On Definitions of Congestion in Communication Networks

By ERIC WOLMAN

(Manuscript received August 18, 1965)

Beginning with the familiar ideas of time-congestion and call-congestion for a full-access trunk-group, this paper considers the relations between various measures of congestion for networks of more general structure. The discussion is based on a simple heuristic model which makes the definitions of time- and call-congestion directly comparable. This model leads to a two-dimensional classification of measures of congestion, which should be useful in treating types of systems not mentioned here. Three importan papers in the literature are analyzed in terms of the proposed classification

I. INTRODUCTION

The concepts of *time-congestion* and *call-congestion* have been in common use since the early days of telephone traffic theory. Both ideas are quite simple when applied to a single set of devices used in telephony, such as a full-access trunk-group. But applications of these ideas in the theory of networks of more general structure, composed of elements arranged both in series and in parallel, have not always been consistent, either internally or with each other. This paper describes an attempt to resolve some of the difficulties, which may have arisen because the simplicity of many "classical" models renders unnecessary some distinctions which are important in the general case.

This section describes the nomenclature used below and the assumptions on which the remainder of the paper rests. The following section treats a useful model in the context of a full-access trunk-group. In the third section, similar ideas are applied to the theory of general communication networks. The fourth section contains a brief discussion of some of the switching literature, in the light of Section III. Some conclusions appear in the fifth section. In order to save space, I propose the abbreviations "CC" for call-congestion and "TC" for time-congestion.

I use the following terminology. A single element of a system carrying traffic is *busy* or *idle*. This choice is binary and tells whether or not the element can honor a request for service. When a set of elements is arranged so as to form all or part of a communication network, the events {path} and {no path} correspond respectively to the existence and non-existence of a chain of idle elements connecting two specified points of the network. A *call-attempt*, which is a request for an idle path connecting two points of a network, is *blocked* if no such path exists. Blocking is a binary concept. "Congestion" refers to a non-vanishing *probability of blocking*, which takes values on the interval [0,1].

Analysis of congestion refers to a specified portion of a communication system. This may be one group of trunks, considered in isolation; it may be the switching network connecting incoming to outgoing trunks in a tandem office; or it may be an entire system. When congestion exists, some call-attempts do, or could, fail. A call-attempt must be made by a *source* and must be directed to a *destination;* sources and destinations are *terminals*. Since we may think of call-attempts as simply appearing as inputs to a particular model, a terminal really marks the boundary of the model, which may or may not include the entities originally responsible for call-attempts.

But for stated exceptions, this paper should be read with the understanding that blocked calls are cleared, i.e., that call-attempts which cannot be served immediately are dismissed and have no effect on the system. With blocked calls delayed, there may be more than one relevant notion of blocking. One may ask for the probability of positive delay or of delay exceeding a fixed amount. If only finite queues are possible, one may want to know the probability of entering the queue or of overflow from the queue. The ideas expressed below can easily be extended to such systems, but for simplicity they are introduced here in the context of loss operation (blocked calls cleared), with or without retrials.

The traffic that a source or class of sources offers to a system is described by means of a random process which specifies the instants at which call-attempts occur. The parameters of such processes (with deterministic traffic as a special case) can vary with time, from source to source, or with the states of their sources, of the network, or of other sources. When an expected number of call-attempts in an interval of time is divided by the length of the interval, the quotient is an average *calling-rate*. The present method of elucidating congestion allows, and takes into account, dependence of a calling rate upon its source and upon the state of its source and of the network; but it is simplest, with one exception treated explicitly, to require independence among different sources. There remains only the question of time-dependence.

Let us assume that the stochastic process which describes the operation of the traffic system of interest is stationary. The method proposed below is based on subdividing a typical interval of time according to the states of various elements of the system. I write as if various quantities were *defined* as ratios of lengths of sub-intervals of an interval of finite length T. For every kind of time-congestion (TC), the definition is actually the limit, under appropriate conditions, of such a ratio as $T \to \infty$. For the sake of shortening complicated statements, this important distinction will not be mentioned in Sections II–IV.

Similarly the definition of call-congestion (CC), which is what many authors mean by "probability of blocking", is often for the stationary case given as the limit as $T \to \infty$ of the ratio of the number of *blocked* call-attempts (in $[0,T)$) to the number of *all* attempts (in the same interval). I propose to describe CC and TC in comparable terms by equating call-congestion in a typical case with such a ratio as

[(expected length of blocked time in T when attempts are possible)
 $\cdot$ (expected calling-rate when attempts are possible but blocked)]
$\div$ [(expected length of time in T when attempts are possible) $\cdot$ (expected
 calling-rate when attempts are possible)].

This procedure, of replacing an expected number of calls by a product of an expected length of time and a conditional expected calling-rate, is valid in cases for which a fraction such as that displayed here has a limit as $T \to \infty$ which agrees with the corresponding true CC as defined above. The procedure is certainly valid when the traffic system can be described by a stationary Markov chain, as is the case for many systems representable by models with finitely or infinitely many sources, blocked calls cleared or delayed, etc. But so far as I am aware, it is not now known either exactly what systems can be so described in a tractable way, or for what other systems the desired agreement holds. Thus further discussion of the applicability of this approach is deferred until Section V. The additional problems, especially in connection with ratios such as that displayed above, that are encountered in *measuring* the various kinds of congestion form a separate topic and are not discussed here.

The preceding paragraphs make it clear that this paper does not encompass traffic that varies in time (except possibly where we need only finite-time-average values of congestion), although such traffic can be quite simply described. By resorting to ensemble averages — that is, to the conceptual experiment of running a traffic system over and over again from time 0 to time t with statistically identical inputs — it is quite possible to define instantaneous values of congestion from various points of view, although it is difficult to motivate a distinction between

CC and TC. However, such generality would again serve only as a distraction from the main issue. I therefore adopt the assumption of stationarity: For present purposes this is a steady-state (equilibrium) theory.

The reader must see now that this introduction is really a sop to Cerberus. It records the framework of ideas in which the following remarks are couched, but very informally and on the assumption that the reader is thoroughly familiar with the concepts and terminology of traffic theory. I hope the loss of precision inherent in such a heuristic approach is outweighed by the gain in simplicity.

II. CONGESTION IN THE SIMPLEST SITUATION

Let us first consider a single full-access trunk-group $\mathcal{G}$. Its relevant properties are represented in Fig. 1. It consists of a number of channels. Subscribers at one end communicate with those at the other over a shared communication system which for our purposes consists of $\mathcal{G}$ alone. (We do not discuss communication between two subscribers at one end of $\mathcal{G}$.) Two particular subscribers are labeled "A" and "B", and we define the *pair* P as the pair (A,B). Traditional notions of blocking and congestion in this situation refer to the state of $\mathcal{G}$. Of course in defining CC from A's point of view we have to distinguish blocked states in which A is busy from those in which A is idle; but the fundamental concept is that of "all trunks of $\mathcal{G}$ busy". This is because the idea is simple and makes sense: Either some trunks are idle or they are all busy; and when all are busy, any subscriber either has a trunk or cannot get one.

Now let us represent a typical period of time T for $\mathcal{G}$ as the sum of periods x, y, u, and v, drawn as intervals in Fig. 2. The label "no path" for intervals x and y means that all trunks of $\mathcal{G}$ are busy, and "path"

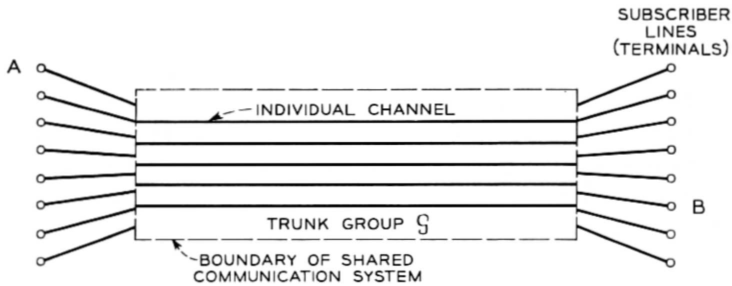

Fig. 1 — Full-access trunk-group.

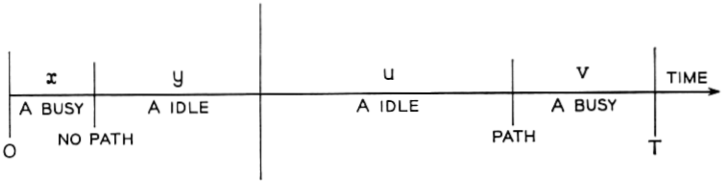

Fig. 2 — Division of time for $\mathcal{G}$.

means that $\mathcal{G}$ has at least one idle trunk. With these conventions, it is clear that the standard definition gives

$$\text{TC} = \frac{x + y}{x + y + u + v} \qquad \text{(Fig. 2)}.$$

(In each such expression, the number in parentheses specifies the figure in which the symbols are defined.)

The intervals x and y in which there is "no path" are lumped together in the formula for TC, as are u and v, because the wire chief's point of view ignores the state of terminal A. "The wire chief's point of view" is a traditional phrase which expresses the possibility of viewing the possible paths between A and B with the interests of the shared communication system rather than of the terminal-pair in mind. This viewpoint reflects a concern for the network itself and its ability to establish connections between pairs of terminals, and is therefore symmetric with respect to A and B. In particular, congestion between A and B is naturally seen as the fraction of time during which no path composed of idle elements exists. The complementary point of view, emphasizing states of the source, is known as that of the "particular subscriber" (a phrase from which I shall often omit the first word). This concept of "point of view" is useful, but requires considerably more discussion in the next section.

Suppose that A's calling rate varies, and is on the average $r(A)$ times as large when there is no path through $\mathcal{G}$ as when there is a path. Of course $r(A)$ can represent either an instantaneous effect or one depending on A's history. (Note that $r(A)$ incorporates the effects of, but is by no means simply related to, a change in A's calling rate triggered by an unsuccessful attempt and persevering until a successful attempt. If $r(A)$ has such a cause, then probably $r(A) > 1$ with human callers; but an automatic calling system might easily be designed to have $r(A) < 1$ in order to reduce the load on the switching equipment caused by unsuccessful attempts.) If A can initiate a call only when idle, then the customary notion of call-congestion, the proportion of A's attempts that are unsuccessful, is that

$$\mathrm{CC} = \frac{r(A)y}{r(A)y + u} \qquad \text{(Fig. 2)}.$$

If subscribers are not all alike, this formula may apply only to A, as suggested by the traditional association of CC with the particular subscriber's point of view.

If indeed A attempts calls only when idle, it may be totally uninteresting to him to consider intervals of time during which he is busy. But A may still want to distinguish between, on the one hand, the fraction of his call-attempts which fail and, on the other hand, the fraction of that time in which he *can* initiate calls during which attempts must fail. Thus it is natural to define the *modified* (or *conditional*) time-congestion as

$$\mathrm{MTC} = \frac{y}{y + u} \qquad \text{(Fig. 2)}.$$

Fig. 2 shows clearly that MTC, like CC, measures congestion from the subscriber's point of view.

What relations hold among these three measures of congestion in this situation? First, it is obvious that CC and MTC agree if and only if $r(A) = 1$, and that CC < MTC if $r(A) < 1$ and *vice versa*. Second, suppose that a relation **R** holds between the modified and ordinary time-congestions, where **R** is either <, =, or >. Multiplying both sides of the formula MTC **R** TC by $(y + u)(x + y + u + v)$, we get $y(x + y + u + v)$ **R** $(x + y)(y + u)$. Subtraction of $(xy + y^2 + yu)$ from both sides yields the result that MTC **R** TC if and only if yv **R** xu. Whenever $uy \neq 0$, it is helpful to rewrite this condition as

$$\mathrm{MTC}\ \mathbf{R}\ \mathrm{TC} \quad \text{if and only if} \quad \frac{v}{u}\ \mathbf{R}\ \frac{x}{y}.$$

(In certain degenerate cases $uy = 0$. For example, if $\mathcal{G}$ contains c trunks and there are c terminals at each end, $y = 0$ because A cannot be idle when there is no path through $\mathcal{G}$.) In the finite-source model analyzed by Engset [Ref. 1, pp. 250–1], where $r(A) = 1$, the events $\{A \text{ busy}\}$ and $\{\text{no path}\}$ are positively correlated, so that $(v/u) < (x/y)$. Therefore MTC < TC, and this in turn implies the well-known fact that CC < TC. But notice that no inequality between MTC and TC inheres in the definitions; it is not inconceivable that in some useful model $\{A \text{ busy}\}$ and $\{\text{no path}\}$ could be negatively correlated events, which would make $(v/u) > (x/y)$ and so reverse the previous inequality.

We care most about the relation between call- and time-congestion. The previous paragraph covers this for the special case in which $r(A) = 1$. The procedure applied above, of cross-multiplication, cancellation, and division by uy, shows that in general

$$\text{CC } \mathbf{R} \text{ TC} \quad \text{if and only if} \quad r(A)\left(1+\frac{v}{u}\right) \mathbf{R} \left(1+\frac{x}{y}\right).$$

Unless both CC and TC agree with MTC, we see that they are not very likely to agree with each other. Also, the most natural assumptions about a traffic system in the absence of information to the contrary would be, first, that $r(A) \geqq 1$, and second, that $(v/u) \leqq (x/y)$ because of a positive correlation between the events {no path} and {A busy}. Unless put in quantitative form, these assumptions lead only to the conclusion that MTC is likely to be smaller than both time- and call-congestion.

For this situation we have defined three types of congestion. The two kinds of time-congestion would agree if the probability of blocking within the shared communication system were independent of the state of the relevant source. CC agrees with MTC when a calling rate is unaffected by the availability of desired paths. CC and TC agree if both agree with MTC, or otherwise when $r(A)[1 + (v/u)] = 1 + (x/y)$. Since these three measures of congestion can differ, all are of interest.

III. CONGESTION IN NETWORKS

A communication network $\mathfrak{N}$ is sketched in Fig. 3; this one happens to be a 3-stage switching network. Before defining anything, we note one fundamental difference from the simple example in Section II. It makes sense to emphasize the dichotomy described by "all trunks of $\mathcal{G}$ busy" and "$\mathcal{G}$ has some idle trunks". The isolated, full-access trunk-

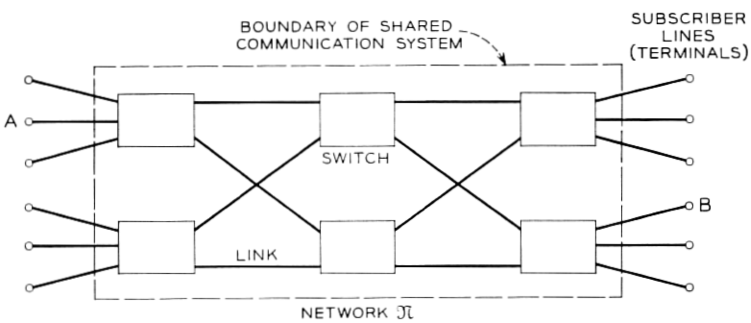

Fig. 3 — Connecting network.

group exemplifies the more general case of congestion in any *one* link, composed of interchangeable (fully accessible) channels, of a communication network. For present purposes we can completely describe such a link at a given instant by saying whether it does or does not contain at least one idle channel. But, except in special cases of little interest, no such concept as "$\mathfrak{N}$ is busy" can be defined in a useful way. Even the wire chief does not say, "My network is busy"; he says, "No paths exist between the following pairs of customers". In a general network, whether it be a switching network, a network of trunk groups, or a combination of both to form a complete communication system, the notion of blocking applies only to specific pairs $P = (A,B)$ of terminals. This is why so many investigators immediately fix their attention on the subgraph $g(P)$ of $\mathfrak{N}$ corresponding to the various routes over which A and B could be connected. Fig. 4 shows one conventional way of representing $g(P)$ for the situation of Fig. 3. The nodes are switches and the branches, links.

Let us take $\{P \text{ idle}\}$ to be the event that both A and B are idle, and $\{P \text{ busy}\}$ the event that either A or B or both are busy; that is, $\{P\,\text{busy}\} = \{P\,\text{not idle}\}$. Then one possible way of subdividing a typical period of time for $\mathfrak{N}$ is shown in Fig. 5. Here the conditions "path" and "no path" refer to the subgraph $g(P)$, so that the entire subdivision of the period T is of interest only to the terminal-pair P. (It will of course relate also to other pairs, if any, whose behavior is in every respect statistically identical with that of P.) Let us call each quantity that is most naturally defined by dividing time as in Fig. 5, *congestion as measured for (or by) the particular pair*. The nature and significance of this convention are considered below.

(One important feature of the representation of Fig. 4 is that no branch of $g(P)$ is uniquely associated with one subscriber. Thus, as seen by the pair P, $g(P)$ is the entire shared communication system, just as $\mathcal{G}$ was in the example of Section II; though of course the *behavior* of $g(P)$ is affected by traffic passing through the remainder of $\mathfrak{N}$. With this viewpoint $v \neq 0$ in Fig. 5; for it is quite possible for a path between A and B to exist in $g(P)$ when P is busy, because the terminal nodes of $g(P)$

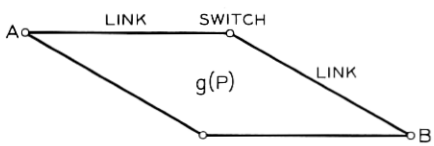

Fig. 4 — Subgraph $g(P)$ of $\mathfrak{N}$.

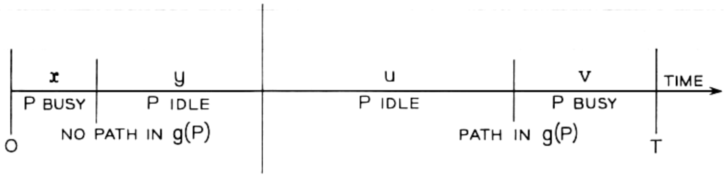

Fig. 5 — Possible division of time for $\mathfrak{N}$.

are the switches to which A and B have access. I have gathered in conversation that certain traffic theorists have a mental picture of $\mathfrak{N}$ like the one in Fig. 6, in which $g(P)$ includes a branch for each subscriber. The only effect of this change, when it exists, is to make the events $\{P \text{ busy}\}$ and $\{\text{path}\}$ incompatible. In this case $v = 0$ in Fig. 5. The distinction between the conventions of Figs. 4 and 6 is conceptually quite important, because Fig. 6 by itself can suggest the *omission* of

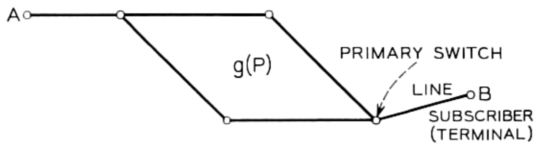

Fig. 6 — Extended subgraph.

v from all our formulae. But the distinction does not affect the character of the results discussed below, and so we assume the model of Fig. 4 without further comment.)

Returning to Fig. 5, it seems natural to define the time-congestion as

$$\text{TC}_2 = \frac{x + y}{x + y + u + v} \qquad (\text{Fig. } 5),$$

where the subscript "2" refers to measurement for the *pair* P. This definition is clearly analogous to that of TC in Section II: For TC_2, being a function of $x + y$ and of $u + v$, is based only on what P sees as its *shared* communication system; the states of the terminals are not taken into account. On the other hand, the intervals $x + y$ and $u + v$ relate to $g(P)$ alone rather than to all of $\mathfrak{N}$. For these reasons we may call TC_2 "the wire chief's definition for the pair P", indicating that the *point of view* is the wire chief's while the *measurement* is for (or by) the particular pair.

Fig. 5 leads to natural definitions of MTC and CC (again as measured for the pair), but only if we think of calls as being attempted between A

and B only when P is idle. The usefulness of this convention, which is discussed below, depends on the physical arrangements of the network. It is hard to think of "attempts" in any *other* way in a connecting network with gas-tube crosspoints, in which one tries to set up a connection between two points by establishing a potential-difference between them; if idle paths exist, the gas tubes along one of them will break down and so connect the two points. Some forms of common control tend to lead toward similar ideas; for example, it may be useful to think of a marker as "attempting" to find a route between two devices only if both devices are idle. Here we define the modified time-congestion as

$$\mathrm{MTC}_2 = \frac{y}{y + u} \qquad \text{(Fig. 5)},$$

and the call-congestion as

$$\mathrm{CC}_2 = \frac{r(P)y}{r(P)y + u} \qquad \text{(Fig. 5)}.$$

As in Section II we may distinguish between TC_2 and these quantities by saying that the latter reflect the viewpoint of a pair of subscribers rather than of the wire chief, and thus by calling them "the subscribers' definitions for the pair P".

With congestion measured for the pair in three ways, what about the network as a whole? Surely we may want to know what fraction of all calls offered to the network are blocked, or what fraction of time finds the average pair unable, for lack of paths in $\mathfrak{N}$, to establish a connection. These questions are easily answered by averaging the previous quantities over all pairs P. The resulting measures of congestion could be described as "for the office, exchange, network, or system" or as the "office average", etc. Because "system" seems too broad a term, and in order not to emphasize switching applications as opposed to trunking, I choose the term "network" to describe congestion averaged over all pairs. This operation yields the quantities

$$\overline{\mathrm{TC}_2} = \text{avg over } P \text{ of } \mathrm{TC}_2\,;$$
$$\overline{\mathrm{MTC}_2} = \text{avg over } P \text{ of } \mathrm{MTC}_2\,;$$

and

$$\overline{\mathrm{CC}^2} = \text{weighted avg over } P \text{ of } \mathrm{CC}_2\,,$$

where the weighting in calculating the network-average call-congestion is proportional to the average calling-rate for each P. (Such weights may be very hard to find when $r(P) \neq 1$.) Notice that, with homogeneity of subscribers, averaging has no effect: For example $CC_2 = \overline{CC_2}$ if all subscribers are alike.

In some networks it is possible for A to try to call B whenever A is idle, regardless of B's state. Step-by-step switching is an obvious example, as is the Bell System voice network including all its subscriber sets. Even with common control, it is possible for a request for connection to initiate search for an idle path even if detection of a busy destination must follow establishment of a path through $g(P)$. In fact there are systems in which it is useful to imagine busy terminals as *initiating* call-attempts. In an electronic central office, for example, a path is sometimes reserved for a connection which is to be established as soon as P becomes idle. In such a system an attempt can occur in the presence of any combination of states of the terminals of a pair, and each such combination may have a different associated calling-rate. The present treatment is restricted to less formidable models. Nevertheless it is important to realize that the situations in which only idle pairs make attempts and in which any idle terminal can make attempts, are only two of a large class of situations, many of the more complicated members of which are of engineering interest.

Calls attempted by an idle source can be successful, or can fail because there is no path in $g(P)$ or because B is busy. In situations for which it is important to distinguish between these sources of failure, the model must surely allow for calls to busy terminals. We arrange this by drawing Fig. 7. In a sense it subdivides time for A with respect to B, since the event {path in $g(P)$} relates to a particular B. Fig. 7 suggests the definitions

$$TC_{(1)} = \frac{x + y}{x + y + u + v} \qquad \text{(Fig. 7)},$$

$$MTC_{(1)} = \frac{y}{y + u} \qquad \text{(Fig. 7)},$$

and

$$CC_{(1)} = \frac{r(A)y}{r(A)y + u} \qquad \text{(Fig. 7)},$$

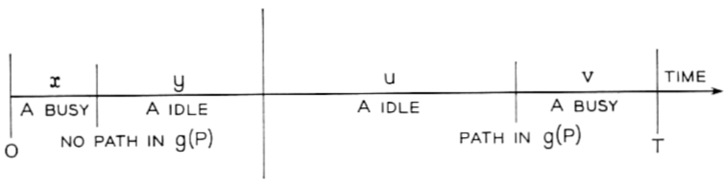

Fig. 7 — Time subdivided for A w.r.t. B.

which may be described as measures of congestion for one particular terminal with respect to another (hence the subscript "(1)"). These quantities agree as closely as possible with those (lacking subscripts) of Section II. Notice also that $TC_{(1)} \equiv TC_2$, a correspondence which does not hold for MTC or CC.

The quantities just defined measure *A*'s difficulties in trying to call *B*. We can measure *A*'s difficulties in making *any* calls by averaging over all *B*. We obtain measures of congestion for the particular terminal:

$$\begin{aligned} TC_1 &= \text{avg over } B \text{ of } TC_{(1)} ; \\ MTC_1 &= \text{avg over } B \text{ of } MTC_{(1)} ; \end{aligned}$$

and

$$CC_1 = \text{weighted avg over } B \text{ of } CC_{(1)} ,$$

where these weights are proportional to the frequencies with which *A* calls the various *B* to whom he can be connected.

One more averaging process takes us to congestion as measured for the network, for the case in which calls to busy destinations are possible. We define

$$\begin{aligned} \overline{TC_1} &= \text{avg over } A \text{ of } TC_1 ; \\ \overline{MTC_1} &= \text{avg over } A \text{ of } MTC_1 ; \end{aligned}$$

and

$$\overline{CC_1} = \text{weighted avg over } A \text{ of } CC_1 ,$$

with weights according to the relative frequencies with which the different *A* place calls. Here, as before, the measurements before and after averaging coalesce when terminals do not differ from each other. In any case, because $TC_{(1)} \equiv TC_2$, we also know that $\overline{TC_1} \equiv \overline{TC_2}$.

For a general network, measures of congestion have been defined for the pair and the network when attempts occur only between idle terminals, and for the terminal and the network when any idle source can attempt calls. In the latter situation, we also have measures of congestion for the particular terminal with respect to one destination. These are always useful because they lead to the particular terminal's measurements; and they have intrinsic interest in those cases, such as the control of electronic switching systems, in which it may be important to know the congestion encountered by messages from one source to each of several destinations. As for differences among the three quantities TC, MTC, and CC, the discussion in Section II on ordering relations carries over, *mutatis mutandis*, to the cases subscripted "(1)" and "2". (Notice, however, what care must be taken when dividing by uy in repeating the calculations involving **R**. The generalization of the cautionary example given above is that $y = 0$ in Fig. 5 whenever the network $\mathfrak{N}$ is non-blocking.) Similar arguments can be applied with caution to the measures obtained by averaging. It is particularly important that no simple formula for $\overline{CC_2}$ can be written using an "$\bar{r}(P)$" found by taking a P-average of $r(P)$; and so on.

Here we pause to summarize the measures of congestion defined above. This is best done as in Table I, which shows how the proposed symbols and terminology are related. A more accurate description of the mode of operation associated with the number 1 in a subscript would be that attempts are made *without regard* to the state of the destination; the heading in the table is shorter and suggests the distinction that is often most important. Classification by "point of view" is omitted entirely: For as used here the phrase is merely dichotomous, the wire chief being associated with TC and the particular subscriber with MTC and CC, so that the names "time-congestion" etc. are adequate by themselves. "Measurement for (or by)" is a new classification which supplies words to go with the subscripts and bars. The word "terminal" is used in column headings instead of "subscriber" because the latter would be too reminiscent of classification by point of view.

We have not yet considered the important case in which a single source A tries to call a "destination" consisting of several terminals, such as a group of outgoing trunks. Quite different paths through the network may lead from A to various equally useful members of the "destination". The discussion of Jacobaeus's work[2] in Section IV covers congestion in this situation.

Before concluding that each of our apparent plethora of definitions is necessary, we must ask this question: Is there really a non-trivial difference between quantities defined for the case in which attempts occur only when P is idle, as in Fig. 5, and those for the case of Fig. 7 in which a busy terminal can be called? We answer this question in terms

TABLE I — MEASURES OF CONGESTION

Mode of Operation:		Attempts when P Idle		Attempts Possible to Busy Terminal		
As measured for:		Pair	Network	Terminal w.r.t. One Destination	Terminal	Network
Type of congestion:	Time:	TC_2	$\overline{TC_2}$	$TC_{(1)}$	TC_1	$\overline{TC_1}$
	Modified time:	MTC_2	$\overline{MTC_2}$	$MTC_{(1)}$	MTC_1	$\overline{MTC_1}$
	Call:	CC_2	$\overline{CC_2}$	$CC_{(1)}$	CC_1	$\overline{CC_1}$

of Fig. 8, which combines Figs. 5 and 7. The intervals x, y, u, and v are as in Fig. 5, but the time when P is busy is labeled according to whether A is busy or idle, and x and v are subscripted correspondingly. The most restricted definition of TC — that with subscript "2" or "(1)" — depends only on {path} or {no path} in $g(P)$, and is the same for both modes of operation. But

$$\mathrm{MTC}_2 = \frac{y}{y+u} \qquad \text{(Fig. 8)}$$

and

$$\mathrm{MTC}_{(1)} = \frac{x_i + y}{x_i + y + u + v_i} \qquad \text{(Fig. 8)}.$$

Cross-multiplication and cancellation as before show that

$$\mathrm{MTC}_2 \ \mathbf{R}\ \mathrm{MTC}_{(1)} \quad \text{if and only if} \quad \left(1 + \frac{v_i}{u}\right) \mathbf{R} \left(1 + \frac{x_i}{y}\right),$$

again assuming that $uy \neq 0$, and after adding 1 to both sides for later convenience. Although the state of B is unspecified when A is busy and P is busy, certainly B is busy when P is busy and A is idle as in the periods x_i and v_i. It follows that

$$\begin{aligned} \frac{u}{u+v_i} \quad \text{(Fig. 8)} &= \frac{\Pr\{P \text{ idle} \mid \text{path}\}}{\Pr\{A \text{ idle} \mid \text{path}\}} \\ &= \frac{\Pr\{B \text{ idle } \& \ A \text{ idle} \mid \text{path}\}}{\Pr\{A \text{ idle} \mid \text{path}\}} \\ &= \Pr\{B \text{ idle} \mid A \text{ idle } \& \text{ path}\}, \end{aligned}$$

and likewise

$$\frac{y}{x_i + y} \quad \text{(Fig. 8)} = \Pr\{B \text{ idle} \mid A \text{ idle } \& \text{ no path}\}.$$

These quantities are the reciprocals of those appearing in the previous

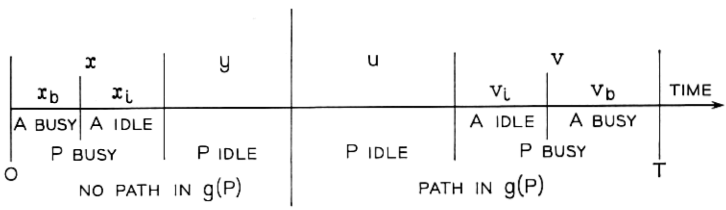

Fig. 8 — Further subdivision of time for $\mathfrak{N}$.

relation. Subtracting both of them from 1, we find that MTC_2 **R** $\text{MTC}_{(1)}$ if and only if

$$\Pr\{B \text{ busy} | A \text{ idle \& path}\} \;\mathbf{R}\; \Pr\{B \text{ busy} | A \text{ idle \& no path}\}.$$

In some models, $\Pr\{B \text{ busy}\}$ is not affected by the existence or non-existence of a path from A when A is idle. But in most systems in which the terminal B contributes a positive fraction of the traffic carried in $g(P)$, $\Pr\{B \text{ busy}\}$ *is* so affected, and MTC_2 must differ from $\text{MTC}_{(1)}$.

A similar argument applies to call-congestion. Because

$$\text{CC}_2 = \frac{r(P)y}{r(P)y + u} \qquad \text{(Fig. 8)}$$

and

$$\text{CC}_{(1)} = \frac{r(A)(x_i + y)}{r(A)(x_i + y) + u + v_i} \qquad \text{(Fig. 8)},$$

we can show that CC_2 **R** $\text{CC}_{(1)}$ when

$$\frac{1 + (v_i/u)}{1 + (x_i/y)} \;\mathbf{R}\; \frac{r(A)}{r(P)},$$

assuming $r(P)uy \neq 0$. Equality in this relation requires either an unlikely coincidence or agreement of both the MTCs and the calling-rate ratios. Only these last need further comment. Whenever it is reasonable to think of the calling rate of P as being composed of calls attempted by A and by B, the ratios $r(A)$ and $r(P)$ are equal if the calling rates of A and B (to each other) respond in the same way to a "no path" condition in $g(P)$. This they may or may not do, but anyway this question is independent of whether calls attempted by P are made up equally of calls from A and from B.

Having observed a genuine difference between the "pair-attempt" and "source-attempt" models, we return to Fig. 8 to tie off one more loose end. The definition of $r(P)$ seems natural; but suppose that, when calls to busy terminals are possible, changes in A's calling rate occur in response to the failure of attempts, with no distinction between busy-terminal and blocked-path failures. It is even possible to imagine average relative calling-rates $r_p(A)$ during {no path} and $r_B(A)$ during {path & B busy}, although relating these ratios to A's rules of operation might be extremely difficult. Such a model would generalize the definition of call-congestion to

$$\text{CC}_{(1g)} = \frac{r_p(A)(x_i + y)}{r_p(A)(x_i + y) + u + r_B(A)v_i} \qquad \text{(Fig. 8)},$$

which still measures the frequency of blocking due only to {no path}. Here I drop this line of thought, which can be extended if necessary in a particular application. (The possibility of defining CC so as to illuminate particular matters of interest is discussed by S. P. Lloyd in an unpublished Bell Laboratories memorandum. He treats especially the problem of correcting the apparently high CC caused by A's "extreme impatience" in retrying very rapidly after an attempt fails.) The main point is that the theorist may want to describe A's behavior as conditional upon the state of the system, whereas A actually responds to the results of his call-attempts or to other *sampled* data on the system's states.

IV. QUANTITIES CALCULATED IN THE LITERATURE

The purpose of this section is to compare the foregoing ideas with discussions in several papers on connecting networks. We begin with the monumental 1950 paper by Jacobaeus.[2] Syski, in his book,[1] achieves wonders of condensation in his excellent summary of this paper, mostly in Section 1 of Chapter 8, "Link Systems". Although Syski's book is available to most readers, I think it will be best to quote certain passages in full from Jacobaeus. The essential remarks are in Chapter 4, "Principles of Congestion Calculation in a Link System", and Chapter 9, "Formulae for Call Congestion in a Link System", of Ref. 2. They correspond to Sections 1.1.2, 1.1.3, 1.2.1, and 1.7.2 of Syski's Chapter 8.

The portion of a system analyzed by Jacobaeus consists of two switching stages and the links joining them. Fig. 9 illustrates his nomenclature, which I use in order to facilitate comparison with his paper; American terms appear in parentheses. An "A-device" is an inlet to a primary switch (strictly, a switch in the first stage considered), and an "A-group" is the set of inlets to one switch. A "B-device" is a primary-secondary link, and a "B-column" is the set of links emanating from one primary switch. Thus a B-device connects a primary outlet to a secondary inlet. A "C-device" is an outlet from a secondary switch, and a "route" is generally a "C-column", which is the set of corresponding outlets of the secondary switches. For example the second C-column is the set of second outlets of secondary switches. A route may consist of a part of a C-column, or of more than one.

Jacobaeus calculates, under various conditions, the probability that an A-device has no access to a particular route. He does this by writing first the probability that p of the C-devices of the route are busy; then multiplying by the probability that the $m - p$ B-devices (links) leading to the idle C-devices are also busy; and finally summing over p. This

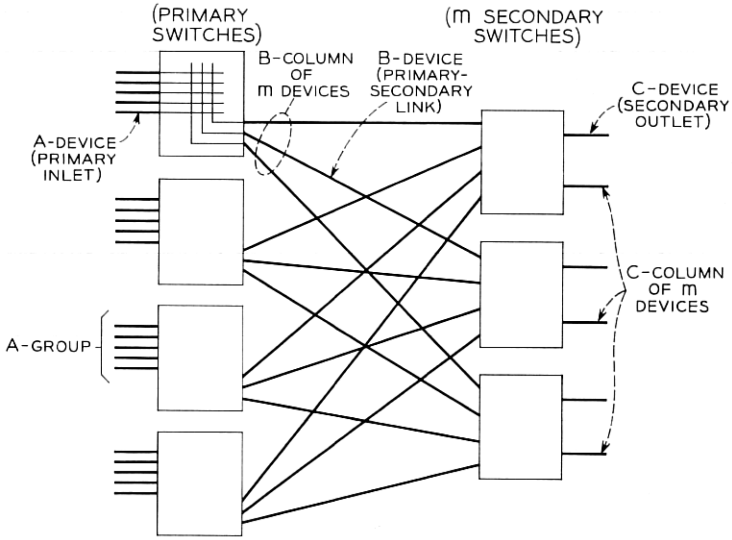

Fig. 9 — Nomenclature of Jacobaeus.

procedure yields his basic congestion-equation [Jacobaeus, p. 11, equation (1); Syski, p. 437, equation (2.1)]. The first ingredient of the product just described is a probability distribution for the number of busy C-devices in a route. The second ingredient is a distribution for the number of busy links in a specified part of a B-column. Three topics account for much of the length of Jacobaeus's paper. One is the correct choice of these two distributions in each situation studied. (It is easier to understand Jacobaeus's discussion of this topic with the aid of Syski's remarks [p. 439] on the meanings of the various traffic parameters used.) The second is the inaccuracy caused by assuming independence of the instantaneous link- and route-loads, as represented by multiplying the two blocking-probabilities. The third is the inaccuracy caused by approximating some of the difficult summations arising from the basic equation.

It is clear that this basic equation, as used in Chapters 5–8, 10, and 11 of Jacobaeus, yields $TC_{(1)}$, where this "congestion" is contributed by the portion of a system included in this model. We see this for the following reasons: First, the basic equation ignores the state of the calling A-device. Second, the goal of a call-attempt is not one destination but a route, consisting of several equally useful C-devices, and usually considered to be loaded with traffic from many sources according to Erlang's B-formula (first loss-formula). Third, calls are placed without prior regard to the state of the desired route. Fourth, the equation holds for a single A-group calling a single route. Thus $TC_{(1)}$ is measured for the particular terminals which share one primary switch, with respect to the class of destinations attainable via one route (which may be a trunk group). This measurement is as particular as possible, given that almost any switching system has *some* symmetries among its terminals.

Application of the basic equation in such a way as to account for finite-source effects appears to yield $CC_{(1)}$. However, since the case in which $r(A) \neq 1$ is never mentioned, this quantity cannot here be distinguished from $MTC_{(1)}$.

These views are partially confirmed by Jacobaeus on p. 23 of his Chapter 9:

> "The congestion formulae derived in the preceding chapters have all referred to the usual congestion concept, time congestion. The quantity used to express the congestion is the fraction of time during which the loading in the switching system is such that a new call cannot be switched. It is also possible to consider the call congestion, the fraction of the total number of calls which are blocked owing to a shortage of means of connection. In a full availability group loaded with random traffic the call congestion is equal to the time congestion, because the probability of a new call is always the same independent of the conditions ruling in the group. The condition for this is that there should be an infinite number of traffic sources, which must be the case if the traffic is to be truly random.
>
> "In the majority of the connection systems treated above, the number of traffic sources in each A-group has been limited. ... It may therefore be expected that the call congestion will have a smaller value than the time congestion."

There is further discussion [pp. 24, 38] of the finite size of an A-group, mostly with respect to the question of independence of sources. Jacobaeus points out that A-devices are truly independent only when they are "direct traffic-producers: subscribers' lines. . . . Otherwise the A-devices are secondary traffic sources, that is they derive their traffic from a large number of traffic sources by way of a concentrating device." We see that the "particular terminal" for which measurements are made may in fact be a "source" within the shared communication system, depending on what portion of the system is selected for study by means of this two-stage model. (Actually Jacobaeus discusses extensions of his method to more stages.)

We note in passing that some of the distributions substituted into the basic equation apply only when blocked calls are held (i.e., remain in the system for a length of time which does not depend on when or whether the desired path becomes available). Jacobaeus discusses this point [pp. 35, 38]. Relevant formulae appear in his Chapter 7; see also equations (2.15–17) of Syski [p. 442]. This situation is represented by the possibility, because of the presence of concentrating primary switches, that an A-device can be busy even when all B-devices for the same A-group are occupied by other sources. This cannot happen in the model of C. Y. Lee, to which we turn next, since he writes [Ref. 3, p. 1300, footnote] that "we restrict our attention to networks consisting of switches of non-blocking type".

The ideas of Jacobaeus form the closest possible analogue in the general case to the simple approach of Section II. In Lee's paper we expect to find measurements for the pair rather than for the particular terminal, since he writes in his Introduction [Ref. 3, p. 1288], ". . . methods for calculating . . . in a gas tube network are . . . given . . ." (see above, p. 2280). Our expectation is confirmed; for, although the links of Lee's subgraphs $g(P)$ are directed, his methods of calculation are symmetric and do not distinguish between sources and destinations. (See especially his Section 4.3, in particular the subsection on end-matching [p. 1311] in which a "voltage is applied to both ends of the network simultaneously (across a single input and a single output)".) I believe that Lee's use of directed links serves merely to specify, out of the geometrically possible ones, the allowed paths for a call in $g(P)$. But Lee does not discuss the pair-vs-terminal issue. In applying his model to line-switched traffic in a trunking network, one would naturally calculate congestion for one terminal, possibly with respect to another. This point is not important, for, as we see below, it cannot really be decided.

The essential attribute of Lee's model is that it ignores the states of terminals. He shows a four-stage switching network together with a subgraph $g(P)$ [Ref. 3, Figs. 3.2–3]; each possible path in the latter consists of just three branches. Such a model corresponds to our Fig. 4 rather than Fig. 6. The quantity found directly by Lee is "$P(i,j)$ the probability of all paths from input i to output j busy" [p. 1300]. This quantity corresponds to our TC_2. The question of pair or terminal discussed above is moot because, as we recall from p. 2282 above, TC_2 agrees with $TC_{(1)}$; their measurements diverge only for MTC and CC.

Lee introduces the important idea of blocking-probability for a *network* as an average of such probabilities over all terminal-pairs [p. 1300, equation (3.1)]. This, the quantity of major interest to Lee in parts of his paper, is of course our $\overline{TC_2}$. Furthermore, Lee sometimes assumes complete interchangeability of terminals [Section 3, p. 1300]: "In this section, we assume $P(i,j)$ to be independent of i and j . . . ." In this case, $TC_2 = \overline{TC_2}$ for every terminal-pair.

Lee's methods apply directly only to the (approximate) calculation of TC. His results are therefore consonant with those of Jacobaeus on TC, although these results are approached by the one on behalf of the pair and by the other on behalf of the source. It also seems clear from Lee's Section 4, in which he applies his methods to problems of retrials and connection time and discusses the results, that his formulae are meant to be taken as yielding CC_2. I think this intention constitutes an implicit assumption of what Jacobaeus (as quoted above) called "truly random traffic"; that is, that $r(P) = 1$ and that there are "infinitely many" sources.

The paper by Grantges and Sinowitz[4] is important because it extends the methods of the earlier literature and launches a practical attack on the problems of independence, network size, and computational complexity from which all researches in this field have suffered. They say [Ref. 4, p. 969] that "the nodes of the graph represent network switches and the directed branches of the graph represent network links". Their Figs. 1 and 2(a) [p. 970] show that, as in my Fig. 4, no branch of a subgraph is associated with any single terminal. Their "program is based on a simplified mathematical model of switching networks developed by C. Y. Lee" [Ref. 4, p. 967]. They assume [p. 969] stationary traffic, non-blocking switches, and complete equivalence among terminals. As discussed above, this last assumption makes network averages agree with measurements for one pair or terminal. I write here of the quantity that is calculated *directly*, and therefore ignore the symmetry that also yields measures of congestion for the whole network.

The previous paragraph shows that the mental picture of Grantges and Sinowitz is compatible with that embodied in this paper. Let us now examine the first paragraph beginning on p. 976 of Ref. 4. The predecessor of that paragraph discusses congestion in the simple context of a full-access trunk-group. The paragraph in question describes the attitude taken by the authors toward the problem of more general networks. They say that " . . . the measure of most concern to the network designer is call congestion. Now if it is assumed that calls originate completely independent of the state of the network, time congestion will equal call congestion. Such an assumption is unjustified if calls cannot originate from busy lines, since time congestion conventionally includes busy line periods while call congestion excludes them." The last clause means "excludes from both numerator and denominator of the ratio", as is clear from all our formulae for TC and CC. The second sentence just quoted must mean that TC and CC would have the same formula; and indeed our expression naturally defining CC would be the same as that for the corresponding TC if call-attempts occurred (as they do in Erlang's infinite-source model) without regard to the state of the network. This possibility is ruled out in the third sentence, which simply observes that the formulae for the two quantities differ. Of course, as our Section III shows, it is *because* calls often cannot originate from busy lines that we conventionally exclude busy-line periods in defining CC. Grantges and Sinowitz continue, "It is, however, reasonable to assume that idle pairs of terminals originate calls at a constant rate independent of the state of the network." (This sentence restricts us to measurements for the particular pair: that is, to TC_2, MTC_2, and CC_2, with the proviso that the network averages are included by symmetry.) "In particular, there must be no change in the calling rate after a blocked call." (In other words $r(P) = 1$.) "If, under this assumption, time congestion is modified to include only periods in which both lines are idle, it will be equal to call congestion." (Here they recognize that $MTC_2 \equiv CC_2$ when $r(P) = 1$.) "Actually, even if the foregoing assumption is not met, the time between calls is likely to be much longer than the time taken by the network to return to equilibrium, so that, again, the modified time congestion will be close to the call congestion." (That is, MTC_2 is near CC_2 when $r(P)$ is near unity.)

In the second paragraph beginning on p. 976, Grantges and Sinowitz go on to say, "With a suitable choice of branch occupancies, Lee's model allows the computation of call congestion. Alternatively, the branch occupancies may be chosen so as not to reflect the requirement that only idle terminals are to be considered, thus allowing the computation of time congestion." Here "suitable" means that branch occupancies are [p. 977, footnote] "chosen to reflect the requirement that the input-output terminals j,k are idle by (usually) subtracting the load contributed by the terminals j,k from the assumed carried link loads". This confirms the claim that the pair's measurement is calculated, since otherwise one would subtract the load contributed by one terminal alone.

It appears from all this that the NEASIM program of Grantges and Sinowitz was designed to find (approximately) either TC_2 or, after correction of branch occupancies, MTC_2, the latter quantity being supposed very close to CC_2 on the ground that $r(P)$ is very near 1. This situation agrees with that of Lee's paper, though the latter does not mention the correction procedure needed to find CC_2. Neither paper mentions the possibility of call-attempts to busy terminals, but the correction procedure is easily modified, as mentioned above, to cover that case. Since the output of the NEASIM program [Ref. 4, p. 975] is a functional relationship between sets of branch occupancies and probabilities that no path exists through $g(P)$, no decision is actually built into the program as to which measure of congestion is calculated. The user makes this choice separately by relating terminal loads to branch occupancies, a process which may have to be iterative and to rely on routing assumptions in order to ensure consistency.

V. DISCUSSION

Before summarizing the implications of this approach we must reconsider the question of applicability mentioned in Section I. It is not really profitable to try to characterize exactly (beyond the need for stationarity) models in which the mathematical limits required by the present method, with its use of conditional calling-rates and finite T, exist and agree in pairs. Instead we should look upon the arguments leading to Table I as exemplifying a point of view which is, if flexibly applied, relevant to a wide range of traffic systems.

Section IV shows that the thirteen independent entries of Table I — (recall that TC_2 and $\overline{TC_2}$ agree with $TC_{(1)}$ and $\overline{TC_1}$ respectively) — suffice, as they stand, for interpreting a variety of investigations in the literature. What can be said about this approach in cases of doubt, or when it is clear that a system requires new and more appropriate definitions? We need only to remember that such sketches as appear in Figs. 2, 5, 7, and 8 are intended merely to represent in a manageable way certain *conceptually* simple processes of measurement. For TC or for MTC such a process requires two clocks, one or both of them controlled by switches whose states reflect the defining states of the network under study. The measurement of CC requires instead two counters, which count respectively blocked attempts and all attempts of the desired categories.

For illustration consider a source A which, after every blocked attempt, makes retrials at a steadily increasing rate until the desired connection is set up. Application of one of our formulae for CC would require knowledge of $r(A)$, a quantity more difficult to evaluate in this case than call-congestion itself. But it is simple in principle just to count attempts and blocked attempts, which define CC directly. What is not so clear is whether this definition is useful. If periods during which a call-attempt of A would be blocked tend to be rare but long-lived, a very few unsuccessful first-attempts can lead to a high value of CC. When the cost incurred through failure to deliver a message increases rapidly with time, as does the retrial rate, this measure of congestion may be appropriate. Under other circumstances it may be preferable to count only first attempts and blocked first-attempts, or to count blocked retrials with a weighting factor that decreases with retrial number. The conceptual framework proposed here should not obscure the fact that measures of call-congestion can always be understood by reference to the idea of counting call-attempts; and similarly for time-congestion and the measurement of time intervals. The hard part in practice is to choose a definition whose behavior accurately reflects the usefulness of a system's performance.

Furthermore, conceptual simplicity does not guarantee utility. Direct measurement of MTC_2 requires a clock that knows when both A and B are idle *and* when there is no idle path in $g(P)$. Especially in a large network, this is likely to be impossible or impractical. One must often *estimate* congestion indirectly and in finite time, rather than measure it. This brings up such matters as sampling bias and efficiency, cost of instrumentation, and so on. Thus in choosing a definition of congestion one must consider not only its *purpose*, as discussed above, but also the cost of, and attainable accuracy in, *measuring* (or estimating) the quantity defined.

These interesting problems are not covered in this paper, which treats only the *concept* of congestion. I think we can attribute much of the confusion which has characterized this subject to the natural tendency of authors to treat highly symmetric models. Neither the row nor the column structure of Table I is significant when traffic is truly random and all terminals are alike. But the principal conclusion of this paper is that the distinctions embodied in Table I should be kept in mind. In other words it is always useful, if only as a precaution, to ask in what mode a system operates, for what entity congestion is to be measured, and what type of congestion is of interest.

A subsidiary conclusion is that there is no sharp distinction between trunking networks and switching networks, but rather a range of salient properties characteristic of various kinds of communication networks. Instead the basic division is between general networks, as treated in Section III, and single links of networks (links composed of one or more channels), for which the simpler discussion of Section II is adequate.

VI. ACKNOWLEDGMENTS

Although responsibility for the peculiarities of this discussion is mine alone, I am indebted to many of my colleagues at Bell Laboratories for their help in clarifying my ideas about congestion theory. Among them I am especially grateful to V. E. Beneš, A. Descloux, and H. O. Pollak for stimulating discussions, and to R. F. Grantges, W. S. Hayward, and John Riordan for helpful comments on earlier versions of this paper.

REFERENCES

1. Syski, R., *Introduction to Congestion Theory in Telephone Systems,* Oliver and Boyd, Edinburgh and London, 1960.
2. Jacobaeus, C., A Study on Congestion in Link Systems, Ericsson Tech., No. 48, 1950, pp. 1–68.
3. Lee, C. Y., Analysis of Switching Networks, B.S.T.J., *34,* November, 1955, pp. 1287–1315.
4. Grantges, R. F. and Sinowitz, N. R., NEASIM: A General-Purpose Computer Simulation Program for Load-Loss Analysis of Multistage Central Office Switching Networks, B.S.T.J., *43,* May, 1964, pp. 965–1004.