RELIABLE CIRCUITS USING LESS RELIABLE RELAYS

BY

E. F. MOORE [1] AND C. E. SHANNON [1]

ABSTRACT

An investigation is made of relays whose reliability can be described in simple terms by means of probabilities. It is shown that by using a sufficiently large number of these relays in the proper manner, circuits can be built which are arbitrarily reliable, regardless of how unreliable the original relays are. Various properties of these circuits are elucidated.

Part I [2]

INTRODUCTION

In an important paper [3] von Neumann considers the problem of constructing reliable computing circuits by the redundant use of unreliable components. He studies several cases, one of which, for example, involves the construction of machines using as a basic component a "Sheffer stroke" organ.[4] Von Neumann shows that under certain conditions it is possible to combine a number of unreliable Sheffer stroke organs to obtain an element which acts like a Sheffer stroke organ of higher reliability. In fact, under certain conditions one can approach perfect operation by means of a sufficiently redundant circuit.

The present paper was inspired by von Neumann's work and carries out a similar analysis for relay circuits. It appears that relays are basically more adaptable to these error-correcting circuits than the neuron-like components studied by von Neumann. At any rate, our results go further than his in several directions.

In the first place, von Neumann needs to assume a certain fairly good reliability in his components in order to get started. With the Sheffer stroke organ, a probability of error less than 1/6 is absolutely necessary, and something like one in a hundred or better is required in the specific error-correcting circuits developed. The methods developed here, on the other hand, will apply to arbitrarily poor relays.

Secondly, the amount of redundancy required in our circuits for a

[1] Murray Hill Laboratory, Bell Telephone Laboratories, Inc., Murray Hill, N. J.

[2] Part II will appear in this JOURNAL for October, 1956.

[3] J. VON NEUMANN, "Probabilistic Logics," California Institute of Technology, 1952. (Also Published in "Automata Studies," edited by C. E. Shannon and J. McCarthy, Princeton University Press, 1956.)

[4] The Sheffer stroke is the logical operation on two variables "not A and not B." It has the property that all logical functions can be generated in terms of it. A Sheffer stroke *organ* is a device with two binary inputs and one binary output which performs this logical operation. An unreliable component of this sort would give the proper output only with a certain probability.

given improvement in reliability is considerably different from that required by von Neumann. For example, in one numerical case that he considers, a redundancy of about 60,000 to 1 is required to obtain a certain improvement in operating reliability. The same improvement is obtained in relay circuits with a redundancy of only 100 to 1. We also show that in a certain sense some of our circuits are not far from minimal. Thus, in the numerical case just mentioned, our results show that a redundancy of at least 67 to 1 is necessary in any circuit of the type we consider. Hence, the actual circuits which achieve this improvement with a redundancy of 100 to 1 are not too inefficient in the use of components.

Another difference is that it is not necessary in the case of relays to use what von Neumann calls the "multiplexing system" in order to approach perfect operation on the final output. With his types of elements, the final output (without multiplexing) always has a definite residual unreliability. With the systems described here, this final probability of error can approach zero.

This paper is not intended for practical design purposes, but rather for theoretical and mathematical insight into the problem. There may, however, be some practical applications. The reliability of a commercial relay is typically very high, for example, one failure in 10^7 operations. However, there are cases where even this reliability is insufficient. In the first place, in large-scale computing machines an extremely large number of individual relay operations may be involved in one calculation, an error in any one of which could cause an error in the final result. Because of this, the Bell Telephone Laboratories' computers have made extensive use of self-checking and error-detecting schemes. A second type of situation requiring extreme reliability occurs when human safety is dependent on correct operation of a relay circuit, for example, railway interlocks, safety circuits on automatic elevators and in guided missiles, etc. It is possible that some of the simpler circuits we describe may be of some use in applications such as these. However, the results of this paper will not be directly applicable to actual relays which wear out with age, but only to idealized relays whose probability of failure are constant in time.

IDEALIZED RELAYS

We will prove results only for idealized relays whose failures can be described in one specific manner by means of probabilities. Their description allows only intermittent types of failures, and allows these only under the assumption that the probability of failure remains constant as time passes.

This idealization does not cover such actually possible cases as relays which wear out with age, relays whose windings burn out, or relays which have been wired into the circuit with an imperfect soldered

connection. It is also assumed that the circuit is not improperly designed or improperly wired and that there are no bits of solder to produce short circuits between different wires.

Since all of the above kinds of errors and failures can actually occur in practice, using real relays, the results of this paper do not strictly apply to such real relays. However, the two kinds of failures considered in this paper do actually occur in relays, so the kinds of circuits suggested are of some possible application.

The first kind of failure allowed is the failure of a relay contact to close, which in actual relays is often due to a particle of dust preventing electrical closure.

The second type of failure is the failure of a contact to open, which in actual relays is usually due to the welding action of the current passing through the contacts. We shall consider relay circuits in which the only causes of errors are of these two types—failure of contacts that should be closed to be actually closed and of contacts that should be open to be actually open. We will assume, in fact, that there are two probabilities associated with a contact on a relay. If the relay is

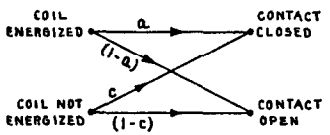


FIG. 1. Schematic representation of the transition probabilities.

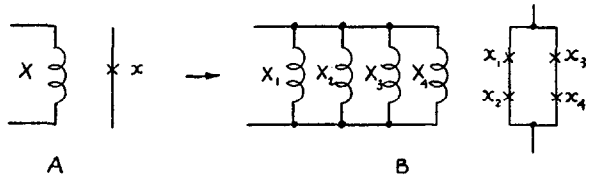


FIG. 2. One proposed way of transforming relay circuits to improve reliability.

energized, the contact is closed with probability a, open with probability $1 - a$. If the relay is not energized, the contact is closed with probability c and open with probability $1 - c$. If a is greater than c, we will call the contact a make contact; if a is less than c we call it a break contact. We assume that different contacts are statistically independent. With actual relays this is probably not too far from the truth for contacts on *different* relays and, indeed, this is all that is required for most of the results we wish to establish. In addition, we shall assume that on the successive times that a relay coil is energized its closures are statistically independent.

A relay of this type governed by probabilities a and c will be called a *crummy*[5] relay. Its probability operation may be represented schematically as in Fig. 1. This will be recognized as similar to diagrams used to represent a simple noisy communication channel, and indeed such a relay can be thought of as a noisy binary channel. The capacity of the corresponding channel will be zero if and only if $a = c$. We will

[5] "Crummy = crumby, esp. lousy," Webster's New International Dictionary. We chose the more modern spelling universally used in comic books.

see later that highly reliable computers can be constructed from a sufficient number of crummy relays if and only if $a \neq c$.

THE GENERAL METHOD OF IMPROVING RELIABILITY

In a general way the analysis we will give depends on constructing networks of contacts which act like a single contact but with greater reliability than the contacts of which they are composed. For example, in Fig. 2*A*, we have a crummy relay X with a make contact x. This relay might appear as a part of a large computing circuit. In Fig. 2*B* we replace this by four crummy relays X_1, X_2, X_3, X_4 whose coils in parallel replace the single coil X, and whose contacts are in the series parallel combination shown, this two-terminal circuit replacing the single previous x contact. If each of these four contacts has the probability p of being closed, it is easily seen that the probability of the four-contact circuit being closed is

$$h(p) = 1 - (1 - p^2)^2 = 2p^2 - p^4.$$

This function is plotted in Fig. 3. It will be seen that it lies above the diagonal line $y = p$ for p greater than 0.618 and lies below the line for

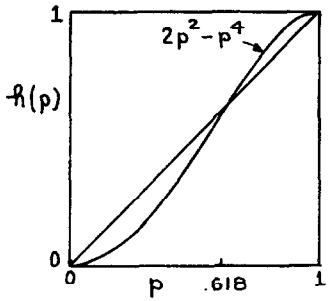
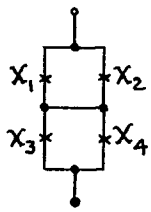


FIG. 3. The function describing the behavior of Fig. 2*B*.



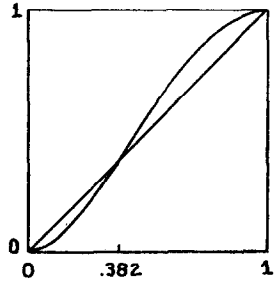


FIG. 4. Another series-parallel circuit and its associated function.

p less than 0.618. This means that if 0.618 is between the a and c of Fig. 1, Fig. 2*B* will act like a relay with better values of a and c, that is, values nearer to zero and one. For example, if the individual relays made errors with probabilities $1 - a = c = 0.01$, the circuit of Fig. 2*B* would make errors when the coils are energized with probability 0.000396, and when the coils are not energized with probability 0.0002. Thus a large improvement in reliability, both when the coil is energized and when it is not energized, is obtained by the use of this circuit.

Figure 4 shows another contact arrangement giving rise to a somewhat different function

$$h(p) = [1 - (1 - p)^2]^2 = 4p^2 - 4p^3 + p^4.$$

Here again, $h(p)$ is the probability of the network being closed, when the individual contacts each have probability p of being closed. The network of Fig. 4 is the dual of that in Fig. 2, and the curve is that obtained by interchanging 0 and 1 in both abscissa and ordinate in Fig. 3.

The bridge network of Fig. 5 gives rise to a symmetrical curve crossing the diagonal at $p = 0.5$. For this network we have:

$$h(p) = 2p^2 + 2p^3 - 5p^4 + 2p^5.$$

All of these networks tend to accentuate the nearness of p to its values 0 or 1 and thus tend to improve reliability. Many other networks have similar properties as we shall see. Furthermore, we will show that it is possible to find a network whose curve, Fig. 6, crosses the diagonal line for a value of p between any two given numbers a and c (no matter how close together) and in fact is less than δ at a and greater than $1 - \delta$ at c, for any positive δ. This means that an arbitrarily good relay can be made from a sufficient number of crummy relays.

It may be seen that this general procedure operates to improve the reliability of either make or break contacts. The only difference is the labeling of the points a and c.

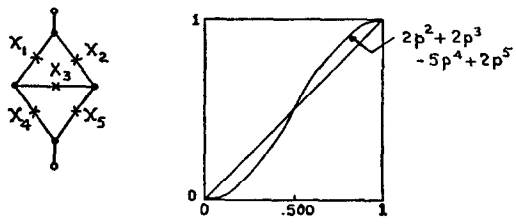


FIG. 5. A bridge circuit and its associated function.

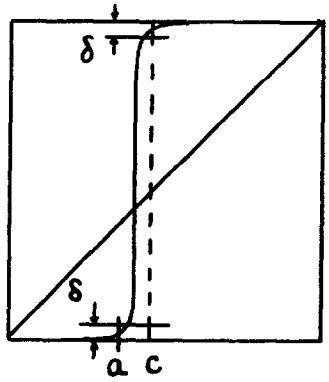


FIG. 6. The general form of curve of attainable functions.

PROPERTIES OF $h(p)$

Consider any two-terminal network made up of contacts each of which has a probability p of being closed. The network will have a probability, say $h(p)$, of being closed. We wish to investigate some of the properties of $h(p)$.

In the first place, $h(p)$ is a polynomial and may be written as follows:

$$h(p) = \sum_{n=0}^{m} A_n p^n (1 - p)^{m-n} \qquad (1)$$

where m is the total number of contacts in the network and A_n is the number of ways we can select a subset of n contacts in the network such that if these n contacts are closed, and the remaining contacts open, then the network will be closed. This is evident since (1) merely sums up the probabilities of the various disjoint ways that the network could be closed.

The first non-vanishing term in (1), say $A_s p^s (1-p)^{m-s}$, is related to the shortest paths through the network from one terminal to the other —s is the length of these paths and A_s the number of them. This is because in (1) all the elements of a subset which contribute to A_s must actually be on the path (otherwise A_s would not have been the first non-vanishing term). We will call s the *length* of the network. It is evident from (1) that near $p = 0$ the function $h(p)$ behaves as $A_s p^s$.

In a similar way, one can work with the probability of the network being open and write

$$1 - h(p) = \sum_{n=0}^{m} B_n (1-p)^n p^{m-n} \tag{2}$$

where B_n is the number of subsets of n contacts such that, if all contacts in a subset are open and the other contacts closed, the network is open. The first non-vanishing term in this series, say $B_t (1-p)^t p^{m-t}$, relates to the smallest cut sets of the network (sets of contacts which, if opened, open the network). Here t is the number of contacts in these minimal cut sets, and B_t the number of such cut sets. The reason is essentially as before. We will call t the *width* of the network. It is evident that, in the neighborhood of $p = 1$, $h(p)$ behaves as $1 - B_t (1-p)^t$.

The function $h(p)$ may also be calculated by other means. For example, fix attention on a particular contact in the network, N. Calculate the probability function for the network obtained from N by replacing this contact with a short circuit, say $f(p)$, and for the network obtained from N by replacing this contact with an open circuit, say $g(p)$. Then clearly,

$$h(p) = pf(p) + (1-p)g(p). \tag{3}$$

Furthermore we will have, whenever $0 \leq p \leq 1$,

$$f(p) \geq g(p). \tag{4}$$

This is intuitively evident since closing a connection certainly cannot decrease the probability of the network being closed. Formally, it follows from the relation (1), noting that the cases where the g network is closed are a subset of those in which f is closed, and consequently the terms in the expression for f dominate those in the expression for g.

If the network in question is planar, it will have a dual. Let $h_D(p)$ be the probability function for this dual network. For each

state of the contacts of the original network let us make correspond in the dual network the state in which corresponding contacts have the opposite value. Then states for which the original network is open correspond to states for which the dual network is closed. If the probability of closure of a contact in the dual network is $1 - p$, where p is the probability of closure in the original network, then the probabilities of corresponding states are equal. Consequently we will have

$$1 - h_D(1 - p) = h(p). \tag{6}$$

An example of this relation between the h functions for a network and its dual is given in Figs. 3 and 4. Either of these graphs can be obtained from the other by inverting, that is, by interchanging 0 and 1 in both abscissa and ordinate.

If the network is self-dual (for example the bridge of Fig. 5),

$$1 - h(1 - p) = h(p). \tag{7}$$

Substituting $p = 1/2$, we find $h(1/2) = 1/2$.

COMBINATION OF TWO NETWORKS

Consider now two networks N_1 and N_2 with functions $h_1(p)$ and $h_2(p)$. If N_1 and N_2 are connected in series, Fig. 7, the resulting net-

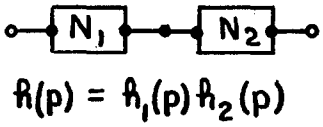


FIG. 7. Connection of two networks in series.

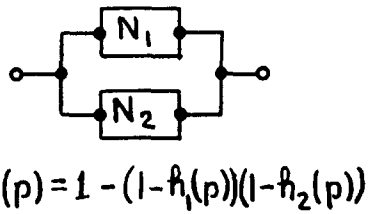


FIG. 8. Connection of two networks in parallel.

work will be closed only if both parts are closed. Hence, the resulting $h(p)$ function will be given by the product $h_1(p)\ h_2(p)$.

If N_1 and N_2 are connected in parallel, Fig. 8, the resulting network will be open only if both parts are open, an event with probability $(1 - h_1)(1 - h_2)$. Hence, the resulting $h(p)$ function for the parallel network will be $[1 - (1 - h_1)(1 - h_2)]$.

A third method of combining the two networks N_1 and N_2 is by "composition." By this we mean replacing each element of N_1 by a copy of N_2, as shown for a typical example by Fig. 9. It is evident that the composite network has an h function given by the composition of the two original h functions:

$$h(p) = h_1(h_2(p)). \tag{8}$$

If N_1 and N_2 are identical and this process is repeated $n-1$ times, we obtain the n^{th} composition of h with itself, which we denote by

$$h^{(n)}(p) = h(h(h\cdots h(p)\cdots)).$$

The value of $h^{(n)}(p)$ can be found readily from the $h(p)$ curve by the staircase construction shown typically in Fig. 10 for $h^{(3)}(p_1)$. Thus, by composition, a greater improvement in reliability may be obtained with networks whose $h(p)$ curve crosses the diagonal but once. This effect, and the improvement by iteration relating to the staircase construction of Fig. 10, are very similar to situations in von Neumann's approach.

BOUNDS ON $h'(p)$

We will now deduce an interesting inequality concerning the slope of possible functions $h(p)$. As a corollary, we will show that any $h(p)$ function can cross the diagonal at most once.

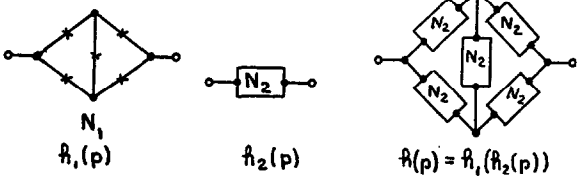


FIG. 9. Composition of two networks.

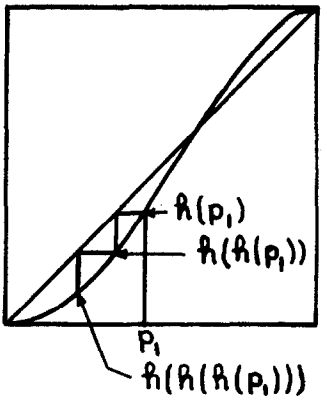


FIG. 10. The effect of iterated composition.

Theorem 1

$$\frac{h'(p)}{(1-h(p))h(p)} > \frac{1}{(1-p)p} \quad \text{whenever} \quad 0 < p < 1, \tag{9}$$

provided $h(p)$ is neither identically zero, identically one, nor identically equal to p.

This will be proved by an induction on the number of contacts in the network. We expand $h(p)$ as in (3) except that we expand it about some contact which lies on a path through the network, and then we assume that either the inequality holds for the functions f and g, or that they are among the three exceptional functions, and then we

prove the inequality for the function h. But since the contact actually lies on a path, the proof of (4) gives that $f(p) < g(p)$ for all p. Also we cannot have $1 - f(p) + g(p) = 0$ for any p, for if so, we would have $f(p) = 1$ and $g(p) = 0$, which implies there is no path through the network of g, and no cut set through the network of f, and hence $f(p) = 1$ and $g(p) = 0$ for all p, hence $h(p) = p$, contradicting the hypotheses of the theorem.

It can be seen that

$$(1 - p)p(f - g)(1 - f + g) > 0 \text{ whenever } 0 < p < 1, \quad (10)$$

since each of the terms is positive. Multiplying out,

$$pf - pg - pf^2 + 2pfg - pg^2 - p^2f + p^2g + p^2f^2 - 2p^2fg + p^2g^2 > 0.$$

Rearranging and factoring

$$-pf^2 + (1 - p)pf - (1 - p)g^2 - (1 - p)pg > \\ - [p^2f^2 + (1 - p)^2g^2 + (1 - p)2pfg].$$

Adding $pf + (1 - p)g$ to each side,

$$(1 - f)pf + (1 - p)pf + (1 - p)(1 - g)g - (1 - p)pg \\ > pf + (1 - p)g - [pf + (1 - p)g]^2 = h - h^2 = (1 - h)h. \quad (11)$$

Now, since by inductive assumption either $\frac{f'}{(1 - f)f} > \frac{1}{(1 - p)p}$, or we have one of the three exceptional functions, we have in any case that $(1 - f)f \leq (1 - p)pf'$ and similarly $(1 - g)g \leq (1 - p)pg'$. Using these in the left member of (11) we obtain

$$(1 - p)p^2f' + (1 - p)pf + (1 - p)^2pg' - (1 - p)pg > (1 - h)h.$$

Dividing by $(1 - p)p$,

$$pf' + f + (1 - p)g' - g > \frac{(1 - h)h}{(1 - p)p},$$

or

$$\frac{d}{dp}(pf + (1 - p)g) > \frac{(1 - h)h}{(1 - p)p},$$

$$\frac{h'}{(1 - h)h} > \frac{1}{(1 - p)p},$$

completing the proof.

If we replace the inequality (9) in the statement of the theorem

by an equality, that is if we set $\frac{y'}{(1-y)y} = \frac{1}{(1-p)p}$, we have a differential equation, the solutions of which form a one-parameter family of curves. The inequality (9) states that the permissible h functions corresponding to contact networks must have slopes greater than these y curves. If we solve this differential equation for the y curves we obtain

$$\frac{y(p)}{1-y(p)} = C\frac{p}{(1-p)}. \tag{12}$$

This family of curves is plotted in Fig. 11 for $C = 1/4, 1/3, 1/2, 1, 2, 3, 4$. Any possible $h(p)$ function must cross curves of this family with

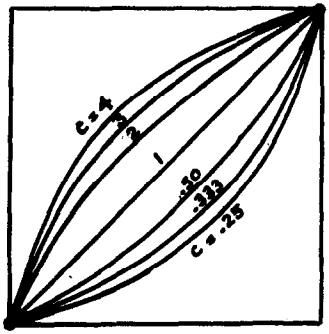


FIG. 11. The family of curves satisfying the equation

$$\frac{y(p)}{1-y(p)} = C\frac{p}{(1-p)},$$

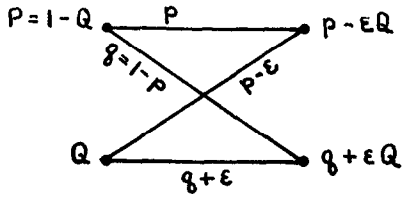


FIG. 12. A binary channel used to obtain an upper bound on the slope $h'(p)$.

a greater slope. Consequently, any $h(p)$ curve can cross one of these curves at most once in the open interval $0 < p < 1$. Since the straight line of slope 1 which goes through the origin is one member of this family, any $h(p)$ curve can cross this line once at most, say at the point $p = p_0$. Then applying the staircase construction as shown in Fig. 10, it can be seen that $h^{(n)}(p)$ approaches 0 as a limit for all $p < p_0$, and approaches 1 for all $p > p_0$. Thus any network whose $h(p)$ curve crosses this diagonal straight line can be composed with itself to obtain a network which improves reliability. In fact if we iterate the composition n times, we will have

$$\lim_{n\to\infty} h^{(n)}(p) = \begin{cases} 1 & p > p_0 \\ p_0 & p = p_0 \\ 0 & p < p_0 \end{cases}$$

where p_0 is the (necessarily unique) diagonal crossing point.

It is possible to place an *upper* bound on the slope $h'(p)$ by a curious

argument involving information theory. Consider the binary channel shown in Fig. 12. The rate of transmission for this channel will be

$$\begin{aligned} R &= H(y) - H_x(y) \\ &= -(p - \epsilon Q)\log(p - \epsilon Q) - (q + \epsilon Q)\log(q + \epsilon Q) \\ &\quad + (1 - Q)(p \log p + q \log q) \\ &\quad + Q[(p - \epsilon)\log(p - \epsilon) + (q + \epsilon)\log(q + \epsilon)]. \end{aligned}$$

For ϵ approaching zero, $(a + \epsilon)\log(a + \epsilon)$ is approximated by its Taylor series

$$a \log a + (1 + \log a)\epsilon + \frac{\epsilon^2}{a} + \cdots.$$

Using this in the above for all terms containing ϵ, we find that the constant terms and first order terms in ϵ vanish. The first non-vanishing terms are given by

$$R = (Q - Q^2)\frac{\epsilon^2}{pq} = \left[\frac{1}{4} - \left(Q - \frac{1}{2}\right)^2\right]\frac{\epsilon^2}{pq}.$$

It is evident from this last expression that R is maximized (when we vary Q) by $Q = 1/2$. This maximum R is, by definition, the channel capacity C. Thus as ϵ approaches zero in Fig. 12, the capacity C is asymptotic to $\frac{\epsilon^2}{4pq}$.

Now consider a crummy relay which has probability p of being closed when the relay is energized and $p - \epsilon$ of being closed when the coil is not energized. The relay may be thought of as a communication channel for which the coil is the input and the contact the output. If ϵ is very small, the capacity will be $\frac{\epsilon^2}{4pq}$. If we have n relays, with the same p and ϵ, the total capacity of this sytem, using the n coils as input and the n contacts as output, is $n\epsilon^2/4pq$, since the capacity of a set of independent channels is the sum of the individual capacities.

We wish to show from these capacity considerations that the probability function $h(p)$ for our contact networks must satisfy

$$\frac{dh}{dp} \leq \sqrt{\frac{n(1 - h)h}{(1 - p)p}}. \tag{13}$$

Consider a network N with n contacts and probability function $h(p)$. Let the individual relays and contacts have probabilities p_1 and ϵ as in Fig. 12. Then the network as a whole acts like a single relay with

parameters $h(p_1)$ and $h_2'(p_1)\epsilon$, (when ϵ is small). As such, it has a capacity $(h'\epsilon)^2/4(1-h)h$. This capacity must be less than or equal to that obtained when these n relays are used in the best possible way. Hence,

$$\frac{(h'\epsilon)^2}{4(1-h)h} \leq \frac{n\epsilon^2}{4(1-p_1)p_1}.$$

This being true for any p_1, we have, rearranging terms, the desired result

$$h' \leq \sqrt{\frac{n(1-h)h}{(1-p_1)p_1}}.$$

If this inequality is changed to an equality, we obtain the differential equation

$$\frac{\sqrt{n}\,dp}{\sqrt{(1-p)p}} = \frac{dh}{\sqrt{(1-h)h}}$$

the solution of which is

$$\sqrt{n}\sin^{-1}(1-2p) = \sin^{-1}(1-2h) + \theta. \tag{14}$$

For a given number of contacts n, a possible $h(p)$ curve must cross the corresponding family of curves (14) always with less or equal slope.

Another sort of upper bound on $h(p)$ functions obtained from n contacts can be found by a different argument. A two-terminal network corresponds to a Boolean function of the n contacts involved. However, it is not possible to realize all Boolean functions using only one make contact for each variable. Suppose we ignore these conditions of realizability and consider the class of all Boolean functions of n variables. For any such Boolean function there will be an $h(p)$ function, $h(p)$ being the probability that the function is equal to one if each variable has the (independent) probability p of being equal to one. Which Boolean functions have $h(p)$ functions with the greatest slopes and show the greatest sharpening effect on probabilities?

A Boolean function of n variables will be called a *quorum* function if there is some s, $0 \leq s \leq n$, such that if less than s of the variables are one the function is zero, and if more than s of the variables are one the function is one.

Theorem 2

If the h curve for any quorum function of n variables, say $h_Q(p)$, crosses the h curve of any other Boolean function of n variables, say $h(p)$, then at the point of crossing p_0 we have

$$h'(p_0) < h_Q'(p_0)$$

that is, the quorum function has the greater slope. Furthermore,

$$h(p) > h_Q(p) \qquad 0 < p < p_0$$
$$h(p) < h_Q(p) \qquad p_0 < p < 1.$$

This theorem says that, in a certain sense, the quorum functions are the best of all Boolean functions for our purposes of increasing reliability.

Proof: For any Boolean function of n variables, the $h(p)$ polynomial is made up of a sum of terms of the form p^iq^{n-i}, a term of this form for each state of the variables for which the Boolean function has the value one with i of the variables equal to one. A quorum function has the value one for all states with i less than s, say, and zero for all states with i greater than s. Hence the $h_Q(p)$ function is of the form

$$h_Q(p) = \sum_{i=0}^{s-1} \binom{n}{i} p^iq^{n-i} + Ap^sq^{n-s}. \quad 0 \leq A \leq \binom{n}{s}$$

Since h is not identical with h_Q but is equal in value to it at p_0, it follows that the h polynomial must miss some terms before (or at) i equals s and have some extra ones after (or at) i equals s. In other words, we can write

$$h(p) = \sum_{i=0}^{n} B_ip^iq^{n-i}$$

with $B_i \leq \binom{n}{i}$ Let $C(p) = \sum_{i=0}^{\tau} B_ip^iq^{n-i} + \alpha p^sq^{n-s}$ where α is B_s or A, whichever is smaller. Then we will have

$$h_Q(p) = C(p) + \sum_{i=0}^{\tau} D_ip^iq^{n-i}$$

$$h(p) = C(p) + \sum_{i=\tau+1}^{n} E_ip^iq^{n-i} \tag{15}$$

where the D_i and E_i are non-negative integers and τ is $s-1$ or s according as B_s or A was smaller.

Now we note that for an expression of the form $u(p) = p^iq^{n-i}$ we have

$$u'(p) = i\,p^{i-1}q^{n-i} - (n-i)p^iq^{n-i-1}$$
$$= \left(\frac{i}{p} - \frac{n-i}{q}\right) u(p) = \frac{i-pn}{pq}\,u(p).$$

Thus $\frac{u'}{u} = \frac{i - pn}{pq}$ is a monotone increasing function of i. Now all the terms in the sum in (15) for h_Q correspond to smaller values of i than those in the sum for h. If we let $u_Q(p)$ stand for any term in the sum in h_Q and $u(p)$ stand for any term in the sum in h, we will have

$$\frac{u_Q'}{u_Q} < \frac{u'}{u}$$

and hence there will exist a constant K such that

$$\frac{u_Q'}{u_Q} < K < \frac{u'}{u},$$

and

$$u_Q' < Ku_Q, \qquad Ku < u'.$$

Summing the first inequality over all the different terms u_Q, and the second over all the u, we obtain

$$\sum u_Q' < K \sum u_Q, \qquad K \sum u < \sum u'.$$

But evaluating at p_0, we have $\sum u_Q = \sum u$, and consequently

$$\sum u_Q' = \sum u',$$
$$h_Q'(p_0) < h'(p_0).$$

The remainder of the theorem follows readily by noting that to contradict it, since the h and h_Q curves are continuous, would require that they cross at a point different from p_0 and in such a way as to contradict the first part of the theorem.

NETWORKS OF A GIVEN LENGTH AND WIDTH

We have seen that the orders of flatness of $h(p)$ in the neighborhoods of $p = 0$ and $p = 1$ are related to the "length" and "width" of the network in question. It is clear that in the case of practical importance, the values of p of interest will be in these neighborhoods, that is, the relays will be initially quite reliable. In this section we will develop some results relating these orders of flatness with the number of elements in the network.

Theorem 3

If a network N has length l and width w it contains at least lw contacts. Equivalently, if $h(p)$ behaves like Ap^l near $p = 0$, and if $1-h(p)$ behaves like $B(1 - p)^w$ near $p = 1$, the corresponding network contains at least lw contacts.

Proof: We associate an integer with each contact in N by the following process. Contacts directly connected to the left terminal of N are labeled "1," contacts connected to those labeled 1 but not already labeled are numbered "2," and so on inductively. In general, a contact will be labeled n if it is possible to find a path to the left terminal through $n - 1$ other contacts but there is no such path through a smaller number.

The set of contacts labeled n for any particular n from 1 to l will be shown to form a cut set of the network. This is true since every path through the network starts at the left terminal with a contact labeled 1 and ends at the right terminal with a contact labeled l or more (if any of the contacts touching the right terminal were labeled with numbers less than l the length of N would be less than l). Along any path, the numbering changes by 0 or ± 1 in going from one contact to the next. Hence every path in going from contacts numbered

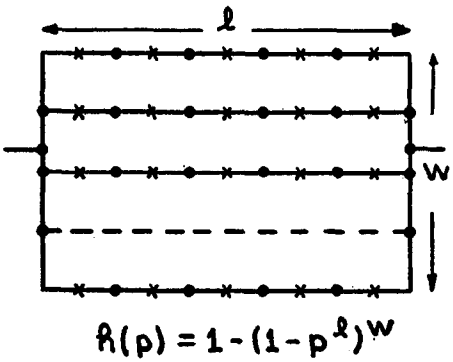


FIG. 13. A series-parallel network of length l and width w.

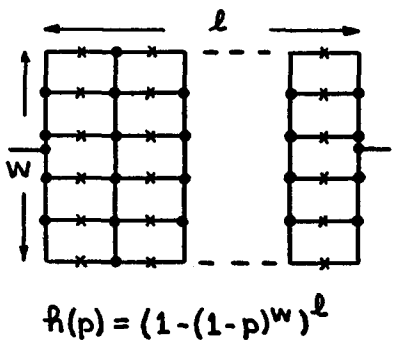


FIG. 14. Another series-parallel network of length l and width w.

1 to those with numbers $\geq l$ must pass through every intermediate value. Consequently if all contacts labeled n (for $1 \leq n \leq l$) are deleted from N, all paths are broken and these contacts thus form a cut set.

Since the network is of width w, every cut set contains at least w contacts. Thus there are at least w contacts labeled 1, at least w labeled 2, $\cdots$, and at least w labeled l. The network therefore contains at least wl contacts.

The alternative statement of Theorem 3 follows from remarks made in connection with Eqs. 1 and 2.

It is possible to achieve the "dimensions" l and w with *exactly* lw contacts in a wide variety of ways. For example, we can make a series chain of l contacts and parallel w copies of this (Fig. 13). Dually, w contacts can be paralleled and l copies of this placed in series (Fig. 14).

Theorem 4

A complete characterization of minimal networks with dimensions l and w is the following. Let Y and Z be the terminal nodes, s_0 be the set consisting of Y alone, and s_l be the set consisting of Z alone. In addition to s_0 and s_l there will be $l-1$ subsets of nodes $s_1, s_2, \cdots, s_{l-1}$. There will be precisely w elements connecting nodes in s_n to nodes in s_{n+1} $(n = 0, 1, \cdots, l-1)$. Finally, if any node in s_j has m elements connecting it to nodes in s_{j-1}, then it has m elements connected to nodes in s_{j+1} $(j = 1, 2, \cdots, l-1)$.

This means that any such minimal network with dimensions l and w can be obtained from the network of Fig. 13 by making appropriate connections among nodes in the same vertical line. When all the nodes in each vertical line are connected together, for example, the result is Fig. 14. Another possibility is shown in Fig. 15.

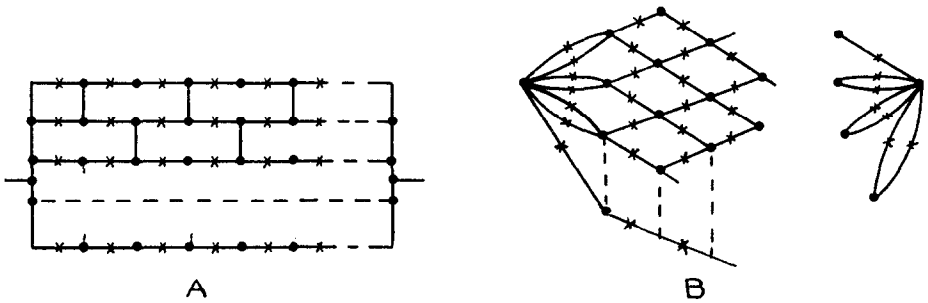


FIG. 15. A hammock network of length l and width w.

To show that any minimal lw network is of the form described in Theorem 4, first note that in our preceding proof, each of the numbered cut sets must contain precisely w elements, and these elements must run between elements of lower numbers and higher numbers. The nodes between elements numbered $j-1$ and j will belong to subset s_j in the above characterization. Now suppose that some node in s_j has m elements going to nodes in s_{j-1} and $m+p$ going to nodes in $s_{j+1}(p > 0)$. The elements numbered $j+1$ form a cut set of w elements. It is easily seen that if the $m+p$ members of this, going from the node in question, are replaced by the m elements going to nodes in s_{j-1}, then we will still have a cut set but one with less than w elements, a contradiction. Consequently any minimal network of dimensions l and w is of the type described in our characterization.

To show the converse, that any network of the type characterized has dimensions l and w, note first that to go from one terminal to the other the path must pass through nodes belonging to $s_1, s_2, \cdots, s_{l-1}$. Hence any path is of length at least l and the network is of length l. Now consider any cut set c. We will show that c contains at least w

elements. Consider the smallest-numbered contacts of c. Suppose one of these is connected from node A in s_{j-1} to node B in s_j. Then either all elements from B to nodes in s_{j-1} are in the cut set or the one in question is not essential to the cut set and may be eliminated, giving a still smaller cut set. In the former case, this group of elements can be replaced by an equal number, those going from node B to members of s_{j+1}, preserving the cut set property. Proceeding in this way, the cut set is gradually worked over toward the right-hand terminal, either reducing or keeping constant the number of elements in the cut set.

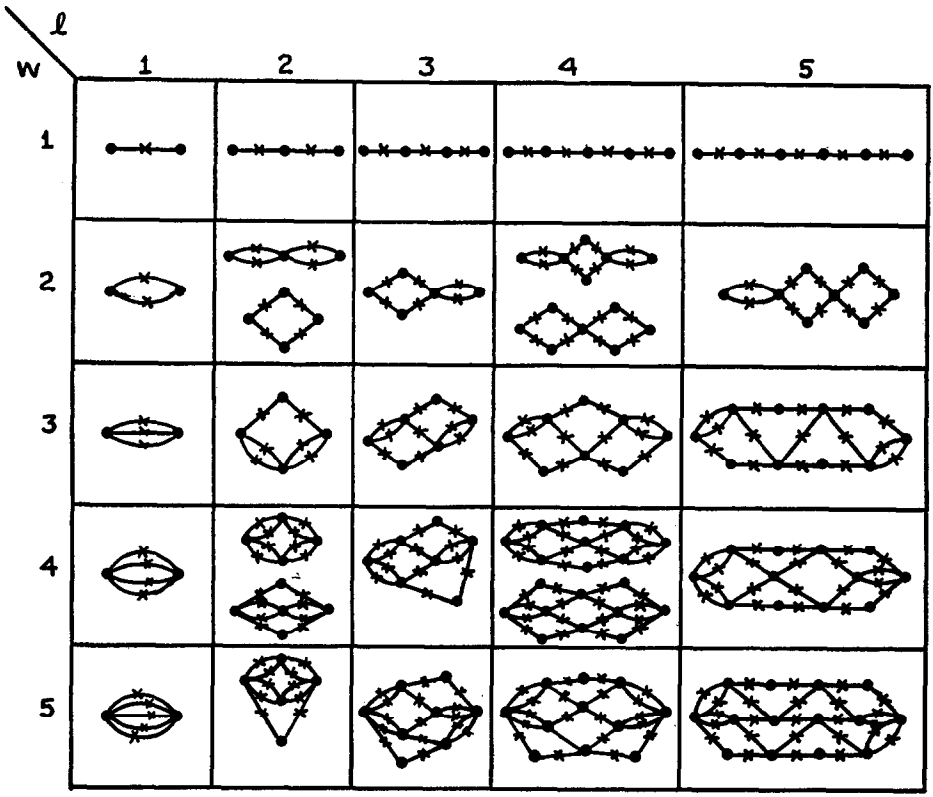


FIG. 16. Hammock networks of various lengths and widths.

When all the elements of the cut set are adjacent to the right-hand terminal there are exactly w members. Consequently there were at least that many in the original cut set, as we wished to prove.

An interesting type of minimal lw network is obtained by putting alternate connections in Fig. 13, leading to the brick-wall appearance of Fig. 15A. When redrawn, pulling together the vertical connections, the network appears as in Fig. 15B, and we will call networks of this type hammock networks. Figure 16 shows some of the simple cases

of hammock networks. It will be seen that if both l and w are even, there are two possible hammock networks with these dimensions. If either or both are odd, there is only one. Furthermore, the dual of a hammock network with length l and width w is a hammock network with length w and width l. These hammock networks are, in a sense, midway between the extreme minimal lw networks of Figs. 13 and 14, having half of the connections required to go from Fig. 13 to Fig. 14. In the case where l and w are equal and odd the (unique) hammock network is self-dual.

(To be continued)