistic graph associated with the ensemble of base graphs. The reliability of a link in this average probabilistic graph is the product of the probability that the link appears in the base-graph ensemble multiplied by the probability that the link is reliable, once it appears. Each of the $\binom{n}{2}$ possible links has the same reliability. Thus, the average probabilistic graph is identical with a complete-base probabilistic graph with link reliability equal to $n\,\overline{d}\,p/\binom{n}{2}$, and node reliability equal to t. The values of S and W are obtained directly from Eqs. 90 and 91:

$$S \sim 1 - tn\left(1 - \frac{p2\overline{d}}{n-1}\right)^{n-1} \tag{112}$$

$$W \sim t^2\left[1-2\left(1 - \frac{p2\overline{d}}{n-1}\right)\left(1 - \frac{tp2\overline{d}}{n-1}\right)^{n-2}\right] \tag{113}$$

Equation 113 can be simplified because

$$\left[1-2\left(1 - \frac{p2\overline{d}}{n-1}\right)\left(1 - \frac{tp2\overline{d}}{n-1}\right)^{n-2}\right] \geqslant \left[1-2\left(1 - \frac{tp2\overline{d}}{n-1}\right)^{n-1}\right] \tag{114}$$

Furthermore, Eqs. 112 and 113 (with Eq. 114) can both be simplified by use of the inequality $(1-x) \leqslant e^{-x}$. Applying this inequality, we obtain

$$\bar{S} \sim 1 - tne^{-\bar{d}2p} = 1 - e^{-(\bar{d}2p - \log n - \log t)} \tag{115}$$

$$\bar{W} \sim t^2[1-2e^{-\bar{d}2pt}] \tag{116}$$

The results given by Eqs. 115 and 116 represent the average behavior of an ensemble of probabilistic graphs with randomly chosen base graphs. But these results leave much to be desired. First of all, we might be tempted to say that Eqs. 115 and 116 show, on the average, that the ensemble is efficient for small p, since $\log 1/q = p + \frac{p^2}{2} + \frac{p^3}{3} + \ldots$. Hence, the upper bound on S (Eq. 83) tends to Eq. 115 as p goes to zero. However, the derivation of the upper bound on S is valid only when d is a constant, and, in general, does not apply to the mean of d when d is a random variable over an ensemble of base graphs. Therefore we are not justified in saying that the ensemble is efficient, on the average, for small p. Another inadequacy of Eqs. 115 and 116 is that the derivation yields no appreciable insight into the properties of the random base-graph ensemble. For example, we obtain no information about necessary path lengths.

These complaints can be eliminated by a more detailed study of the average reliability of probabilistic graphs associated with a second base-graph ensemble.

5.2 A RANDOM-BASE PROBABILISTIC GRAPH WITH A FIXED NUMBER OF LINKS

a. Construction of Base-Graph Ensemble

We shall now consider a different base-graph ensemble than that studied in section 5.1. Let us start with a set of n nodes, labeled 1, 2, ..., n, and a set of L links, 1, 2, ..., L. A random-base graph is assembled from these n nodes and L links by the following procedure. A population of size n (corresponding to the n nodes) is sampled 2L times, with replacement after every second sample. The first two samples drawn, say n_1 and n_2, specify the location of link 1; that is, link 1 is placed between nodes n_1 and n_2. Likewise, the next two samples drawn specify the location of link 2, and so on, until finally the last two samples drawn specify the location of link L. Every base graph, therefore, has L links and n nodes. Note that parallel links can occur, but slings are excluded by the technique of replacing only after every second sample. The purpose of choosing the base graphs in this fashion is that if one terminal of a link is known, no information is available about the other terminal node except that it is distributed over the n - 1 remaining nodes in the graph with equal probability.

b. Analysis of Random-Base Probabilistic Graph

The term "member," as used in this section, refers to a linear graph in the probabilistic graph ensemble associated with one base graph in the base-graph ensemble. The measure (probability) associated with a member is the probability of obtaining the base graph multiplied by the probability of erasing that set of links and nodes which yields the member.

The random-base probabilistic graph will be analyzed in the following manner. Consider the calculation of $\overline{S}$. In every member of every probabilistic graph associated with the ensemble of base graphs, we designate one existing node as node I. If no nodes exist, we throw away that member of the ensemble. This elimination is conservative because we have assumed that $S = 1$ in the vacuous graph. This labeling of an existing node decreases the probability of the existence of each of the other nodes below t. Thus, the probability that node i exists, given that at least one node, not i, exists, is equal to

$$\frac{t(1-(1-t)^{n-1})}{1-(1-t)^n} < t \tag{117}$$

The numerator of the left-hand side of Eq. 117 is the joint probability that node i and at least one additional node exist. The denominator is the probability that at least one node exists. It is shown in Appendix II-A that the left-hand side of Eq. 117 is bounded below by $t(1-1/n)$.

We are now going to determine a lower bound on $\overline{S}$ by calculating the probability of those members for which node I is connected by paths of length, at most, $j + 1$ to every other existing node. We do this by "building out" from node I a structure of existing nodes and links in which node I is connected to at least 2k different nodes by paths of length 1, and is connected to at least $2k^2$ different nodes by paths of length 2, or less, and so on, until node I is connected to all existing nodes by paths of length $j + 1$, or less. Thus, every node is connected to every other node by a path in the structure of length that is less than or equal to $2j + 2$. We then calculate the probability of obtaining this structure, and use it as a lower bound to $\overline{S}$.

Let us calculate the probability, $P_1(2k)$, that at least 2k different existing nodes are attached to node I by an existing link. (Another way to view $P_1(2k)$, one which clarifies our use of the term "existing," is as the measure of those members for which at least k different nodes are each attached by a link to the node chosen as node I.) The calculation of $P_1(2k)$ proceeds as follows. The number of links attached to node I in the various base graphs is a random variable equal to the number of times node I was chosen during the 2L selections. The mean of this random variable is 2d. Let us define B(b, 2d) as the probability of obtaining a base graph in which the number of links attached to node I exceeds a fraction b, $0 < b < 1$, of the mean number, 2d. We shall postpone the calculation of B(b, 2d) and the choice of a value for b. Note, however, that B(b, 2d) approaches 1 as b approaches 0.

Let us now consider only those base graphs for which the number of links attached to node I exceeds 2bd. In the probabilistic graphs associated with these base graphs, some of the links may be destroyed. We define E(e, 2bd) as the probability of obtaining members for which a fraction e, $0 < e < p$, or more, of the links attached to node I exist. Discard those members for which this is not true.

The 2ebd, or more, links attached to node I have nodes other than node I at their far end (no slings are allowed), but some of these nodes may be duplicates. We define G(g, 2ebd) as the probability that a fraction g, $0 < g < 1$, or more, of the nodes are different. Throw away all of the members that do not have at least 2gebd different nodes attached to node I by existing links (before node destruction).

As we have indicated, node destruction must be taken into account. This is the final step in calculating the probability that at least 2k intact nodes are connected by intact paths, each of length 1, to node I. Define H(h, 2gebd) as the probability that a fraction h, $0 < h < t$, of the different nodes connected to node I exist. Discard all members for which this is not true.

Therefore, out of all of the members of probabilistic graphs associated with the ensemble of base graphs, we have kept only those members for which at least 2hgebd different nodes are linked directly to node I. For convenience, set k = hgebd. The total measure of these members provides a lower bound on $P_1(2k)$, since some of the discarded members may not meet the specified conditions but may still contribute to $P_1(2k)$. We have

$$P_1(2k) \geq B(b, 2d) \times E(e, 2bd) \times G(g, 2ebd) \times H(h, 2gebd) \tag{118}$$

where $0 < b < 1$, $0 < e < p$, $0 < g < 1$, and $0 < h < t$. Note that each of the probabilities are conditional on the truth of the event associated with the preceding probability. Let us call the reliable structure that we are building out from node I a "bush." The set of 2k reliable nodes plus node I are called "level 1" of the bush.

We shall now build out from level 1 of our bush in a manner similar to that by which we reached level 1, starting, of course, with k nodes rather than 1. Thus, $B_2(b, 4dk)$ is the probability (with link destruction neglected) that at least (2bd)(2k) links are attached to our level-1 nodes. A slight change is now made. The 4dbk links may not all be different, because it is possible for a link to connect two of the level-1 nodes, and hence be counted twice. (This could not happen when we were building level 1 because of the exclusion of slings. In that case, all 2bd links provided distinct choices of the n-1 remaining nodes.) However, we are extremely conservative if we assume that only one half of the 4bdk links are different. Now we define $E_2(e, 2bdk)$ as the probability that at least a fraction e of the 2bdk links exists. Similarly, $G_2(g, 2ebdk)$ is the probability that at least 2gebdk nodes attached to the existing links are different, and $H_2(h, 2gebdk)$ is the probability that at least a fraction h of these nodes exists. The set of $2k^2$ nodes plus the nodes of level 1 is called "level 2" of the bush.

We can now proceed to level 3, and, indeed, to level j, of our bush in a manner identical with that by which we reached level 2. At level j, we have at least $2k^j$ different, existing nodes connected by existing paths of length, at most, j to the original node. We also have a lower bound on the probability of reaching level j. The level, j, at which we stop this bush-building procedure is governed by the behavior of the probability $G_j(g, 2ebdk^j)$. It becomes less and less probable that a fixed fraction, g, of the nodes chosen will be different as the number of choices approaches the order of the number of nodes in the graph. The explicit choice of j will be discussed later, but we stop while G_j is still close to 1.

At the j^{th} level, we have k^j nodes connected together, but some of the existing nodes in the graphs may still not be included. Therefore, one more step is necessary. The probability that $2ebdk^j$ or more different links exist and are connected to level-j nodes is similar to before. Now, however, we do not calculate the probability that a fraction g of the chosen nodes are different. Instead, we calculate the probability that the chosen nodes include all of the existing nodes in the graph. Call this probability M(z), where we have set $z = 2ebdk^j$. Thus, the probability of obtaining a j^{th}-level bush times M(z) gives a lower bound on $\overline{S}$. This lower bound includes only those cases for which two nodes are connected by intact paths of length less than 2j + 2.

c. Bounding of Probabilities Used in Analysis

We must now calculate bounds on the various probabilities used in the preceding analysis. We first introduce a useful bound on the tails of the binomial distribution. The exponent in this bound has been shown to be asymptotically correct by Shannon (26). Let N Bernoulli trials be made with probability of success P. Then, the probability that the number s, of successes exceeds a fraction α, $0 < \alpha < P$, of the total number of trials is:

$$\left.\begin{aligned} &\Pr\{s \geq \alpha N\} \geq 1 - e^{-NR(\alpha, P)} \\ &R(\alpha, P) = T_P(\alpha) - H(\alpha) \\ &T_P(\alpha) = -\alpha \log P - (1-\alpha) \log (1-P) \\ &H(\alpha) = -\alpha \log \alpha - (1-\alpha) \log (1-\alpha) \end{aligned}\right\} \tag{119}$$

$H(\alpha)$, a function of one variable, should not be confused with H(h, x), a function of two variables. The two terms in the exponent $R(\alpha, P)$, $T_P(\alpha)$, and $H(\alpha)$, are plotted as functions of α in Fig. 19. It can be seen that $R(\alpha, P)$ is positive, and decreases from $\log (1/1-P)$ to 0 as α goes from 0 to P.

We shall now use Eq. 119 to calculate B(b, 2d), the probability that node I is selected at least 2bd times. We find that the probability that node I is selected as one terminal of a link during two drawings, without replacement, from a population of size n is equal to the probability that it is selected on the first drawing, 1/n, added to the probability, $(1-1/n) \times (1/n-1)$, that it is not selected on the first drawing but it is selected on the second. The probability of node I being a terminal is, therefore,

$$\frac{1}{n} + \left(1-\frac{1}{n}\right) \times \frac{1}{n-1} = \frac{2}{n} \tag{120}$$

Since there are L drawings for links, the mean number of times node I is drawn is $2L/n = 2d$, which agrees with our previous statement. Now, through the use of Eq. 119 with $N = L$, $P = 2/n$, and $\alpha = 2bd/L = 2b/n$ ($< 2/n$) (see Appendix II-B), we obtain

$$B(b, 2d) \geq 1 - e^{-d(1-b)^2/2} \tag{121}$$

Next, let us calculate $B_i(b, 4dk^i)$, the probability of obtaining at least $4bdk^i$ links attached to the nodes of level i, having already built the bush out to level i. First, note that the number of links attached to a set of nodes is independent of the reliability of

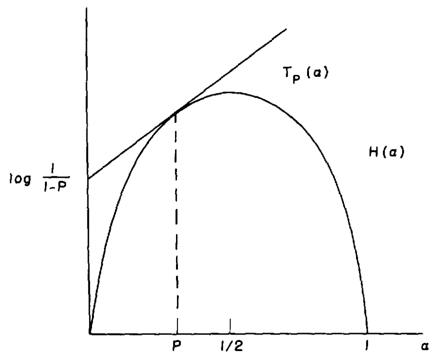


Fig. 19. Behavior of the exponent in binomial distribution bound.

any previously selected set of links and nodes. Second, remember that the level-i nodes include all previously selected nodes, including node I. Thus, it is conservative to "forget" the information that previous levels have achieved their quotas of attached links, and simply calculate the probability that $4bdk^i$ links be attached to $2k^i$ nodes. Note that this "forgetting" would not be conservative if the previous nodes were not included with the i^{th}-level nodes, because the fixed number, L, of links in the graph requires that if some nodes have a larger set of attached links, other nodes must have a smaller set. For example, suppose we know that the original node has at least 2bd attached links. Then, if the original nodes were not included with the level-1 nodes, the probability that the level 1 nodes have 4bdk attached links would be slightly decreased.

The probability $B_i(b, 4dk^i)$ can now be obtained from Eq. 121, since it is only necessary to substitute $4k^i/n$ for $2/n$ as the probability, and $4bdk^i$ for 2bd as the fraction αL. Both changes are effected by substituting $2dk^i$ for d. We have

$$B_i(b, 4dk^i) \geq 1 - e^{-dk^i(1-b)^2} \tag{122}$$

It is shown in Appendix II-C that a very conservative bound on the product of the B's occurring at all levels of the bush is

$$B(b, 2d) \times B_1(b, 4dk) \times \ldots \times B_j(b, 4dk^j) \geq 1 - 2\, e^{-d(1-b)^2/2} \tag{123}$$

Let us proceed to calculate $E_i(e, 2bdk^i)$ and $H_i(h, 2egbdk^i)$, the probabilities of obtaining existing links and nodes, respectively, at the i^{th} level. It is conservative to "forget" that links and nodes have existed at previous levels because this knowledge can only increase the probability that the i^{th}-level links and nodes also exist. Equation 119 can be applied directly, and it yields

$$E_i(e, 2bdk^i) \geq 1 - e^{-2bdk^i R(e,p)}, \qquad 0 < e < p \tag{124a}$$

$$H_i(h, 2gebdk^i) \geq 1 - e^{-2gebdk^i R(h,t)}, \qquad 0 < h < t \tag{124b}$$

Conservative bounds on the product of E_i's and H_i's for $i = 0, 1, \ldots, j$ are similar to the bound in Eq. 123. Hence

$$E(e, 2bd) \times E_1(e, 2bdk) \times \ldots \times E_j(e, 2bdk^j) \geq 1 - 2\, e^{-2bd\, R(e,p)} \tag{125a}$$

$$H(h, 2gebd) \times H_1(h, 2gebdk) \times \ldots \times H_{j-1}(h, 2gebdk^{j-1}) \geq 1 - 2\, e^{-2gebd\, R(h,t)} \tag{125b}$$

In Eq. 125b, H_j is not included because the probability that the j^{th} level connect to all of the existing nodes is covered by M(z). Notice that the bounds in Eqs. 123, 125a, and 125b are all independent of j.

We must now consider $G_i(g, 2ebdk^i)$, the probability that a fraction, g, of the nodes selected at the i^{th} level will be different. Again, we claim that it is conservative to forget that previous sets of nodes have been different, and calculate G_i by carrying out $2ebdk^i$ independent Bernoulli trials. The justification is a little more difficult here than in the previous cases. Recall that we are considering those members that have at least $2ebdk^i$ existing links attached to the i^{th}-level nodes. Some of these links are links that connect the i^{th}-level nodes back to node I, and hence we know that they must be present. We can consider that each one of these special links is connected to a different node, these nodes being just the nodes in level i, which, by convention, we include in level i + 1. The remaining links are attached to nodes chosen at random, and must produce the additional different nodes that are necessary to give us $2gebdk^i$ different nodes in level i + 1, before node destruction. If we consider all $2ebdk^i$ nodes to be chosen at random, we get a conservative number of different nodes, since we have then ignored the subset of nodes that we know are different.

Thus we can calculate $G_i(g, 2ebdk^i)$. For convenience, let $y = 2ebdk^i$. Make y drawings with replacement from a population of size n. (We are conservative in not using replacement after every other trial, since replacement after every trial increases the probability of repeated nodes.) Define $G_i(g, y)$ as the probability that a fraction g, or more, of the y drawings are different. The calculation of $G_i(g, y)$ is one variation of the classical occupancy problem. Let x be the number of nodes that are not chosen during the y drawings (n - x is the number of different nodes chosen), and let P(s;y, n) be the probability that x = s. Then

$$G(g, y) = \sum_{s=0}^{n-gy} P(s;y,n) = 1 - \sum_{s=n-gy}^{n} P(s;y,n) \tag{126}$$

The tail of the distribution, P(s;y, n), as required by Eq. 126, is obtained in an interesting manner. The k^{th} binomial moment of a discrete distribution $\{p_i\}$ is defined as

$$B_k = \sum_{i=k}^{\infty} \binom{i}{k} p_i \tag{127}$$

Since $\binom{i}{k}$ is a monotonic increasing function of i, the right-hand side of Eq. 127 is bounded below by setting $\binom{i}{k} = \binom{k}{k} = 1$ for all i. Thus

$$B_k \geq \sum_{i=k}^{\infty} p_i \tag{128}$$

Let $B_k(y, n)$ be the k^{th} binomial moment of the P(s;y, n) distribution. Then, from Eqs. 126 and 128,

$$G(g, y) \geq 1 - B_{n-gy} \tag{129}$$

From a problem of Riordan (24), we have

$$B_k(y, n) = \binom{n}{k} \left[1-\frac{k}{n}\right]^y \tag{130}$$

Substituting Eq. 130 in Eq. 129 and simplifying (see Appendix II-D), we obtain the desired bound on G(g, y):

$$G(g, y) \geq 1 - e^{-y(1-2g)\log(n/gy)}, \quad \text{if } gy \leq \frac{n}{2} \tag{131}$$

Now, if j is such that $2gebdk^{j-1} \leq n/2$, then Eq. 131 can be employed to obtain the probability of choosing different nodes at all levels of the bush. Thus, conservatively, we have

$$G(g, 2ebd) \times G_1(g, 2ebdk) \times \dots \times G_{j-1}(g, 2ebdk^{j-1})$$
$$\geq 1 - 2\, e^{-2ebd(1-2g)\log(n/2gebd)} \tag{132}$$

where $2gebdk^{j-1} \leq n/2$. However, if j is chosen so that

$$2k^{j-1} < hn/2 \tag{133}$$

but

$$2k^j = h(2gebdk^{j-1}) \geq h\ n/2 \tag{134}$$

then Eq. 132 is valid only when Eq. 134 is an equality. When Eq. 134 is an inequality, the right-hand side of Eq. 132 can be used as a lower bound on obtaining at least hn/2 (rather than $2k^j$) different nodes (before node destruction) in the j^{th} level. Simply reduce the fraction of different nodes in the G_{j-1} term to $g'(g'<g)$, where $2g'ebdk^{j-1} = hn/2$. This reduction is conservative, so the right-hand side of Eq. 132 is unchanged.

It is now only necessary to calculate the value of M(z), the probability of choosing all reliable nodes in z choices with replacement from the population of size n. To be on the safe side, we shall require that all n nodes be chosen. The probability that node 1 is selected at least once in z trials is

$$1 - \left(1-\frac{1}{n}\right)^z \tag{135}$$

Then the probability that node 2 is selected, given that node 1 is selected, is less than the probability that node 2 is selected without this condition. However, let us ignore the condition even though it is not conservative. Then, Eq. 135 holds for each of the n nodes, and

$$M(z) \leq \left[1 - \left(1-\frac{1}{n}\right)^z\right]^n \tag{136}$$

Note that the right-hand side of Eq. 136 is bounded below by

$$\left[1 - \left(1-\frac{1}{n}\right)^z\right]^n \geq [1-e^{-z/n}]^n \tag{137}$$

However, Feller (3) claims that the right-hand side of Eq. 137 represents the asymptotic behavior of M(z).

$$M(z) \sim [1-e^{-z/n}]^n \geq 1 - ne^{-z/n} \tag{138}$$

Here, M(z) is the last probability needed to complete our analysis of the random probabilistic graph when $\overline{S}$ is the criterion. It would have been desirable to require that only the reliable nodes in any base graph be connected to level j. But the conditional probability distribution of the number of nodes is not known.

It is now possible to write the bound on $\overline{S}$ explicitly. Let us determine the value of z to be used in Eq. 138. The value of j at which we stop building our bush is specified by Eqs. 133 and 134. Hence, there are at least hn/2 nodes in the j^{th} level, so that the number of existing, attached links, z, is given by

$$z = ebdhn/2 \tag{139}$$

From Eqs. 123, 125a, 125b, 132, 138, and 139, and with the use of the inequality $(1-x_1)(1-x_2)\ldots(1-x_n) \geq 1-(x_1+x_2+\ldots+x_n)$ when all the x_i are positive (which is proved by induction on n), we find that the bound on $\bar{S}$ is

$$\bar{S} \sim 1 - \Big(2e^{-d(1-b)^2/2} + 2e^{-2bd\,R(e,p)} + 2e^{-2ebd(1-2g)\log(n/2gebd)} + 2e^{-2gebd\,R(h,t)} + ne^{-ebhd/2} \Big) \tag{140}$$

where $0 < b < 1$, $0 < e < p$, $0 < g < 1/2$, and $0 < h < t$.

The parameters b, e, g, and h in Eq. 140 might well be chosen close to their upper bounds so that each of the terms in the sum on the left-hand side would be of equal magnitude. However, for any fixed values of those parameters, there is a value of n above which the last term dominates. If δ is a number just greater than zero, the bound then assumes the form

$$\bar{S} \sim 1 - e^{-[(pt+\delta)d/2 - \log n]} \tag{141}$$

Thus, for p small (see section 5.1) and with $t = 1$, the random-base probabilistic graph requires, on the average, a density four times greater than the necessary density. For $t < 1$, even greater densities are required. For p that is not small, the random-base probabilistic graph is inefficient (or this bound is inefficient).

Moreover, we must remember that we have obtained an additional condition on this bound. No paths of length greater than $2j + 2$ are used to connect nodes. From Eq. 133,

$$j - 1 < \frac{\log hn/4}{\log k} \tag{142}$$

Since $k = hgebd$, the path length is bounded by

$$\text{Path length} < 4 + 2\,\frac{\log hn/4}{\log hgebd} \tag{143}$$

Therefore the sufficient path length varies directly as the logarithm of the number of nodes, and inversely as the logarithm of link density. Notice that if path length is bounded, Eq. 143 imposes a much more severe restriction on density than does reliability, because the density is forced to grow directly with the number of nodes.

The analysis of the random-base probabilistic graph is completed by bounding the average value of the probability of weak connectivity, $\overline{W}$. This can be done rather simply by following our previous methods. Then, by restricting our random ensemble of base graphs, a much tighter bound will be obtained.

Let us consider the simple case. Since the average behavior of the graph is insensitive to node names, $\overline{W} = \overline{P}_{12}$. Therefore, consider the average probability of a path between the two nodes 1 and 2, which exist with probability t^2. We build out a bush from one of the terminal nodes, say node 1, just as we built out a bush from node 1 in the case of $\bar{S}$. Again, we stop at level j given by Eqs. 133 and 134. This time, however, we require only that node 2 be included in the nodes attached to the j^{th}-level nodes. Call this probability Q(z). Q(z) is the probability that node 2 is chosen on one of the z selections from a population of size n, with replacement. Clearly,

$$Q(z) = 1 - \left(1-\frac{1}{n}\right)^z \geq 1 - e^{-z/n} \tag{144}$$

and hence, by using the z of Eq. 139, and substituting Q(z) for M(z) in Eq. 140, we obtain

$$\overline{W} \geq t^2\Big[1 - \Big(2e^{-d(1-b)^2/2} + 2e^{-2bd\,R(e,p)} + 2e^{-2ebd(1-2g)\log(n/2gebd)} + 2e^{-2gebd\,R(h,t)} + e^{-behd/2}\Big)\Big] \tag{145}$$

where $0 < b < 1$, $0 < e < p$, $0 < g < 1/2$, and $0 < h < t$. Notice that Eq. 145 is independent of n. The first and last terms in the parentheses are the restrictive terms. Hence, if δ is, again, a number slightly greater than 0, then

$$\overline{W} \geq t^2\Big[1 - \Big(2e^{-d(1-b)^2/2} + 2e^{-(pt-\delta)d/2}\Big)\Big] \tag{146}$$

Unless pt is small, the first term dominates. Again, for $t = 1$ and p very small, the average behavior is efficient.

A tighter bound on $\overline{W}$ than Eq. 146 can be obtained by slightly modifying the random base-graph ensemble. Since the term that dominates Eq. 146, unless pt is small, is the probability of obtaining at least 2bd links attached to node I, in order to obtain a larger bound, it is necessary to purge the ensemble of all base graphs for which a node exists with less than 2bd attached links. It should be mentioned that although many such base graphs exist, their measure is small. However, other selection procedures can be thought of for forming n-node, L-link base-graph ensembles with at least 2bd links attached to every node, and with the property that information about one end of a link gives no information about the other end.

Let us suppose that we have such a purged ensemble, and calculate a lower bound on $\overline{W} = \overline{P}_{12}$. Rather than building out a bush from node 1 alone, we shall build out two bushes, one from node 1 and one from node 2, and then calculate the probability that the two bushes have an existing node in common. Let us consider the building of the bushes. The probabilities, $B_i(b, 4dk^i)$, need no longer be taken into account, because we are assured of obtaining the $4bdk^i$ links attached to the $2k^i$ i^{th}-level nodes by our method of purging the base-graph ensemble. Furthermore, the link and node reliability probabilities, $E_i(e, 2bdk^i)$ and $H_i(h, 2gebdk^i)$, are unaffected. It is only necessary to take a second look at $G_i(g, 2ebdk^i)$, the probability of obtaining a fraction, g, of different nodes. We claim that it is conservative to forget that the ensemble has been purged and use the previously calculated values of G_i. The proof of this claim rests on the supposition that nodes with a small number of attached links are unlikely to be chosen, and hence are unlikely to contribute to the number of different nodes. Eliminating these infrequently chosen nodes increases the probability of choosing a set of different nodes.

It is now necessary to obtain T(z), the probability that if two sets of z choices are made independently from a population of size n, with replacement, then they have an existing node in common. First, let us calculate a lower bound on the number of existing nodes. As before, the probability that the number of existing nodes is hn or greater is simply H(h, n), where $0 < h < t$. This probability can be absorbed in the product of Hi's in Eq. 125b.

Second, let us calculate T(z). Enumerate the existing nodes 1, 2, ..., hn. The probability that node 1 is chosen at least once in both selections is, from Eq. 135,

$$\left[1 - \left(1-\frac{1}{n}\right)^z\right]^2 \tag{147}$$

The probability that node 1 is not chosen in each selection is one minus expression 147. Now, the probability that node 2 is not chosen at least once in each selection, given that node 1 was not chosen at least once, is smaller than the unconditional probability that node 2 is not chosen, because if one node is not chosen, the probability of other nodes being chosen is increased. Consequently, ignoring the conditions on the selection of nodes 2, 3, ..., hn gives a conservative estimate of T(z).

$$T(z) \geq 1 - \left\{1 - \left[1 - \left(1-\frac{1}{n}\right)^z\right]^2\right\}^{hn} \tag{148}$$

Simplifying Eq. 148 (see details in Appendix II-E), we obtain for T(z):

$$T(z) \geq 1 - e^{-hn(z/n)^2/8}, \qquad z \leq \frac{3}{4}n \tag{149a}$$

$$T(z) \geq 1 - e^{-hn[z/n-0.7]}, \qquad z > \frac{3}{4}n \tag{149b}$$

From Eq. 139, $z/n = ebhd/2 > 3/4$, and so Eq. 149b applies. Hence, if we use Eqs. 145 and 149b, and recall that two bushes must be built, the bound on $\overline{W}$ becomes

$$\overline{W} \geq t^2\Big[1-\Big(4e^{-2bd\,R(e,p)}+4e^{-2ebd(1-2g)\log(n/2gebd)} +4e^{-2gebd\,R(h,t)}+e^{-hn[ebnd/2-0.7]}\Big)\Big] \tag{150}$$

The controlling terms in Eq. 150 are the link- and node-reliability terms. Thus, if δ is any (small) positive number, e and h can be chosen so that

$$R(e,p) > \log(1/1-p) - \delta \tag{151a}$$

$$R(h,t) > \log(1/1-t) - \delta \tag{151b}$$

Then, if we choose d sufficiently large that, for example, $k = hgebd > 5$, and choose $b = 1-\delta$, the bound on $\overline{W}$ becomes

$$\overline{W} \geq t^2[1-8e^{-2d[\log(1/1-p)-2\delta]}-4e^{-2ged[\log(1/1-t)-2\delta]}] \tag{152}$$

where g is chosen sufficiently small to cause the third term on the right-hand side of Eq. 150 to be approximately equal to the second term. We see from Eq. 152, that the purged ensemble is efficient on the average for all values of p when t is sufficiently close to 1, and hence that the second term in Eq. 152 dominates the third.

To summarize, for t close to 1, the purged ensemble can be efficient, on the average, when W is the criterion of reliability. In particular, the purged ensemble is more efficient than the unpurged ensemble. A further discussion of results from the random-base probabilistic graphs will be found in Section VI.

VI. CONCLUSION

6.1 DISCUSSION OF RESULTS AND SUGGESTIONS FOR FURTHER RESEARCH

Let us consider the results for small graphs presented in Section III. Section 3.1 dealt with the procedure for exact analysis, and it was pointed out that in all but very small or very symmetric graphs (such as the complete graph) exact analysis is a long task involving much "bookkeeping." It seems fair to say that the exact analysis of other graphs, except for simple configurations like the ring graph and the star graph, is not an interesting direction for research. Systematic procedures for determining all of the trees, or paths, or cutsets of a graph would be useful in applying the inclusion-exclusion technique.

A more interesting aspect of small-graph study is synthesis. The theorem proved in section 3.2 is an indication of results that might be obtained. The basic problem, here, is to determine that configuration of n-nodes and L-links which has the highest value of S or W when it is used as a base graph. Node destruction may or may not be allowed. It is not clear whether or not there is a uniformly best solution, or whether or not the optimum configuration will be a function of p (and t).

An insight into the synthesis problem might prove to be of great value in deriving tight bounds on S and W for small graphs. For example, $P_2(n, L)$, the upper bound on S derived in section 3.3, can be realized by a graph only if every set of $n - 1$ links is a tree. This is not possible if L is greater than n. If additional constraints on realizable graphs were known, this bound might be improved.

The results from large graphs, presented in Sections IV and V, provide a great deal of information about both necessary and sufficient link densities. An immediate application is found in the design of communication networks that are intended to handle messages of low information content, but of high priority, such as an order to fire a pre-aimed missile. The necessary and sufficient channel redundancy can be calculated as soon as message-center and channel reliabilities are known.

In particular applications, criteria of reliability other than S or W might be desired. One possibility might be the probability that a majority of existing nodes, rather than all of the nodes, are strongly connected. This criterion would be of value, for example, for communicating with individual soldiers in the field. Such changes can be accomplished by straightforward modification of the bounds used in Sections IV and V.

More research is still necessary, however, to bring closer together the necessary and sufficient densities, particularly when t is not 1. Future research might involve the study of additional configurations, or the improvement of upper and lower bounds. Furthermore, research on the interplay between reliability, path length, and link density, although it has been illuminated by the study of random-base probabilistic graphs in Section V, needs to be pursued further. For example, can an efficient graph be obtained when path length is bounded and node destruction allowed? If not, what is the minimum growth of path length with number of nodes in an efficient graph?

The study of random-base probabilistic graphs has explained one phenomenon, the independence of the necessary link density from the number of nodes in the graph, when $\overline{W}$ is the criterion. Therefore it is only necessary to choose the link density large enough so that the terminal nodes are connected to level-1 nodes with high probability. From then on, bushes can be built into the graph from the two sets of level-1 nodes until between them they contain a large number of the nodes of the graph, and thus have a high probability of overlapping. If the number of nodes is increased, it is only necessary to build the bushes to higher levels, and this can be done with probability close to 1. Clearly, the number of nodes affects only the path length, and not the necessary density.

Furthermore, the study of random-base probabilistic graphs has raised other questions. Does a random ensemble behave efficiently, on the average, only for small p, when S is the criterion? Is it possible to purge the ensembles, as was done in Section V for the bounding of $\overline{W}$, to obtain good efficiency for all p? These questions, which are of great interest, all apply to extensions of existing results.

New directions may also be investigated. Finding the capacity of a graph when the capacity of the links is known, is an interesting, but difficult, problem. So is the matter of necessary and sufficient link densities when the graph is harassed by an intelligent, rather than a random, opponent. Special cases of nonindependent link failure should be investigated. An important example of this is the unavailability of channels that are already being used to transmit messages, when these messages are being relayed through several switching centers (path length greater than 1).

6.2 OTHER APPLICATIONS OF OUR MODEL

In this research we have been studying an abstract model of a communication network. But the results are equally valid when they are applied to any situation for which the model is an acceptable idealization. In particular, the results apply to any situation in which the existence of a path between sets of objects is the desired event, and both the paths and objects satisfy our assumption of independent failures with known probabilities.

One such application, which has little relation to communication networks, is the study of "rumor networks" (suggested by Professor I. D. Pool, of the Department of Economics and Social Science, M. I. T.). In this "network," members of a human population are the nodes, and acquaintances between pairs of members form the links. There is a probability, p, that if one member knows of a rumor, he will repeat it to an acquaintance. One wishes to find the probability that all members of a population will hear a rumor. If we assume that the number and identity of each person's acquaintances are random, then the results of section 5.1 can be applied.

It is to be expected, however, that other applications will require different emphases, hence different constraints to be placed upon probabilistic graphs. The methods developed in this report may then form a basis for the new inquiry.

APPENDIX I

TABULATION OF UPPER BOUNDS ON S

The entries in Table 2 are upper bounds on the probability of strong connectivity in an n-node L-link probabilistic graph, with node reliability p = 0.1, 0.3, 0.5, 0.7, or 0.9, and link reliability t = 1. The principal entries are $P_2(n, L)$ as defined by Eq. 46 or 50. The numbers are taken from Pearson's (19) tables (see Eq. 56).

Some values of the upper bound given by Eq. 57 have also been calculated; these values are entered in parentheses in Table 2.

For comparative purposes, values of S(n), the exact probability of strong connectivity in an n-node graph, have been taken from Gilbert's table (7); these are marked by asterisks in Table 2.

Table 2. Upper bounds on S.

n	L	p = 0.1	p = 0.3	p = 0.5	p = 0.7	p = 0.9
2	1	.1	.3	.5	.7	.9
2	1	.1*	.3*	.5*	.7*	.9*
3	2	.01	.09	.25	.49	.81
3	3	.028	.216	.5	.784	.972
3	3	.028*	.216*	.5*	.784*	.972*
4	3	.001 (>1)	.027	.125 (.646)	.343	.729 (.892)
4	4	.0037 (>1)	.0837	.3125 (.596)	.6517	.9477 (.963)
4	5	.0086 (>1)	.1631	.5000 (.627)	.8369	.9914 (.988)
4	6	.0159 (>1)	.2557	.6563 (.688)	.9295	.9987 (.9962)
4	6	.0129*	.2187*	.5938*	.8925*	.9958*
4	7	.0257 (>1)	.3529	.7734 (.751)	.9712	.9998 (.9987)
5	4	.0001 (>1)	.0081	.0625 (.789)	.2401	.6561 (.891)
5	6	.0013 (>1)	.0705	.3438 (.572)	.7443	.9842 (.981)
5	8	.0050 (>1)	.1941	.6367 (.662)	.9420	.9996 (.997)
5	10	.0128 (>1)	.3504	.8281 (.766)	.9894	1.0000 (1.000)
5	10	.0081*	.2563*	.7109*	.9575*	.9995*
6	5	.00001	.0024	.0313	.1681 (.4358)	.5905
6	7	.00018	.0288	.2266	.6471 (.7354)	.9743
6	9	.00089	.0988	.5	.9012 (.8608)	.9991
6	11	.0028	.2103	.7256	.9784 (.9329)	1.000
6	13	.0065	.3457	.8666	.9959 (.9689)	1.000
6	15	.0127	.4846	.9408	.9993 (.9857)	1.000
6	15	.0062*	.3169*	.8157*	.9850*	.9999*

APPENDIX II

INEQUALITIES USED IN SECTION V

A. We want to show that

$$\frac{t(1-(1-t)^{n-1})}{1-(1-t)^{n}} \geqslant t\left(1-\frac{1}{n}\right) \qquad \text{(II-1)}$$

Since

$$t\,\frac{1-(1-t)^{n-1}}{1-(1-t)^{n}} = \frac{t}{1+\dfrac{(1-t)^{n-1}-(1-t)^{n}}{1-(1-t)^{n-1}}} = \frac{t}{1+\dfrac{t}{(1-t)^{-n+1}-1}} \qquad \text{(II-2)}$$

and

$$(1-t)^{-n+1}-1 = t(n-1) + \text{positive terms}$$

therefore

$$\frac{t}{1+\dfrac{t}{(1-t)^{n-1}-1}} \geqslant \frac{t}{1+\dfrac{1}{n-1}} = t\left(1-\frac{1}{n}\right) \qquad \text{(II-3)}$$

B. We want to show that

$$1 - e^{-LR(2b/n,\,2/n)} \geqslant 1 - e^{-d(1-b)^{2}/2} \qquad \text{(II-4)}$$

From Eq. 119, we have

$$R\left(\frac{2b}{n},\frac{2}{n}\right) = \frac{2b}{n}\log\frac{n}{2} + \left(1-\frac{2b}{n}\right)\log\left(\frac{n}{n-2}\right) - \frac{2b}{n}\log\frac{n}{2b} - \left(1-\frac{2b}{n}\right)\log\frac{n}{n-2b} \qquad \text{(II-5)}$$

After simplification, Eq. II-5 becomes

$$R(2b/n,\,2/n) = \left(1-\frac{2b}{n}\right)\log\frac{n-2b}{n-2} + \frac{2b}{n}\log b \qquad \text{(II-6)}$$

Let $a = 1 - b$, and expand the log terms to obtain

$$LR(2b/n,\,2/n) = \left(1-\frac{2b}{n}\right)\left(\frac{2a}{n-2} - \frac{1}{2}\left(\frac{2a}{n-2}\right)^{2} + \dots\right) - \frac{2b}{n}\left(a + \frac{a^{2}}{2} + \dots\right) \qquad \text{(II-7)}$$

Now, if the first series in Eq. II-7 is cut off at a negative term, it makes the right-hand side too small. The second series is bounded as follows:

$$\left(a + \frac{a^2}{2} + \frac{a^3}{3} + \ldots\right) < a + \frac{1}{2}(a^2+a^3+\ldots) = a + \frac{a^2}{2b}$$

Thus, Eq. II-7 is bounded by

$$R(2b/n, 2/n) \geq \left(1 - \frac{2b}{n}\right)\left(\frac{2a}{n-2} - \frac{1}{2}\left(\frac{2a}{n-2}\right)^2\right) - \frac{2ba}{n} - \frac{a^2}{n}$$

By simple algebraic manipulations, we obtain

$$\left(1-\frac{2b}{n}\right)\frac{2a}{n-2} - \frac{2ba}{n} = \frac{2a^2}{n-2} \qquad \text{(II-8)}$$

Furthermore,

$$\frac{2a^2}{n-2} - \left(1-\frac{2b}{n}\right)\frac{1}{2}\left(\frac{2a}{n-2}\right)^2 - \frac{a^2}{n} \geq \frac{a^2}{n} - \frac{2a^2}{(n-2)^2} \qquad \text{(II-9)}$$

For $n \geq 8$, the terms on the right-hand side of Eq. II-9 are greater than

$$\frac{a^2}{n} - \frac{2a^2}{(n-2)^2} \geq \frac{a^2}{n} - \frac{2a^2}{4n} = \frac{a^2}{2n}$$

Whence

$$LR(2b/n, 2/n) \geq L\frac{a^2}{2n} = d(1-b)^2/2 \qquad \text{Q. E. D.} \qquad \text{(II-10)}$$

C. We want to show that

$$(1-e^{-a})(1-e^{-ak}) \ldots \left(1-e^{-ak^j}\right) \geq 1 - 2e^{-a} \qquad \text{(II-11)}$$

If $x_i \geq 0$, then $(1-x_1)(1-x_2) \ldots (1-x_n) \geq 1 - (x_1+x_2\ldots+x_n)$, and if $k \geq 2$, then $k^i \geq ik$. Using these inequalities, we have

$$(1-e^{-a}) \ldots \left(1-e^{-ak^j}\right) \geq 1 - (e^{-a}+(e^{-a})^2+\ldots(e^{-a})^j)$$

Summing the geometric series, we obtain

$$(1-e^{-a}) \ldots \left(1-e^{-ak^j}\right) \geq 1 - \frac{e^{-a}}{1-e^{-a}} \qquad \text{(II-12)}$$

It is clear that if $e^{-a} \leq 1/2$, Eq. II-11 is true. If e^{-a} is $> 1/2$, the right-hand side of Eq. II-11 is negative, and hence the bound is worthless.

D. We want to show that if $gy \leq n/2$, then

$$\binom{n}{n-gy}\left(1-\frac{n-gy}{n}\right)^y \leq e^{-y(1-2g)\log(n/gy)} \qquad \text{(II-13)}$$

From Riordan (25), or from Shannon (26), we have

$$\binom{n}{r} \leq e^{-n\left(\frac{r}{n}\log\frac{r}{n}+\frac{n-r}{n}\log\frac{n-r}{n}\right)} \tag{II-14}$$

Consequently, if we set $a = gy/n$, Eq. II-13 is bounded by

$$\binom{n}{n-gy}\left(\frac{gy}{n}\right)^y \leq e^{-n(a\log a+(1-a)\log(1-a)-(y/n)\log a)} \tag{II-15}$$

However, (a) log (a) < (1-a) log (1-a), for $0 < a < 1/2$. This can be proved by showing that the derivative of $f(a) = (1-a)\log(1-a) - a\log a$ has only two roots that are symmetric about $a = 1/2$. By calculating f(a) for some a, we find that $f(a) > 0$, for $0 < a < 1/2$. By using this inequality in Eq. II-15, we obtain

$$\binom{n}{n-gy}\left(\frac{gy}{n}\right)^y \leq e^{-n(y/n-2a)\log 1/a} \qquad \text{Q. E. D.} \tag{II-16}$$

E. We want to show that if $T(z) = 1 - \left\{1 - \left[1 - \left(1-\frac{1}{n}\right)^z\right]^2\right\}^{hn}$, then

$$T(z) \geq 1 - e^{-hn(z/n)^2/4} \qquad z \leq \frac{3}{4}n \tag{II-17a}$$

$$T(z) \geq 1 - e^{-hn(z/n-0.7)} \qquad z > \frac{3}{4}n \tag{II-17b}$$

In either case,

$$T(z) \geq 1 - \{1-[1-e^{-z/n}]^2\}^{hn} = 1 - e^{-hz}(2-e^{-hz})^{hn} \tag{II-18}$$

By taking the term in parentheses to be 2, we obtain

$$T(z) \geq 1 - e^{-hn(z/n-\log 2)} \tag{II-19}$$

from which Eq. II-17b follows directly.

For Eq. II-17a, let $x = 1 - e^{-z/n}$, and rewrite the logarithm of the term in parentheses in Eq. II-18:

$$\log(2-e^{-hz})^{hn} = hn\log(1+x) \leq hn\left(x - \frac{x^2}{2} + \frac{x^3}{3}\right) \tag{II-20}$$

If we use the inequalities,

$$\frac{z}{n} - \frac{1}{2}\left(\frac{z}{n}\right)^2 \leq x \leq \frac{z}{n} - \frac{1}{2}\left(\frac{z}{n}\right)^2 + \frac{1}{6}\left(\frac{z}{n}\right)^3 \leq \frac{z}{n} \tag{II-21}$$

the bound on Eq. II-20 becomes

$$\log(2-e^{-hz})^{hn} \leq hn\left(\frac{z}{n} - \left(\frac{z}{n}\right)^2 + \left(\frac{z}{n}\right)^3\right) \tag{II-22}$$

Inserting Eq. II-22 in Eq. II-18, we obtain

$$T(z) \geq 1 - e^{-hn(z/n)^2(1-z/n)} \qquad \text{(II-23)}$$

from which Eq. II-17a follows.

Acknowledgment

The author wishes to take this opportunity to thank his supervisors and fellow members of the Statistical Communication Theory Group, Research Laboratory of Electronics, M.I.T., for their advice and encouragement during the progress of this research. In particular, he is deeply grateful to Professor Yuk-Wing Lee, who devotes so much thought and effort to aiding students in making the difficult transition from being absorbers of knowledge to becoming independent workers, and to Professor Edward Arthurs for encouraging him to enter the field of graph theory, and for his subsequent interest. The author is also greatly indebted to his fellow-student, David A. Chesler, who, in many discussions, helped to clarify and strengthen his results.

References

1. W. Feller, An Introduction to Probability Theory and Its Applications, Vol. 1 (John Wiley and Sons, Inc., New York, 1950).
2. Ibid., p. 75; cf. Eq. (6.7).
3. Ibid., p. 76; cf. Problem 8.
4. G. W. Ford and G. E. Uhlenbeck, Combinatorial problems in the theory of graphs, Proc. Nat. Acad. Sci. 42, 122-128; 203-208 (1956).
5. E. N. Gilbert, Enumeration of labelled graphs, Can. J. Math. 8, 405-411 (1956).
6. E. N. Gilbert, Random graphs, Internal Technical Memorandum, Bell Telephone Laboratories, Inc., New York, Nov. 24, 1958 (to be published in Ann. Math. Stat.).
7. Ibid., see Table I.
8. F. Harary, The number of linear, directed, rooted, and connected graphs, Trans. Am. Math. Soc. 78, 445-463 (1955).
9. F. Harary, Graph theory and electric networks, Trans. IRE, vol. PGCT-6, pp. 95-109 (May 1959).
10. I. M. Jacobs, Connectivity in random graphs — Two limit theorems, Quarterly Progress Report No. 52, Research Laboratory of Electronics, M.I.T., Jan. 15, 1959, pp. 64-70.
11. I. M. Jacobs, Path and tree probabilities for a complete graph, Quarterly Progress Report No. 53, Research Laboratory of Electronics, M.I.T., April 15, 1959, pp. 76-83.
12. I. M. Jacobs, An upper bound on tree probability in a probabilistic graph, Quarterly Progress Report No. 54, Research Laboratory of Electronics, M.I.T., July 15, 1959, pp. 113-117.
13. M. Kochen, Extension of Moore-Shannon model for relay circuits, IBM J. 3, 169-186 (April 1959).
14. H. Mine, Reliability of physical systems, Trans. IRE, vol. PGCT-6, pp. 138-150 (May 1959).
15. E. F. Moore and C. E. Shannon, Reliable circuits using less reliable relays, J. Franklin Inst. 262, 191-208; 281-287 (1956).
16. F. Moskowitz, The analysis of redundancy circuits, Communications and Electronics 39, 627-632 (November 1958).
17. F. Moskowitz and J. B. McLean, Some reliability aspects of systems design, Trans. IRE, PGRQC-8, pp. 7-35 (September 1956).
18. J. von Neumann, Probabilistic logics and the synthesis of reliable organisms from unreliable components, Automata Studies, edited by C. E. Shannon and J. McCarthy, Annals of Mathematics Studies No. 34 (Princeton University Press, Princeton, N.J., 1956), pp. 43-98.
19. K. Pearson (ed.), Tables of the Incomplete Beta-Function (Cambridge University Press, Cambridge, 1934).
20. G. Pólya, Kombinatorische Anzahlbestimmungen für Gruppen, Graphen, und chemische Verbindungen, Acta Math. 68, 145-253 (1937).
21. F. Reza, Some topological considerations in network theory, Trans. IRE, vol. PGCT-5, pp. 30-41 (March 1958).
22. J. Riordan, An Introduction to Combinatorial Analysis (John Wiley and Sons, Inc., New York, 1958).
23. Ibid., p. 5; cf. Eq. (7).
24. Ibid., p. 100; cf. Problem 2.

25. Ibid., p. 17; cf. Example 14.

26. C. E. Shannon, Information Theory Seminar Notes, Massachusetts Institute of Technology, 1956 (unpublished).

27. L. Weinberg, Kirchhoff's "Third and Fourth Laws," Trans. IRE, vol. PGCT-5, pp. 8-30 (March 1958).

28. O. Wing, Reliability study of communication systems, Technical Report T-32/B, Columbia University, New York, Sept. 15, 1958.