A "Thermodynamic" Theory of Traffic in Connecting Networks

By V. E. BENEŠ

(Manuscript received October 22, 1962)

Two new theoretical models for representing random traffic in connecting networks are presented. The first is called the "thermodynamic" model and is studied in detail. The second model is formulated in an effort to take methods of routing into account and to meet certain drawbacks of the "thermodynamic" model in describing customer behavior; since it is more realistic than the first, it leads to results that are vastly more complicated and must be described in another paper.

The "thermodynamic" model is worth considering for four reasons:

(1) It is faithful to the structure of real connecting systems. Indeed it is an improvement over many previous models in that it only considers physically accessible states of the connecting network, while the latter suffer the drawback that a large fraction of the network states on which calculation is based are physically meaningless, being unreachable under normal operation.

(2) It gives rise to a relatively simple theory in which explicit calculations are possible.

(3) The "thermodynamic" model provides a good simple description of traffic in the interior of a large communications network.

(4) It has an analogy to statistical mechanics which permits us to be guided by the latter theory as we try to use the model to understand the properties of large-scale connecting systems.

The two models to be described differ in only one respect. In the first (the "thermodynamic") model, the system moves from a state x to a state y that has one more call in progress, at a rate λ; the effective calling-rate per idle inlet-outlet pair is thus proportional to the number of paths usable in x from that inlet to that outlet. In the second model, the calling-rate per idle inlet-outlet pair is set at λ, and is then spread over the paths usable in x from that inlet to that outlet in accordance with some routing rule. This provides a mathematical description of routing, and avoids the unwelcome feature that a single customer's calling-rate depends on the state of the network.

The "thermodynamic" model is based on the single postulate that the "equilibrium" probabilities of the states of the connecting network maximize the entropy functional for a fixed value of the traffic carried. These probabilities have the same geometric or exponential form as the canonical Maxwell-Boltzmann distribution of statistical mechanics. The theory developed applies to any connecting network regardless of its structure or configuration. The number of calls in progress is analogous to the energy of a physical system. As in statistical mechanics, important averages can be expressed as logarithmic derivatives of a generating function analogous to the partition function of physics.

It is possible to give an interpretation of the maximum entropy postulate in terms of random behavior at the inlets and outlets of the connecting network; this interpretation leads to a stochastic progress z_t of the familiar Markov type, for which the canonical distribution is invariant. The transition rate matrix of z_t is self-adjoint in a suitable inner product space, so that the approach of z_t to equilibrium is easily studied, with resulting applications to traffic measurement.

I. INTRODUCTION

Like the physicist, the traffic engineer is faced with the study of an extremely complex system which is best described in statistical terms. The great success of the theoretical methods of statistical physics has given rise to a fervent hope, sometimes voiced among traffic theorists, that similar methods exist and can be found for the study of congestion. Indeed, the problems are much the same: one desires a small amount of "macroscopic" information about averages, based in a rational way on vast complexities of detail. A. K. Erlang was probably influenced by statistical mechanics when he introduced his method of "statistical equilibrium" into traffic theory. This method has had great success in dealing with problems of the birth-and-death type, like trunking and queueing, but as applied to more complex cases it has led mostly to algebraic and combinatory difficulties. Nothing as elegant or powerful as statistical mechanics has resulted so far.

We shall present two traffic models in this paper. The first is the outcome of a deliberate attempt to ape the methods of physicists in statistical mechanics, and thus to realize, at least in part, the hope mentioned above. It is called the "thermodynamic" model, and it is treated in detail. The second model is introduced later in the paper in an attempt to avoid certain drawbacks that appear in the interpretation of the "thermodynamic" model. Since it has independent interest and leads to involved, more realistic results, it is studied in detail in a later paper.

The approach taken in the "thermodynamic" model bears a close analogy to the methods of statistical mechanics, and is based on the single postulate that the "equilibrium" probabilities of the states of the connecting network maximize the entropy functional for a fixed value of the traffic carried. We develop a theory, briefly summarized in the next paragraph, by deducing the consequences and interpreting the meaning of the one basic assumption.

The state probabilities that maximize the entropy for a given carried load form a distribution function over the set of states that has the same geometric or exponential form as the *canonical* (or Maxwell-Boltzmann) distribution of statistical mechanics. The theory applies to any connecting network, regardless of its structure or configuration. The number of calls in progress is analogous to the energy of a physical system. As in statistical mechanics, important averages can be expressed as logarithmic derivatives of a generating function analogous to the partition function of physics. It is possible to give an interpretation of the maximum entropy postulate in terms of random behavior at the inlets and outlets of the connecting network. This interpretation leads to a stochastic process z_t of the familiar Markov type, and is such that any stochastic process based on it satisfies the maximum entropy postulate. The transition rate matrix A of z_t is self-adjoint in a suitable inner-product space; its characteristic values are real and non-positive, and can be studied by classical variational methods. In terms of these characteristic values the approach of z_t to equilibrium can be studied, with resulting applications to traffic measurements. It turns out that the covariance of any function of z_t is strictly positive. The paper ends with a time-dependent or non-stationary generalization of the maximum entropy postulate that has close analogies with the statistical "derivation" of thermodynamics.

II. PRELIMINARIES

In order to give an adequate summary and discussion of our theory in Section III, it is necessary to present first its concepts, terminology, and notation. Virtually all the notions about to be discussed have appeared in earlier papers by the author[1,2] so only a brief résumé is given here.

Let S be the set of possible (or permissible, or both) states of a connecting network, and let $x, y, \cdots$ be variables ranging over S. The elements of S are partially ordered by inclusion $\leqq$, where $x \leqq y$ means that x can be obtained from y by removing zero or more calls. Furthermore, the states $x \in S$ can be arranged in an intuitive manner in the *state-diagram*, the Hasse figure for the partial ordering $\leqq$. This figure is constructed by partitioning the states in rows according to the number of

calls in progress. The unique zero state (in which no calls are in progress) is placed at the bottom of the figure; above it comes the row of states with one call in progress; and so on. The figure is completed by drawing a graph with the states as nodes, and with lines between states (in adjacent rows) that differ in exactly one call. In an earlier work[2] we made the assumption that in a given state at most one call could be in progress between a given inlet and outlet; it is convenient to discard this assumption here.

If the connecting network under study is in a state x, it can move only to states which are *neighbors* of x, i.e., are obtainable from x by adding a new call or terminating a call in progress. It is useful to divide the neighbors of x into two sets A_x and B_x where

A_x = set of states immediately *above* x, i.e., accessible from x by adding a new call,

B_x = set of states immediately *below* x, i.e., accessible from x by a hangup.

For any set X, the notation $| X |$ is used to denote the number of elements of X. The states $x \in S$ can be defined[2] as sets of chains on a graph, one chain for each call in progress. Hence it is natural to use $| x |$ to mean the number of calls in progress in x. The kth level L_k is the set of all states with k calls in progress, i.e.,

$$L_k = \{x \in S: \quad | x | = k\}.$$

III. SUMMARY AND DISCUSSION

We start, in Section IV, with a brief informal discussion of what is meant, heuristically as well as precisely, by "equilibrium."

In Section V we formulate and discuss the *maximum entropy postulate*, according to which a suitable "equilibrium" distribution $\{q_x, x \in S\}$ of probability over S is obtained by choosing the probability vector q so as to maximize the entropy functional

$$H(q) = -\sum_{x \in S} q_x \log q_x$$

for a given value of the average number of calls in progress, i.e., for

$$\sum_{x \in S} | x | q_x = m.$$

Various heuristic arguments are adduced to support the *prima facie* reasonableness of this principle. In Section VI it is shown that the maximizing probability vector q is given by

$$q_x = \frac{\lambda^{|x|}}{\Phi(\lambda)}, \qquad x \in S$$

$$\Phi(\xi) = \sum_{x \in S} \xi^{|x|}$$

and λ is a constant determined uniquely by the equation

$$m = \lambda(d/d\lambda) \log \Phi(\lambda).$$

Because of their close similarity to corresponding notions from statistical mechanics, the vector q and the function $\Phi(\cdot)$ are henceforth called the *canonical distribution* and the *partition function*, respectively.

In Section VII we have collected together various properties of the partition function, most of them based on the partial ordering $\leqq$ of S. Among these are expressions for $\Phi(\cdot)$ in terms of the Möbius function for $\leqq$, and in terms of several sets of "characteristic polynomials" associated with $\leqq$ and S.

The canonical distribution q is placed in a dynamic context in Section VIII. This is done by defining a Markov process z_t (taking values on S) for which q forms a *stationary* distribution. The transition rate matrix A (infinitesimal generator) of this process allows one to give interpretations of this dynamic context in terms of calling rates and mean holding-times. An informal description of the process z_t is this: if it is in state x, it moves to a state $y \in A_x$ at a rate λ, and to a state $y \in B_x$ at a rate set at unity by convention.

A full discussion of the analogy between the "thermodynamic" theory of traffic and statistical mechanics is given in Section IX. For purposes of illustration, we mention that the number of calls in progress corresponds to the energy of a statistical mechanical system, and that the constant λ is related to the calling rate and corresponds to the temperature (up to a monotone transformation).

The reasonableness of z_t as a description of an operating connecting network is discussed and criticized in detail in Section X. Two possible interpretations of the inlets and outlets are considered: in one, the inlets and outlets are the ultimate terminals of the system, beyond which there is no more switching equipment; in the other, the inlets and outlets are switching centers such as PBX's, frames, or individual crossbar switches, acting as sources of traffic for a network under study. In the first interpretation, there can be at most one call in progress on an inlet or an outlet; in the second, there may be several.

Regardless of which interpretation of the inlets and outlets is adopted, the transition rate matrix A for z_t must be interpreted as saying that

the calling rate between an inlet and an outlet in a given state x is proportional to the number of free paths that x provides between that inlet and outlet. This assumption is unacceptable for the interpretation of inlets and outlets as ultimate terminals; it is not entirely unreasonable if the inlets and outlets are local switching centers.

Section XI is devoted to describing, as an alternative to z_t, a Markov stochastic process on S based on the assumption that the calling rate between an idle inlet terminal and an idle outlet terminal is a constant λ. This calling rate is then spread over the possible ways of realizing the call in question in the current state of the network in accordance with some method of routing. A mathematical description of such a method of choosing routes for calls is given. This description leads directly to a transition rate matrix Q for a process x_t in which every idle inlet-outlet terminal pair has a calling rate λ in every state. The possibility that z_t may be a useful perturbation of x_t is considered.

In Section XII it is observed that the rate matrix A for z_t is a self-adjoint operator in a suitable finite-dimensional inner product space. This implies that the characteristic values of A are real and nonpositive, and leads to bounds on the rate of approach of z_t to equilibrium. These bounds can be applied to estimate the covariance of z_t, and the sampling error incurred in measuring carried loads by averaging z_t, or discrete observations of z_t. In particular it is shown that the dominant (i.e., that of smallest nonzero magnitude) characteristic value r_1 of A satisfies

$$-(m/\sigma^2) \leqq r_1 < 0,$$

where m and σ are (respectively) the mean and standard deviation of the load associated with the equilibrium probability vector q for z_t, so that

$$m = \sum_{x \epsilon S} | x | q_x ,$$

$$\sigma^2 = \sum_{x \epsilon S} (| x | - m)^2 q_x .$$

In Section XIII we give a formula for the covariance of any process f_t defined by applying a function $f(\cdot)$ to z_t, i.e.,

$$f_t = f(z_t).$$

This covariance is *always* positive. Applications of this formula to traffic averages are described briefly in Section XIV. Finally, Section XV considers a time-dependent generalization of the variational principle on which the "thermodynamic" theory of traffic is based.

We conclude this section with an appraisal of the "thermodynamic"

theory presented herein. This will take the form of a list of comments, first *pro*, then *con*, and then a defense.

1. There exist theories[3,4] for connecting networks in which it is assumed that the links of the system are busy or idle with a given probability, all *independently* of one another. It can be verified that an overwhelming fraction of the states of the system so considered are in fact *not* physically meaningful states that the system can reach under normal operation. The theory presented here is based only on permitted, physically meaningful states, and so is not open to this serious objection.

2. The theory provides a *uniform* method of treating any connecting network in that the calculation of equilibrium probabilities *always* reduces to that of the partition function. In most other treatments the nature of the algebraic process of calculating probabilities depends heavily on a detailed account of the network configuration; in our theory it depends on the network only via the numbers $| L_1 |, | L_2 |, \cdots$.

3. The maximum entropy principle can be given a certain informal, *a priori* justification. It provides a "conservative, worst possible case" approach to problems and processes of fantastic complexity. This is because it can be interpreted as enjoining that an "equilibrium" distribution of probability for given carried traffic correspond to a condition of maximum ignorance of the actual state of the connecting network.

4. The canonical distribution q that results from the maximum entropy postulate can be embedded in a dynamic model of traffic by defining a Markov process z_t for which q is the invariant or stationary distribution. This dynamic model is described by a transition rate matrix which is a self-adjoint operator, a fact which makes it possible to study the time-dependent behavior of z_t in a simple approximate way, with applications to traffic measurement, for instance.

5. A very serious drawback of the "thermodynamic" theory is that its natural interpretation in terms of calling rates appears to be unreasonable in most practical cases. For this reason it will probably remain an amusing curiosity, rather than become an engineering tool.

6. The problem of calculating the partition function $\Phi(\cdot)$ is, as in statistical mechanics, very difficult except in cases of unrealistic simplicity. Thus, even if its assumptions are granted, our "thermodynamic" theory does not afford much progress in calculating quantities of interest.

7. The theory can take into account only one of the many different possible methods of routing calls in operating networks. Thus it cannot help the designer choose among alternative methods.

By way of defense against the objections just raised, these points can be made:

(*i*) Comment 5, that the interpretation of the "thermodynamic" theory in terms of calling rates is unreasonable, depends on a natural, but not necessarily valid or compelling, assignment of causes for new calls.

(*ii*) Although the calculation of $\Phi(\cdot)$ is hard, it is at least a definite combinatory problem solvable in principle by counting; thus at least part of the problem of obtaining state probabilities is disposed of.

(*iii*) It is doubtful whether routing methods make as much as an order of magnitude of difference in carried loads in large systems; hence it is reasonable to ignore them in a relatively crude theory such as the present one. (See however, Beneš, Ref. 11.)

The theory presented in this paper should be judged by its success in practice as well as by its agreement with our preconceptions. I believe that in spite of the major failings mentioned, the theory musters interest enough to warrant its presentation to traffic engineers, if only because its concepts and results may prove useful in more realistic approaches.

IV. EQUILIBRIUM

Quantities that are of interest in the design of a connecting network, such as the average load carried, the variance of the load, or even the probability of blocking, can often be calculated from a knowledge of some "equilibrium" or "stationary" state probabilities $\{q_x, x \in S\}$ for the network of interest. These probabilities are usually assumed or proven to be of "equilibrium" type in the sense that they have some physically reasonable invariance property.

Since the concepts of stationarity and equilibrium can assume many precise forms of varying strength, it is important to consider briefly some of these senses. The strongest notion, of course, is that of strict stationarity of a stochastic process, defined by the condition that all the finite-dimensional distributions be independent of time, i.e., be translation-invariant. A whole class of weaker notions can be obtained by requiring only that the distributions of dimension not greater than n be invariant. The notion of wide-sense stationarity, defined by the condition that the covariance depend only on the difference of its arguments, is still another concept of stationarity, formulated for a moment rather than a distribution. Again, Markov processes are described as homogeneous or stationary if their transition probability operators are time-invariant.

"Equilibrium" is a word that usually connotes a stable, quasi-static random behavior which is perhaps a condition of attraction for a process, in the sense that a process not in equilibrium tends toward it. Ergodic

Markov processes with denumerable state spaces are typical examples. It is to be remarked, though, that use of the word "equilibrium" usually implies a nondegenerate limiting behavior for a process y_t under study as $t \to \infty$. Thus a time-homogeneous Markov process may not have a genuine "equilibrium" distribution because it in some sense "blows up," e.g., the process may take values on the integers and the probability mass may move out toward $+\infty$, even though the transition probabilities are time-independent. In such a case, clearly, no first-order distribution can be assigned which is time-invariant.

The analytical expression of "equilibrium" often takes the form of a statement to the effect that an operator has zero as a characteristic value. Perhaps the most familiar example of such a statement arises in the case of a Markov process in continuous time with a transition rate matrix A; the equilibrium equation is $Aq = 0$, for a probability vector q.* This equation, together with its connections to semigroups, to Markov processes, and to the notion of statistical equilibrium used in traffic theory, is discussed immediately below.

A traditional analytical method in telephone traffic theory is that of "statistical equilibrium," due to Erlang.[5] This method may be described heuristically as follows: A notion of equilibrium is defined by the property that the rate of flow of probability into (or onto) a state equals that out of (or from) the state; this equilibrium is expressed in a set of equations among the state probabilities, the so-called statistical equilibrium equations; the "equilibrium" state probabilities are then taken to be (or defined by) the solution of these equations.

The method of statistical equilibrium can be interpreted in the mathematically rigorous context of semigroups of positive operators, here the matrices of transition probabilities $\{Q(t),\ t \text{ real}\}$ for a Markov process x_t taking values in S, with

$$Q(t) = (q_{xy}(t))$$

$$q_{xy}(t) = \Pr\{\text{state of system is } y \text{ at } t \text{ if it was } x \text{ at } 0\}.$$

The *generator* A of the semigroup is the matrix of transition *rates* or the derivative

$$A = \lim_{t \to 0} \frac{A(t) - I}{t} = Q'(0).$$

The matrix A expresses the relative probabilities of the various changes

* We are using the convention $(Aq)_x = \sum_y a_{yx} q_y$, rather than the more usual $(Aq)_x = \sum_y a_{xy} q_y$.

that can take place in a small amount of time, and indeed

$$Q(t) = I + At + o(t) \quad \text{as} \quad t \to 0.$$

In terms of the generator A the equation of statistical equilibrium can be written as $Aq = 0$, which expresses the fact that the vector q of state probabilities is an eigenvector of A corresponding to a zero eigenvalue of A. From the semigroup property

$$Q(t + s) = Q(t)Q(s)$$

it follows that

$$Q(t) = \exp At$$

$$\sum_y q_y q_{yx}(t) = q_x, \qquad x \in S, \qquad t \text{ real}$$

the last equation expressing the *invariance* of q under the transition probability matrices $Q(\cdot)$.

V. THE MAXIMUM ENTROPY PRINCIPLE

In the method of statistical equilibrium, the state probabilities are calculated *a posteriori* from a prior equation expressing an equilibrium or invariance principle. This equation is either postulated or is derived from assumptions that lead to a Markov stochastic process as a model for the operating network.

In the present work we use a variational principle rather than an equilibrium principle as a basis for calculating "equilibrium" state probabilities. In drawing this distinction we refer only to the immediate form of the assumptions and derivations, and imply no absolute distinction, since an "equilibrium" principle can almost always be given a "variational" form. For example, if A is a transition rate matrix for an ergodic Markov process, and A is self-adjoint with respect to an inner product $(\cdot, \cdot)$, then the "equilibrium" probability vector q, i.e., the solution of $Aq = 0$ is equally well described as the vector which maximizes the Rayleigh quotient

$$\frac{(Aq, q)}{(q, q)}.$$

It will turn out that the probabilities $\{q_x, x \in S\}$ derived from our variational principle also have an invariance property expressible, as in the example given, in terms of the self-adjoint generator A of a Markov semigroup by the equation $Aq = 0$. This equation can be interpreted

as a "statistical equilibrium" equation, and the elements of A related to calling-rates and hangup rates, in the various states $x \in S$.

However, instead of starting with a suitable matrix A to represent the infinitesimal dynamic behavior, and solving $Aq = 0$ in order to obtain an equilibrium distribution $\{q_x, x \in S\}$ over the states of the system, we shall directly choose a certain q, to be used as an "equilibrium" distribution for calculating quantities of interest, according to this criterion: The entropy functional

$$H(q) = -\sum_{x \in S} q_x \log q_x$$

is to be as large as possible subject to the conditions

$$q_x \geqq 0, \qquad x \in S$$

$$\sum_{x \in S} q_x = 1,$$

$$\sum_{x \in S} | x | q_x = m,$$

where m is a given number, the average load carried. The first two conditions ensure that only *bona fide* probability distributions are considered, while the third enjoins that q give rise to m as the mean number of calls in progress in equilibrium. This criterion or method for choosing a probability distribution over S we call the *maximum entropy principle*; it is exactly analogous to that used in statistical mechanics, provided that the number of calls in progress is interpreted as the energy of the mechanical system. We have already stated that this principle leads to a unique q which is exactly the same as would be obtained by a particular choice of A, given later, and solving $Aq = 0$; this matrix A has a definite interpretation in terms of system behavior during small periods of time.

A measure of justification for using the maximum entropy principle can be obtained from five arguments:

(1) Insofar as a high value of the entropy functional is an indication of a low degree of information, so far can use of the principle be interpreted as postulating that an equilibrium distribution $\{q_x, x \in S\}$ corresponds to a condition of maximum ignorance subject to a given average number of calls in progress. The principle may thus be said to represent a "safe" or "worst case possible" approach to the problem. Exactly the same principle is used in statistical mechanics to obtain the canonical distribution. In both cases it is a reasonable and systematic way of throwing up our hands.

(2) The principle is appealing for the obvious reasons of unity, uniformity, simplicity, and elegance.

(3) It leads to a theoretical structure similar to that of statistical mechanics. As in the physical theory, statistical quantities of interest are calculated from a *partition function*, characteristic of the network under study, that depends on purely combinatory properties of that network.

(4) The principle of maximum entropy leads to a unified theory applicable to all connecting networks. That is, the resulting "equilibrium" distribution depends algebraically on the structure of the network in a way that in a sense is uniform for all networks.

(5) The principle can be given a dynamic context in terms of Markov processes. This context permits the study of the approach to equilibrium in time, with important applications to sampling error.

VI. THE CANONICAL DISTRIBUTION

In the next few sections we develop some of the principal consequences of the maximum entropy principle, and examine their similarity to statistical thermodynamics. In the present section we determine the distribution $\{q_x , x \in S\}$ which maximizes $H(q)$ for a given average load carried. The following lemma is no doubt well-known, especially to physicists; since its application in traffic theory is relatively new, its detailed proof is included for completeness.

Lemma 1: Let $f(\cdot)$ be a nonnegative function defined on S, and let

$$\Psi(\xi) = \sum_{x \in S} \xi^{f(x)}.$$

The maximum of

$$H(q) = -\sum_{x \in S} q_x \log q_x$$

subject to the conditions

$$q_x \geqq 0, \qquad x \in S$$

$$\sum_{x \in S} q_x = 1 \tag{1}$$

$$\sum_{x \in S} q_x f(x) = m_f , \qquad (m_f \text{ a given positive number in the closed convex hull of the range of } f(\cdot),) \tag{2}$$

is

$$H(q) = \log \Psi(\omega) - m_f \log \omega,$$

where ω is the unique positive solution of

$$\omega(d/d\omega) \log \Psi(\omega) = m_f .$$

The maximum is achieved by the choice

$$\begin{aligned} q_x &= [\omega^{f(x)}/\Psi(\omega)], \\ &= \exp\{-a_1 f(x) - a_2 - 1\}, \qquad x \in S \end{aligned} \tag{3}$$

where a_1, a_2 *are (the values of Lagrange's multipliers) determined by any two of the relations*

$$a_2 = \log \Psi(e^{-a_1}) - 1,$$

$$m_f = \sum_{x \in S} f(x) \exp(-a_1 f(x) - a_2 - 1),$$

$$\omega = \exp\{-a_1\}.$$

Proof: With a_1 and a_2 as Lagrange's multipliers, we form the expression

$$h = -\sum_{x \in S} q_x \log q_x - a_1 \sum_{x \in S} f(x) q_x - a_2 \sum_{x \in S} q_x,$$

differentiate with respect to each q_x, and set the resulting derivatives equal to zero. This gives the equations

$$\partial h/\partial q_x = -(\log q_x + 1 + q_1 f(x) + a_2) = 0, \qquad x \in S \tag{4}$$

whose solution is (3). The multipliers a_1 and a_2 are to be determined from the conditions $\sum_{x \in S} q_x = 1$ and $\sum_{x \in S} f(x) q_x = m_f$. The first gives

$$1 = e^{-a_2 - 1} \sum_{x \in S} \exp[-a_1 f(x)],$$

$$a_2 = \log \Psi(e^{-a_1}) - 1,$$

while the second yields

$$\begin{aligned} m_f &= e^{-a_2 - 1} \sum_{x \in S} f(x) \exp[-a_1 f(x)] \\ &= \frac{\sum_{x \in S} f(x) \exp[-a_1 f(x)]}{\sum_{x \in S} \exp[-a_1 f(x)]}. \end{aligned}$$

Setting $\omega = \exp\{-a_1\}$, it is found that ω should be a solution of the equation

$$\omega(d/d\omega) \log \Psi(\omega) = m_f > 0. \tag{5}$$

From the fact that

$$\begin{aligned} \frac{d^2}{da^2} \log \Psi(e^{-a}) &= \frac{1}{\Psi(e^{-a})} \sum_{x \in S} \left(f(x) + \frac{d}{da} \log \Psi(e^{-a}) \right)^2 e^{-af(x)} \\ &> 0 \end{aligned}$$

it is easily shown (Khinchin,[6] p. 77) that there is exactly one solution ω of (5), and that ω is positive.

A relative extremum of $H(q)$ in $q \geqq 0$ subject to (1) and (2) must satisfy equations (4). Since these have only one solution there is only one such extremum. To show that it is a maximum it is enough to show that the matrix of second derivatives of $H(q)$ with respect to the components q_x of q is negative definite. However, this is straightforward, since

$$\frac{\partial^2 H}{\partial q_x \partial q_y} = \begin{cases} 0 & \text{if } x \neq y \\ -\dfrac{1}{q_x} & \text{if } x = y \end{cases}.$$

In Lemma 1 we let

$$f(x) = |\, x \,| = \text{number of calls in progress in state } x$$

and we obtain

Theorem 1: Let $m > 0$; let

$$\Phi(\xi) = \sum_{x \epsilon S} \xi^{|x|};$$

and let λ be the unique (positive) root of

$$m = \lambda(d/d\lambda) \log \Phi(\lambda).$$

The maximum of $H(q) = -\sum_{x \epsilon S} q_x \log q_x$, subject to the conditions that q be a probability vector over S and that $\sum_{x \epsilon S} |\, x \,| q_x = m$, is

$$H_{\max} = \log \Phi(\lambda) - m \log \lambda$$

and is achieved by the vector q with components

$$q_x = \frac{\lambda^{|x|}}{\Phi(x)}, \qquad x \epsilon S.$$

This is the distribution of probability over S that is determined uniquely by the maximum entropy principle; as noted before, it is the *canonical* distribution. The function $\Phi(\cdot)$ is called the *partition function* of the connecting network whose states form the set S. Since m determines λ uniquely and *vice versa*, we can use λ as the parameter that determines the average traffic level instead of m. Indeed, m is a monotone increasing the function of $\lambda \geqq 0$. Also it can be seen that moments of the distribution of the number of calls in progress (other than the mean) can be calculated from $\Phi(\cdot)$ by logarithmic differentiation.

VII. PROPERTIES OF THE PARTITION FUNCTION

In this section we exhibit various identities and relationships that are typical of the partition function $\Phi(\cdot)$. This function is the generating function of the number of states in a given level; that is

$$\Phi(\xi) = \sum_{k=0}^{w} \xi^k \mid L_k \mid, \qquad w = \max_{x \epsilon S} \mid x \mid.$$

Thus the problem of calculating λ, $\Phi(\cdot)$, and q in our model reduces to the calculation of the sequence

$$\mid L_0 \mid, \mid L_1 \mid, \cdots$$

and *vice versa*.

Remark 1:

$$\mid L_k \mid = \frac{1}{k} \sum_{y \epsilon L_{k-1}} \mid A_y \mid$$

$$\sum_{x \epsilon S} \mid A_x \mid = \sum_{x \epsilon S} \mid x \mid = \Phi'(1)$$

The first part of this result was proven as Theorem 1 in Ref. 2, and it implies the second part.

The Möbius function $\mu(\cdot)$ of the partially ordered system $(S, \leqq)$ is defined recursively by

$$\mu(0) = 1, \qquad \mu(x) = -\sum_{y<x} \mu(y) \quad \text{if} \quad x > 0, \qquad x \epsilon S.$$

We have remarked in previous work (Ref. 2, Section VII) that if S is a class of network states, then $\mu(\cdot)$ takes on the simple form

$$\mu(x) = (-1)^{\mid x \mid} \mid x \mid !.$$

We define the generating function $M(\cdot)$ by

$$M(\xi) = \sum_{x \epsilon S} \xi^{\mid x \mid} \mu(x).$$

Since

$$\Phi(\xi) = \sum_{x \epsilon S} \xi^{\mid x \mid} = \sum_{x \epsilon S} (-\xi)^{\mid x \mid} \frac{\mu(x)}{\mid x \mid !}$$

it can be seen that (except for a change of sign in the generating variable) $\Phi(\cdot)$ is the exponential generating function associated with $M(\cdot)$. Thus we have

Remark 2:

$$M(\xi) = \int_0^\infty e^{-u}\Phi(-\xi u)\,du$$

Proof:

$$\begin{aligned} M(\xi) &= \sum_{k=0}^{w} \xi^k \mid L_k \mid \cdot (-1)^k k! \\ &= \sum_{k=0}^{w} (-\xi)^k \mid L_k \mid \int_0^\infty e^{-u} u^k\,du \\ &= \int_0^\infty e^{-u}\Phi(-\xi u)\,du. \end{aligned}$$

In analogy with Birkhoff,[7] p. 15, (12), we define for each $x \in S$ a characteristic polynomial by the recursion formula*

$$p_x(\xi) = \xi^{|x|} - \sum_{y<x} p_y(\xi).$$

This is related to the Möbius function $\mu(\cdot)$ by the fact that if $\mu_y(\cdot)$ denotes the Möbius function for the set $\{x: x \geqq y\}$, then

$$p_x(\xi) = \sum_{y \leqq x} \xi^{|y|}\mu_y(x).$$

However, the partial ordering of the cone $\{x: x \geqq y\}$ is again of the same form as that of S; i.e., there are exactly $(\mid x - y \mid)!$ ascending chains between y and x, all of length $\mid x - y \mid$. Hence, by ref. 7, p. 15, (11),

$$\mu_y(x) = \frac{(-1)^{|x-y|}}{(\mid x - y \mid)!} = \mu(x - y)$$

and

$$\begin{aligned} p_x(\xi) &= \sum_{y \leqq x} \xi^{|y|}\mu(x - y) \\ &= \xi^{|x|} + \sum_{y<x} \xi^{|y|}\mu(x - y). \end{aligned}$$

Let now

$$\begin{aligned} q_x(\xi) &= \sum_{y<x} \xi^{|y|} \\ P_x(\xi) &= \sum_{y<x} p_x(\xi). \end{aligned}$$

* Actually, Birkhoff's polynomial $p_x[\xi]$ equals $\xi p_x(\xi)$. The definition we use is more convenient for our purposes.

The Möbius inversion formula gives

$$\xi^{|x|} = \sum_{y<x} \mu(x - y) q_y(\xi)$$

$$p_x(\xi) = \sum_{y<x} \mu(x - y) P_y(\xi).$$

To calculate $q_x(\cdot)$ explicitly, we note that if $0 \leqq k \leqq |x|$, then, using the cup $\cap$ for set intersection,

$$|L_k \cap \{y: y \leqq x\}| = \binom{|x|}{k},$$

i.e., there are exactly $\binom{|x|}{k}$ states with k calls up below any state x. Hence

$$q_x(\xi) = \sum_{y<x} \xi^{|y|} = (1 + \xi)^{|x|} - \xi^{|x|}.$$

Let us write

$$\xi^{|x|} = \sum_{y<x} r_y(\xi),$$

where $r_y(\cdot)$ are functions to be determined. Using the Möbius inversion formula once more, we find that one choice of the r_x's is

$$r_x(\xi) = \sum_{y<x} \mu(x - y) \xi^{|y|},$$

$$= p_x(\xi) - \xi^{|x|},$$

so that

$$\xi^{|x|} = \sum_{y<x} (p_y(\xi) - \xi^{|y|})$$

$$= P_x(\xi) - q_x(\xi)$$

and

$$P_x(\xi) = (1 + \xi)^{|x|}$$

$$= \sum_{k=0}^{|x|} \binom{|x|}{k} \xi^k$$

$$= \sum_{k=0}^{|x|} \xi^k \cdot \text{number of elements of } L_k \text{ less than or equal to } x.$$

It is apparent that

$$\Phi(1 + \xi) = \sum_{x \in S} (1 + \xi)^{|x|} = \sum_{x \in S} \sum_{k=0}^{|x|} \binom{|x|}{k} \xi^k.$$

Since for $0 \leqq k \leqq |x|$ there are precisely

$$\binom{|x|}{k}$$

elements in L_k that are below x, we have

$$\Phi(1 + \xi) = \sum_{x \epsilon S} P_x(\xi).$$

The preceding results yield the following identities for $\Phi(\cdot)$:

$$\begin{aligned}
\Phi(\lambda) &= \sum_{x \epsilon S} \lambda^{|x|} = \sum_{x \epsilon S} \sum_{y < x} \mu(x - y) q_y(\lambda) \\
&= \sum_{x \epsilon S} \sum_{y < x} r_y(\lambda) \\
&= \sum_{x \epsilon S} \sum_{y < x} \{p_y(\lambda) - \lambda^{|y|}\} \\
&= \sum_{x \epsilon S} P_x(\lambda) - \sum_{x \epsilon S} q_x(\lambda) \\
&= \Phi(1 + \lambda) - \sum_{x \epsilon S} \sum_{y < x} \lambda^{|y|}.
\end{aligned}$$

VIII. A REVERSIBLE MARKOV PROCESS FOR WHICH THE CANONICAL DISTRIBUTION IS INVARIANT

We shall describe an ergodic reversible Markov process z_t, taking values in the set S of states, and having the property that its stationary distribution over S is *precisely* the canonical distribution derived from the maximum entropy postulate. This Markov stochastic process can be used to place the canonical distribution into a dynamic context by exhibiting it as invariant under a semigroup of positive operators, viz., the transition matrices of the Markov process in question. The transition rate matrix A of this process, i.e., the generator of the semigroup, then provides several interpretations (cf. Section X) of this dynamic context in terms of behavior at the terminals of the networks, i.e., in terms of calling rates and mean holding-times.

Let $x \epsilon S$ be a possible state of the network. In Section II we have introduced the sets of states A_x and B_x with

A_x = set of states immediately *above* x, i.e., accessible from x by adding a new call,

B_x = set of states immediately *below* x, i.e., accessible from x by a hangup.

The process z_t to be considered can be described heuristically by saying that if $z_t = x$ then $z_{(\cdot)}$ is moving to each $y \epsilon A_x$ at a rate $\lambda > 0$.

to each $y \in B_x$ at a rate unity, and to any other state at a rate zero. Its transition rate matrix $A = (a_{xy})$ is given by

$$a_{xy} = \begin{cases} -\,|\,x\,| - \lambda\,|\,A_x\,| & \text{if} \quad y = x \\ 1 & \text{if} \quad y \in B_x \\ \lambda & \text{if} \quad y \in A_x \\ 0 & \text{if} \quad y \notin A_x \cup B_x \quad \text{and} \quad y \neq x. \end{cases}$$

With this matrix we can define a Markov process z_t in the usual way. (Cf. Doob.[8]) A discussion and critique of possible physical interpretations of the rate matrix A is given in Section X.

The probabilistic interpretation of the rate matrix A is that if $z_t = x$ there is a conditional probability $\lambda h + o(h)$ that $z_{t+h} = y$, for $y \in A_x$; there is a conditional probability $h + o(h)$ that $z_{t+h} = y$, for $y \in B_x$; there is a conditional probability $1 - \lambda h\,|\,A_x\,| - h\,|\,x\,| - o(h)$ that $z_{t+h} = x$; all other events have a conditional probability $o(h)$, as $h \to 0$. The constant λ is the calling rate per idle path.

An alternative informal description of the Markov process z_t is as follows: the length of time spent in any state x is a random variable independent of all other lengths of time spent in a state, having a negative exponential distribution with a mean

$$\frac{1}{|\,x\,| + \lambda\,|\,A_x\,|}.$$

At the end of a stay in x, a new state is chosen (independently of everything except x) from $A_x \cup B_x$ according to the probabilities

$$\frac{\lambda}{|\,x\,| + \lambda\,|\,A_x\,|} \qquad \text{for elements of } A_x$$

$$\frac{1}{|\,x\,| + \lambda\,|\,A_x\,|} \qquad \text{for elements of } B_x.$$

The equation $Aq = 0$ is the matrix-vector form of the statistical equilibrium equations for the process z_t. These equations can be written out and solved explicitly, as follows: $Aq = 0$ is equivalent to

$$\{|\,x\,| + \lambda\,|\,A_x\,|\}q_x = \sum_{y \in A_x} q_y + \lambda \sum_{y \in B_x} q_y, \qquad x \in S. \tag{6}$$

We find by substitution that q_x taken proportional to $\lambda^{|x|}$ gives a solution. Hence the unique normalized (to be a *probability* vector) solution is

$$q_x = \frac{\lambda^{|x|}}{\sum_{x \in S} \lambda^{|x|}} = \frac{\lambda^{|x|}}{\Phi(\lambda)}.$$

This is precisely the canonical distribution of probability over S which was obtained earlier from the maximum entropy principle. Thus, one sense in which the canonical distribution is an *equilibrium* distribution is that it is invariant under the transition probability matrices of z_t.

It will be noticed that the vector q has components which satisfy not only the statistical equilibrium equation (6) for z_t, but also the much stronger condition

$$q_x a_{xy} = q_y a_{yx} \qquad x,y \in S,$$

which is an analog of the principle of detailed balance. In the language of probability, this condition is that of *reversibility*; that is, it is equivalent to the condition that the process z_t look the same whether seen forward or backward in the sense that for any two times t and s

$$\Pr \{z_t = x \quad \text{and} \quad z_s = y\} = \Pr \{z_t = y \quad \text{and} \quad z_s = x\}.$$

The reversibility of z_t has important statistical consequences, explored in Sections XII–XIV. However, an immediate consequence is the following form of the Boltzmann H-theorem for z_t:

Remark: Let

$$H_x(t) = -\sum_{y \in S} q_{xy}(t) \log q_{xy}(t),$$

where

$$q_{xy}(t) = \Pr \{z_t = y \mid z_0 = x\}.$$

Then

$$(d/dt)H_x(t) \geqq 0.$$

The proof of this is well-known, being just Pauli's proof of the quantum-mechanical H-theorem from the principle of detailed balance. (See Tolman,[9] p. 464.)

IX. ANALOGY WITH STATISTICAL MECHANICS

As its name suggests, the canonical distribution of probability over S, implied by the maximum entropy principle, resembles the canonical ensemble of statistical mehcanics and thermodynamics. This analogy extends to several other concepts arising either in traffic theory or in statistical mechanics, and will now be described. It is assumed that the reader is familiar with the rudiments of statistical mechanics; a lucid account can be found in Khinchin.[6]

Let us consider a conservative mechanical system embedded in a heat bath, and assume that it is described by a canonical ensemble. It can

exchange energy with its surroundings; its energy is a randomly varying quantity. The basic identification we make is of the *number of calls in* progress in a connecting network with the *energy* of this mechanical system. In other words, new calls in the operating network are analogous to increments of energy in the mechanical system, while hangups represent decrements of energy. The *average energy* is identified with the *average load* carried by the network.

The surfaces of constant energy in the phase-space of the mechanical system are analogous to the levels L_k, i.e., the sets consisting of the various states with k calls in progress for $k = 0, 1, 2, \cdots$. The number $| L_k |$ of ways of putting up k calls, on which our theory rests, is the analog of the area of a surface of constant energy. Just as the canonical density function is constant over the surfaces of constant energy and maximizes the entropy for a given average energy, so is the canonical probability vector q constant over each L_k and maximizes $H(q)$.

The partition function of statistical mechanics is defined (cf. Ref. 6, p. 79) by

$$Z(\alpha) = \int_{\Gamma} e^{-\alpha H(x)}\, dV(x),$$

where Γ is the phase-space, $x \in \Gamma$ is a typical state, $H(x)$ is the total energy of state x (here given by the Hamiltonian function), and dV is the volume element of phase-space. In a similar way, the partition function $\Phi(\cdot)$ is the generating function of the numbers $| L_k |$, $k = 0, 1, 2, \cdots$. The set S of states corresponds to the phase-space Γ, $H(x)$ is analogous to $| x |$, the volume measure on Γ is analogous to the counting measure on S, and $e^{-\alpha}$ replaces ξ.

In Khinchin's development[6] of statistical mechanics the temperature is defined as inversely proportional to the unique root θ of the equation

$$(d/d\theta) \log Z(\theta) = \text{average energy}.$$

Specifically, the absolute temperature T is given by

$$\theta = (kT)^{-1},$$

where k is Boltzmann's constant. In our model for a connecting network the analog (modulo a logarithmic transformation) of θ is the solution λ of

$$(d/d\lambda) \log \Phi(\lambda) = \text{average load carried}.$$

Thus it is tempting to identify $(\log \lambda)^{-1}$ as proportional to the "temperature" of the traffic system.

The matrix A, introduced in Section VIII as the "transition rate"

matrix for the process z_t, provides a sense in which the canonical distribution q is of "equilibrium" type. The *reversibility* of z_t is analogous to the detailed balance property of transition matrices in statistical mechanics. (Cf. Tolman,[9] pp. 165 and 521.) This property also implies that a form of the Boltzmann H-theorem is valid for z_t, as we saw in Section VIII.

The analogies between our thermodynamic model of traffic and statistical mechanics can be collected in the following tabulation:

STATISTICAL MECHANICS	TRAFFIC THEORY
Energy	Calls in progress
Partition function	Generating function $\Phi(\cdot)$ of number of ways of putting up k calls, $0 \leqq k \leqq w$
Entropy	$-\sum_{x \epsilon S} q_x \log q_x$
Temperature	$\{\log$ (calling rate per idle path)$\}^{-1}$
Area of surface of given energy	$\mid L_k \mid$ = number of ways of putting up k calls
Transition rate matrix	A
Detailed balance	Reversibility of z_t
Equilibrium	$Aq = 0$
Heat bath	Idle customers' needs
Phase space Γ	Set S of possible states
Volume measure on Γ	Counting measure on S

X. DISCUSSION AND CRITIQUE

It is now reasonable to consider possible physical interpretations of the stochastic process z_t and of the transition rate matrix A in terms of items describing behavior at the inlets and outlets of the connecting network, such as calling rates, holding-times, and routing rules. Obviously, transitions of z_t from a state x into B_x represent hangups, while transitions from x into A_x represent new calls; the entries of A indicate the "rates" at which these events occur in the different states. However, the reasonableness, and so the acceptability, of z_t as a model for traffic depends on the interpretations of z_t and A in physical terms. Hence we must inquire whether (and how) the rates entered in A can be viewed as realistically describing the terminations of calls in progress, the occurrence of new calls between inlets and outlets of the network, and their routing or disposition.

In general, to construct a Markov process as a model for traffic in a

connecting network whose states form the set S, it is usually sufficient to give, for each state $x \in S$, and each inlet u and outlet v,

(i) the hangup rate for the various calls in progress in x,

(ii) the calling rate between u and v in state x,

(iii) the method for disposing of requests that encounter congestion, receive busy tone, etc.,

(iv) the method for choosing routes of new calls.

A particular choice of the items (i)–(iv) leads to a transition rate matrix, and so to a Markov process. We shall assess the reasonableness of z_t as a model for traffic in terms of items (i)–(iv) above by exhibiting two choices of (i)–(iv) that both lead to the rate matrix A of z_t.

In the dynamic model z_t described in Section VIII, the role of the inlets and outlets is open to (at least) two different interpretations, each of which induces a corresponding interpretation of the transition rate matrix A.

One possible interpretation of the inlets and outlets is to take them seriously as actual terminals or customers' lines. They are then the outermost portions of the network under study, the original sources for traffic that enters the system, beyond which there is no more connecting or switching equipment. From any inlet, or to any outlet, there can be at most one call in progress. In this case the rate matrix A can be interpreted as saying that in a state x each call in progress is terminating at a unit rate, that the calling rate from an idle inlet u to an idle outlet v is

$$\lambda \cdot \text{number of available paths from } u \text{ to } v \text{ in state } x$$
$$= \lambda \cdot \text{number of states covering } x \text{ which include a } (u,v) \text{ call,}$$

and that of the possible routes for a new call one is chosen at random (equal probability for all). The reader can verify that this choice of (i)–(iv) does in fact lead to the rate matrix A. Note that this description does not provide for the generation of blocked calls.

The choice of a unit hangup rate per call in progress is tantamount to measuring time in units of mean holding-time, with the convenience that carried and offered loads come out in the standard units of erlangs. This unit hangup rate can be obtained as a consequence of assuming that the holding-times are negative-exponentially distributed with mean unity, mutually independent, and independent of the random process describing new calls. This assumption of "negative exponential holding-times" is a standard one in congestion theory. (See e.g., Syski,[10] p. 9.)

More interesting (and questionable!) is the fact that under this interpretation the calling rate in a state x between an idle inlet u and an idle

outlet v depends on the number of ways in which a call from u to v could be put into the network in state x. This calling rate can therefore change in time as the state changes, even if u and v remain idle. It can be argued that this is an unrealistic feature, and that therefore z_t is not a wholly reasonable model for telephone traffic in a network whose inlets and outlets are interpreted as terminals or customers' lines. For surely the idle calling parties do not know the state of the system, nor the number of paths available for a call between them, and so they cannot (let alone *do* not) adjust their calling rates accordingly.

In a sense, it would be more intuitive and reasonable to assign a calling rate λ to each idle pair (u,v) of terminals (an inlet u and an outlet v) irrespective of the state x of the system. This basic calling rate for each idle pair (u,v) is then distributed over the states that cover x and realize (u,v) [assuming that (u,v) is not blocked, so that there are such states] in accordance with some routing rule. A stochastic process x_t on S based on this idea is described in Section XI, and is studied in detail in a work (Beneš[11]) to appear later.

From an *a priori* viewpoint, x_t is a more reasonable model for traffic than z_t. The objection (described above) to letting the calling rate for an idle pair depend on the state is severe. Nevertheless it does not necessarily destroy the usefulness of the process z_t for describing traffic. Three comments are relevant here:

(1) If all calls can be put up in at most one way, then x_t and z_t coincide.

(2) If calls can be put up in only a few ways, it may often be possible or useful to regard z_t as a small perturbation of x_t obtained by raising various calling rates. This idea is explored in Section XI.

(3) Even if z_t is not in any precise sense a small perturbation of the *a priori* reasonable model x_t, it deserves to be considered as a model of traffic. It must not be forgotten that the usefulness of a theory rests more on its success in predicting than on its meeting criteria of reasonableness that are adduced *a priori*.

However, it is possible to give the inlets and outlets a second interpretation, different from the one that assigns them the role of "outermost terminals." This interpretation makes z_t a fairly reasonable model of traffic, in the *a priori* sense we are discussing. It consists in letting each inlet or outlet represent a point from which several or many calls can be in progress to other points in the system. Physically, such an inlet or outlet might be a PBX or central office serving a locality. As such, it would itself contain a connecting network which is left out of account in the model. It no longer necessarily makes sense to speak of busy and idle inlets, or outlets.

To give an intuitive rationale for this interpretation and for the assumption about calling rates that corresponds to it, let us pick an inlet-outlet pair (u,v) and think of u and v as (possibly geographically separated) points between which there may be several calls in progress. For example, the network under study might be a toll network, and u and v might be local central offices acting as sources of traffic for the toll system. Or, for a second example, u and v might be distinct switches in a large network inside a central office.

In such situations, it is natural to expect that if in a state there are *many* paths available for a call from u to v, then there is a *larger* probability that a requested call from u to v arise in the next small interval of time h than if there were very few paths between u and v available. In other words, it is reasonable that the calling rate in x for (u,v) calls be a monotone increasing function of the number of paths available in x for such calls.

A particularly simple monotone function is the linear one, and we shall assume that the calling rate for an idle pair (u,v) in x is

$$\lambda \cdot \text{number of paths available in } x \text{ for } (u,v) \text{ call},$$

and that of the available paths one is chosen at random. Again, no provision is made for the generation of blocked attempts, since these will not affect the state probabilities when blocked calls are refused.

We observe that A_x can be partitioned and written as

$$A_x = \bigcup_{(u,v)} A_x(u,v),$$

where

$$A_x(u,v) = \{y\colon y \text{ covers } x \text{ and realizes } (u,v)\}$$

with

$$| A_x(u,v) | = \text{number of paths available in } x \text{ for a } (u,v) \text{ call}.$$

Since routes for new calls are chosen at random we find that the transition rate from x to $y \in A_x$ is exactly λ, so that this second interpretation also leads to the rate matrix A.

XI. A MARKOV MODEL BASED ON TERMINAL-PAIR BEHAVIOR

We now revert to interpreting inlets and outlets as the ultimate terminals of the connecting network. In Section X it was suggested that under this interpretation an *a priori* reasonable model (a stochastic process x_t) can be obtained by postulating an effective calling rate $\lambda > 0$ per idle inlet-outlet pair. This can be done by assuming that each

idle inlet calls an arbitrary outlet at a rate λ, and *vice versa*, with attempted calls to busy terminals rejected with no change of state. The total attempt rate in a state x (excluding calls to busy terminals) is

$$\lambda \cdot \left\{ \begin{matrix} \text{number of idle} \\ \text{inlet-outlet pairs in } x \end{matrix} \right\}.$$

If I is the set of inlets, and Ω that of outlets, with I and Ω disjoint, this has the quadratic form

$$\lambda(\mid I \mid - \mid x \mid)(\mid \Omega \mid - \mid x \mid).$$

As before, we assume a unit hangup rate per call in progress, with blocked calls rejected. The description of x_t can be completed, finally, by specifying a method of routing. This we do by introducing a "routing matrix" $R = (r_{xy})$ with the following properties: Let x be a state, and let Π be the partition of A_x induced by the equivalence relation $\sim$ of "having the same calls up, possibly on different routes"; then

$$\begin{aligned} r_{xy} &\geqq 0 \\ r_{xy} &= 0 \quad \text{unless} \quad y \in A_x \\ \sum_{y \in Y} r_{xy} &= 1 \quad \text{for} \quad Y \in \Pi. \end{aligned}$$

We note that $\sum_{y \in S} r_{xy}$ is exactly the number $s(x)$ of attempts which would be "successful" if they arose in state x, and that Π consists of exactly the sets $A_x(u,v)$ for $\{(u,v)\}$ idle and unblocked in x.

The routing matrix R is to have this interpretation: each time the call $\{(u,v)\}$ is to be completed in state x, a state y is chosen independently from $A_x(u,v)$ with probability r_{xy} and the call is routed so as to take the system to state y.

The foregoing assumptions lead to a rate matrix Q for x_t defined by

$$q_{xy} = \begin{cases} 1 & \text{if} \quad y \in B_x \\ \lambda r_{xy} & \text{if} \quad y \in A_x \\ - \mid x \mid - \lambda s(x) & \text{if} \quad y = x \\ 0 & \text{if} \quad y \in (A_x \cup B_x)' \quad \text{and} \quad y \neq x. \end{cases}$$

This matrix is exactly like A except that for $y \in A_x$ the rate from x to y is not λ but (the in general smaller quantity) λr_{xy}, and that the diagonal terms are correspondingly increased so as to keep row sums equal to zero. For each $Y \in \Pi$, r_{xy} for $y \in Y$ represents a distribution of the calling rate of some idle unblocked pair (u,v) over $A_x(u,v) = Y$. Indeed A results from Q if all the r_{xy} are replaced by unity. The process x_t can be defined in terms of its rate matrix Q.

The assumptions leading to the rate matrix Q and to the process x_t have much *a priori* appeal; x_t itself is discussed in detail in a forthcoming paper[11] already mentioned. Here we shall merely consider whether z_t may be regarded as a perturbation of x_t. Since each process is determined by its respective rate matrix, and since we are interested mostly in equilibrium behavior, we restrict attention to asking how different are the respective equilibrium distributions over S for x_t and z_t. Thus, if p and q are probability row-vectors satisfying $Qp = 0$ and $Aq = 0$ respectively, how different is p from q?

To give a precise estimate, we introduce the norms

$$\| M \| = \sum_{x,y} | m_{xy} |$$

$$\| v \| = \sum_{x} | v_x |$$

for matrices and vectors, respectively. Since $Ap = (A - Q)p$ and $Qq = (Q - A)q$, we find

$$\| p - q \| \leqq \frac{2 \| Q - A \|}{1 - \| Q - A \|}.$$

The norm of $Q - A$, in turn, can be seen to be

$$\begin{aligned} \| Q - A \| &= 2\lambda \sum_{x \epsilon S} \sum_{y \epsilon A_x} (1 - r_{xy}) \\ &= 2\lambda \sum_{x \epsilon S} (| A_x | - s(x)) \\ &= 2\lambda \{\Phi'(1) - \sum_{x \epsilon S} s(x)\} \end{aligned}$$

where

$s(x)$ = number of pairs that are idle and not blocked in x.

Letting

μ = max number of ways a call can be realized

we find $| A_x | \leqq \mu s(x)$, and hence

$$\begin{aligned} \| Q - A \| &\leqq 2\lambda(\mu - 1) \sum_{x \epsilon S} s(x) \\ &\leqq 2\lambda(\mu - 1)\Phi'(1). \end{aligned}$$

Let

$$w = \max_{x \epsilon S} | x |$$

so that $\Phi'(1) \leqq w \mid S \mid$, and

$$\| Q - A \| \leqq 2\lambda(\mu - 1)w \mid S \mid.$$

The average contribution (per state) to $\| Q - A \|$ is then

$$\frac{\| Q - A \|}{\mid S \mid} \leqq 2\lambda(\mu - 1)w.$$

XII. THE APPROACH TO EQUILIBRIUM

It is known from the theory of Markov processes that the matrix $Q(t) = (q_{xy}(t))$ of the transition probabilities

$$q_{xy}(t) = \Pr \{z(t + s) = y \mid z(s) = x\}, \qquad t \geqq 0$$

of the process z_t satisfies the Kolmogorov equations

$$(d/dt)Q(t) = AQ(t) = Q(t)A, \qquad Q(0) = I,$$

and that the study of the *time-dependent* (as opposed to the asymptotic, or equilibrium) behavior of z_t can be carried out in terms of the characteristic values of A. Knowledge of the transition probabilities is essential, for example, in calculating the sampling error incurred in such load averages as

$$\frac{1}{n}\sum_{j=1}^{n} \mid z_{n\tau} \mid, \qquad \frac{1}{T}\int_0^T \mid z_t \mid dt, \tag{7}$$

where τ is the interval between successive discrete observations of $\mid z_t \mid$, and $(0,T)$ is an interval of continuous observation of $\mid z_t \mid$. In this section we study the manner in which z_t approaches equilibrium in terms of the two principal characteristic values of A, i.e., that of largest, and that of smallest nonzero, magnitude. Applications to estimating the covariances of functions of z_t, and to studying sampling error for the traffic averages in (7), are described in Sections XIII and XIV, respectively.

Our study of the approach to equilibrium is based on the observation that the matrix A of transition rates for the process z_t is symmetrizable, i.e., is a self-adjoint operator in a suitably chosen inner-product space of finite dimension $\mid S \mid$. The probabilities

$$q_x = \frac{\lambda^{|x|}}{\Phi(\lambda)} = \frac{1}{\mu_x}$$

are all strictly nonnegative, and we use their reciprocals μ_x as weights in defining an inner product,

$$(r,s) = \sum_{x \epsilon S} r_x \bar{s}_x \mu_x , \tag{8}$$

and a norm,

$$\| s \| = (s,s)^{\frac{1}{2}}.$$

We now remark that for all states x,y from S,

$$q_y a_{yx} = q_x a_{xy}$$

or alternatively

$$a_{yx}\mu_x = a_{xy}\mu_y .$$

Indeed, this remark is the basis for the solution q given in Section VIII for the statistical equilibrium equations (6) of the process z_t; it has the important consequence that A is self-adjoint with respect to the inner product defined by (8), viz.

Lemma 2: $(Ar,s) = (r,As)$, *for any* $| S |$ *-vectors* r,s.

Proof: A is a real matrix, so

$$\sum_x \sum_y a_{yx} r_y \bar{s}_x \mu_x = \sum_x \sum_y r_y \bar{s}_x a_{xy} \mu_y$$

$$= \sum_y r_y \mu_y \sum_x a_{xy} \bar{s}_x = \sum_y r_y \overline{\sum_x a_{xy} s_x \mu_y}$$

In a similar way we prove

Lemma 3:

$$(Ar,s) = -\tfrac{1}{2} \sum_x \sum_y a_{yx} q_y (\overline{s_x \mu_x - x_y \mu_y})(\mu_x r_x - \mu_y r_y).$$

Proof: Since the matrix whose elements are $a_{yx}\mu_x$ is symmetric, we have

$$(Ar,s) = \sum_x \sum_y a_{yx} \mu_x r_y \bar{s}_x$$

$$= \tfrac{1}{2} \sum_x \sum_y a_{yx} \mu_x (r_y \bar{s}_x + r_x \bar{s}_y).$$

Now

$$\sum_x \sum_y a_{yx} q_y \mu_x^2 \bar{s}_r r_x = 0$$

because $Aq = 0$, and

$$\sum_x \sum_y a_{yx} q_y \mu_y^2 \bar{s}_y r_y = \sum_y \sum_x a_{yx} \mu_y \bar{s}_y r_y$$

$$= 0$$

because $\sum_x a_{yx} = 0$. This proves the lemma.

Theorem 2: The characteristic values of A are real and nonpositive. Zero is a simple characteristic value corresponding to the characteristic vector q, normalized to unity.

Proof: The result follows from the known properties of self-adjoint transformations. (See Ref. 12, pp. 153–155.)

The characteristic values of A will all be of the Rayleigh quotient form

$$\frac{(Av,v)}{(v,v)} \leqq 0$$

for some vector v; by Lemma 3 this form is nonpositive. The probability vector solution q of $Aq = 0$ is unique so that zero is a simple characteristic value. Furthermore, if $0 > r_{\max} = r_1 \geqq \cdots \geqq r_{|S|-1} = r_{\min}$ is an arrangement of the characteristic values in decreasing order, the variational description of the characteristic values (Ref. 12, p. 111) implies that with $\| v \|^2 = (v,v)$,

$$r_{\max} = r_1 = \max \{(Av,v) \mid v \perp q, \| v \| = 1\}$$

$$r_{\min} = r_{|S|-1} = \min \{(Av,v) \mid \| v \| = 1\}.$$

The alternative notations $r_{\max}$ and $r_{\min}$ identify the two "dominant" characteristic values, and are introduced for later convenience to enhance the symmetry of the theory.

One can now estimate r_1 from below by substituting suitable trial vectors in the Rayleigh quotient. Choosing a vector v with components

$$v_x = \frac{|x| - m}{\sigma\mu_x}, \qquad x \in S,$$

where

$$m = \sum_{x \in S} |x| q_x = \lambda \frac{\partial}{\partial \lambda} \log \Phi(\lambda)$$

$$\sigma^2 = \sum_{x \in S} (|x| - m)^2 q_x = \lambda^2 \left(\frac{\partial^2}{\partial \lambda^2} + \frac{\partial}{\partial \lambda}\right) \log \Phi(\lambda),$$

it is easily seen that $(q,v) = 0$, that $\| v \| = 1$, and that

$$(Av,v) = -\frac{1}{2} \sum_x \sum_y a_{yx} q_y \left(\frac{|y| - m}{\sigma} - \frac{|x| - m}{\sigma}\right)^2$$

$$= -\frac{1}{2\sigma^2} \sum_y q_y (|y| + \lambda |A_y|).$$

In equilibrium, the average rate of new calls equals the average rate of

hangups, as can be verified from the equilibrium equations $Aq = 0$. That is,

$$\sum_{y \epsilon S} | y | q_y = \lambda \sum_{y \epsilon S} q_y | A_y |,$$

and we find

$$-\frac{m}{\sigma^2} \leqq r_1 < 0,$$

a generalization of a result known (Ref. 13, p. 147) for the simple busy signal trunk group (classical Erlang model).

In general, letting $f(\cdot)$ be any function defined on the set S of states, but not identically a constant, we define

$$m_f = \sum_{x \epsilon S} f(x) q_x$$

$$\sigma_f^2 = \sum_{x \epsilon S} (f(x) - m_f)^2 q_x .$$

Choosing now a vector v with components

$$v_x = \frac{f(x) - m_f}{\sigma_f \mu_x}$$

we obtain (by repetition of previous reasoning)

$$-\frac{1}{2\sigma_f^2} \sum_y q_y \left(\sum_{x \epsilon B_x} [f(x) - f(y)]^2 + \lambda \sum_{x \epsilon A_y} [f(x) - f(y)]^2 \right)$$

as a lower bound for r_1.

We now define a set of *vector-valued* functions $\{c_x(t), x \epsilon S, t \geqq 0\}$ by the condition

$$c_{xy}(t) = q_{xy}(t) - q_y , \qquad y \epsilon S.$$

The function $c_x(\cdot)$ describes the approach to equilibrium from the initial state x at time $t = 0$.

Theorem 3: For $t \geqq 0$

$$\| c_x(0) \| \exp (r_{\min} t) \leqq \| c_x(t) \| \leqq \| c_x(0) \| \exp (r_{\max} t).$$

Proof: Since q_y and $q_{xy}(t)$ are both distributions in y, we have

$$(c_x(t), q) = \sum_y q_{xy}(t) - q_y = 0$$

so that $c_x(t) \perp q$. Also

$$\frac{d}{dt} \| c_x(t) \|^2 = 2 \sum_y [q_{xy}(t) - q_y] \frac{d}{dt} q_{xy}(t) \mu_y$$

$$= 2 \left(c_x , \frac{d}{dt} c_x \right)$$

$$= 2(c_x , Ac_x)$$

since $(d/dt)c_x = Ac_x$; that is for each $y \in S$,

$$\frac{d}{dt} c_{xy} = \frac{d}{dt} q_{xy} = \sum_z q_{xz}(t) a_{zy}$$

$$= \sum_z [q_{xz}(t) - q_z] a_{zy} .$$

Hence, $\| c_x \|$ being nonzero, we find

$$2r_{\min} \leqq (d/dt) \log \| c_x \|^2 \leqq 2r_{\max}$$

and Theorem 3 follows by integration. The argument just given is essentially reproduced from Kramer.[14]

XIII. COVARIANCES OF FUNCTIONS OF z_t

For the purposes of this section it is convenient to introduce an inner product $(\cdot , \cdot)'$, closely related to but different from $(\cdot , \cdot)$ of the previous section, and defined by

$$(r,s)' = \sum_{x \in S} r_x \bar{s}_x q_x .$$

The associated norm is denoted by $\| r \|' = (r,r)'^{\frac{1}{2}}$. The point of the "prime" notation is explained by the fact that the transpose A' of A is self-adjoint with respect to $(\cdot , \cdot)'$.

Remark: *Where A' is the transpose of A*

$$(A'r,s)' = (r,A's)'.$$

Proof:

$$\sum_x \sum_y a_{xy} r_y \bar{s}_x q_x = \sum_x \sum_y a_{yx} \bar{s}_x r_y q_y$$

$$= \sum_y r_y \sum_x a_{yx} \bar{s}_x q_y = \sum_y r_y \overline{\sum_x a_{yx} s_x q_y} .$$

Let $f(\cdot)$ be a function defined on S, and define a stochastic process f_t by the condition

$$f_t = f(z_t).$$

Theorem 4: The covariance of f_t is given by

$$R_f(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} (f, A'^n f)',$$

where the vector f is defined by

$$f_x = f(x) - \sum_{x \epsilon S} f(x) q_x ,$$

$$= f(x) - m_f .$$

Proof: The covariance of f_t is

$$\sum_x \sum_y q_x q_{xy}(t) f_x f_y = (f, Q(t)'f)'$$

$$= (f, (\exp tA)'f)'$$

$$= \sum_{n=0}^{\infty} \frac{t^n}{n!} (f, A'^n f)'$$

with $Q(t)'$ denoting the transpose, and not the derivative, of $Q(t)$. The covariance of f_t is thus the exponential generating function of the series of numbers

$$(f, A'^n f)' \qquad n = 0, 1, 2, \cdots .$$

These can be calculated with the help of the following results:

Lemma 4: Let the matrix elements of A^n be $a_{xy}^{(n)}$. Then

$$q_x a_{xy}^{(n)} = q_y a_{yx}^{(n)}.$$

Proof:

$$q_x a_{xy}^{(n)} = q_x \sum_{u_1, \cdots, u_{n-1}} a_{xu_1} a_{u_1 u_2} \cdots a_{u_{n-1} y}$$

$$= \sum_{u_1, \cdots, u_{n-1}} a_{u_1 x} a_{u_2 u_1} \cdots a_{y u_{n-1}} q_y$$

$$= q_y \sum_{u_{n-1}, \cdots, u_1} a_{y u_{n-1}} \cdots a_{u_2 u_1} a_{u_1 x}$$

$$= q_y a_{yx}^{(n)}.$$

Lemma 5: Let Q be the diagonal matrix of elements q_x, $x \epsilon S$. Then

$$(w, A'^n w)' = (A^n Q w, Q w).$$

Proof:

$$\sum_x w_x \overline{\sum_y a_{xy}^{(n)} w_y} q_x = \sum_x (Qw)_x \sum_y \mu_y a_{xy}^{(n)} \overline{(Qw)_y}$$

$$= \sum_x (Qw)_x \sum_y a_{yx}^{(n)} \overline{(Qw)_y} \mu_x .$$

From the three preceding results we obtain

Theorem 5: The covariance of f_t is

$$R_f(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} (A^n Qf, Qf),$$

where the vector f is as in Theorem 4, and Q is the diagonal matrix of elements q_x, $x \in S$.

It is readily seen that A^n, $n \geqq 1$, is again a self-adjoint operator with respect to $(\cdot,\cdot)$, and that its characteristic values are precisely the nth powers of those of A. Also, for any vector v and $n \geqq 0$

$$(A^n v,v) \begin{cases} \leqq 0 & \text{if } n \text{ is odd} \\ \geqq 0 & \text{if } n \text{ is zero or even} \end{cases}$$

so that by the variational description of characteristic values we have

$$\left.\begin{matrix} r_{\min}{}^n, & n \text{ odd} \\ r_{\max}{}^n, & n \text{ even} \end{matrix}\right\} \leqq \frac{(A^n v,v)}{(v,v)} \leqq \begin{cases} r_{\max}{}^n, & n \text{ odd} \\ r_{\min}{}^n, & n \text{ even} \end{cases} \tag{10}$$

provided that $v \perp q$ (in the inequalities involving $r_{\max}$). Returning now to the vector Qf of Theorem 5, we find

$$\begin{aligned} \| Qf \|^2 &= \sum_x q_x^2 f_x^2 \mu_x \\ &= \sum_{x \in S} (f(x) - m_f)^2 q_x = \sigma_f^2 \end{aligned}$$

and

$$\sum_x q_x (Q_f)_x \mu_x = \sum_x q_x f_x = 0,$$

so that $Qf \perp q$. Letting $v = Qf$ in (10), we obtain

$$\sigma_f^2 r_{\min}{}^n \leqq (A^n Qf, Qf) \leqq \sigma_f^2 r_{\max}{}^n, \quad n \text{ odd},$$

$$\sigma_f^2 r_{\max}{}^n \leqq (A^n Qf, Qf) \leqq \sigma_f r_{\min}{}^n, \quad n \text{ even}.$$

Unfortunately, these inequalities do not give useful bounds for the covariance $R_f(\cdot)$. However, such bounds can be obtained from the formula of Theorem 5 in an elegant way by applying the spectral theorem to A.

Theorem 6: Let $\alpha_1, \cdots, \alpha_k$ denote the distinct *characteristic values of A, and let E_i, $i = 1, \cdots, k$, denote the perpendicular projection on the subspace of all solutions $Ar = \alpha_i r$. Then the covariance $R_f(\cdot)$ of f_t is given by*

$$R_f(t) = \sum_{i=0}^{k} (E_i Qf, Qf) e^{\alpha_i t},$$

with $1 \leqq k \leqq |\ S\ |$, Q *the diagonal matrix of elements* q_x, $x \in S$, *and* f *given by* $f_x = f(x) - m_f$.

Proof: By the spectral theorem for self-adjoint operators (Ref. 12, p. 56) we can write

$$A = \sum_{i=0}^{k} \alpha_i E_i$$

and

$$A^n = \sum_{i=0}^{k} \alpha_i{}^n E_i .$$

We can now calculate with formula (9) of Theorem 5:

$$\begin{aligned} R_f(t) &= \sum_{n=0}^{\infty} \frac{t^n}{n!} (A^n Qf, Qf) \\ &= \sum_{n=0}^{\infty} \sum_{i=0}^{k} \frac{(\alpha_i t)^n}{n!} (E_i Qf, Qf) \\ &= \sum_{i=0}^{k} (E_i Qf, Qf)\ e^{\alpha_i t} . \end{aligned}$$

This proves Theorem 6. Since we know that zero is among the characteristic values (indeed, it is a simple one), one of the α's, say α_0, will be zero. We may reasonably expect $R_f(\cdot)$ to approach zero for large t; hence the constant, i.e., α_0, term of $R_f(\cdot)$ should be zero. This can be seen as follows: the subspace associated with zero consists of vectors proportional to the equilibrium vector q, because zero is a simple characteristic value; but we have already verified that $q \perp Qf$; hence

$$(E_0 r, Qf) = 0, \qquad \text{all } r.$$

Using this we prove the

Corollary 1: $R_f(t) \geqq 0$ *for all* t, *and in fact*

$$0 \leqq \sigma_f{}^2\, e^{r_{\min}|t|} \leqq R_f(t) \leqq \sigma_f{}^2\, e^{r_{\max}|t|}, \qquad \text{all } t.$$

Proof: Since the E_i of Theorem 6 are perpendicular projections, they are linear, self-adjoint, and positive in the sense of Ref. 12, p. 140; the usual term for positive is nonnegative semidefinite. Hence

$$(E_i r, r) \geqq 0$$

for any vector r. Since $(E_0 r, Qf) = 0$ if E_0 is associated with the zero characteristic value, the result follows from Theorem 6, using

$$\sum_{i=0}^{k} E_i = I,$$

$$\sum_{i=0}^{k} (E_i Qf, Qf) = \| Qf \|^2 = \sigma_f^2.$$

XIV. APPLICATIONS TO SAMPLING ERROR

Let us suppose that n samples of the process $f_t (= f(z_t))$ are observed during an interval of equilibrium of z_t at intervals τ apart, and that the normed sum

$$n^{-1} S_n = n^{-1} \sum_{j=1}^{n} f_{j\tau}$$

is used as an estimate of $E\{f_t\}$. We find that

$$\text{Var } \{S_n\} = \sum_{j=-n}^{n} (n - |j|) R_f(j\tau),$$

where $R_f(\cdot)$ is the covariance of f_t. By using the identity

$$\sum_{j=-n}^{n} (n - |j|) e^{-2|j|u} = n \operatorname{ctnh} u - \frac{1 - e^{-2nu}}{2} \operatorname{csch}^2 u,$$

$$= v_n(u),$$

together with Corollary 1 of Section XIII, we find that

$$\sigma_f^2 v_n(-\tfrac{1}{2}\tau r_{\min}) \leqq \text{Var } \{S_n\} \leqq \sigma_f^2 v_n(-\tfrac{1}{2}\tau r_{\max}).$$

In a similar way, if f_t is observed continuously over an interval $(0,T)$ of equilibrium of z_t and the time average

$$M(T) = \frac{1}{T} \int_0^T f(z_t)\, dt$$

is used as an estimate of $E\{f_t\}$, then

$$\text{Var } \{M(T)\} = 2T^{-2} \int_0^T (T - t)\, R_f(t)\, dt,$$

and Corollary 1 gives

$$\sigma_f^2 \int_0^T (T - t)\, e^{r_{\min} t}\, dt \leqq \text{Var } \{M(T)\} \leqq \sigma_f^2 \int_0^T (T - t)\, e^{r_{\max} t}\, dt.$$

XV. A GENERALIZATION

As an extension of the maximum problem posed and solved in Section V we shall seek functions

$$q_x(t),\ x \in S, \qquad t_1 \leqq t \leqq t_2, \qquad t_1 < t_2$$

such that for each t in $[t_1, t_2]$

$$\sum_{x \in S} q_x(t) = 1, \qquad q_x(t) \geqq 0$$

$$\sum_{x \in S} |x| \, q_x(t) = m(t) > 0$$

$$\int_{t_1}^{t_2} H(q(t)) \, dt = \text{maximum.}$$

In other words, we look for a time-dependent distribution of probability over S with prescribed mean values for the function $|\cdot|$ on S, such that the integral of the entropy functional over (t_1, t_2) is a maximum.

The Euler equations for this problem assume the trivial form (with $L_1(\cdot)$ and $L_2(\cdot)$ as Lagrange's multipliers):

$$(\partial H/\partial q_x) - L_1(t) \, |x| - L_2(t) = 0, \qquad x \in S$$

or, writing out the H-derivative,

$$\log q_x(t) + 1 + L_1(t) \, |x| + L_2(t) = 0, \qquad x \in S.$$

The argument of Lemma 1 following equation (4) shows that $q_x(\cdot)$ is given by

$$q_x(t) = \frac{\lambda(t)^{|x|}}{\Phi(\lambda(t))} \, t_1 \leqq t \leqq t_2$$

where $\lambda(\cdot)$ is the unique solution of the equation

$$m(t) = \frac{\sum_{x \in S} |x| \, \lambda(t)^{|x|}}{\Phi(\lambda(t))} = \left(u \frac{d}{du} \log \Phi(u) \right)_{u=\lambda(t)}.$$

This solution has the form of the canonical distribution at each time point in $[t_1, t_2]$, and Theorem 1 in effect is just the special case of this result that arises when $m(t) \equiv m$. It is apparent that the form of this solution does not depend on what interval $[t_1, t_2]$ was considered, so we may assume that $m(\cdot)$, and hence also $\lambda(\cdot)$ and $q(\cdot)$, are defined on the real axis.

Let us define the matrix-valued function $A(t)$ by $A(t) = (a_{xy}(t))$ where

$$a_{xy}(t) = \begin{cases} 1 & \text{if } \ y \in B_x \\ \lambda(t) & \text{if } \ y \in A_x \\ -|x| - \lambda(t) \, |A_x| & \text{if } \ x = y \\ 0 & \text{otherwise.} \end{cases}$$

In other words $A(t)$ is obtained from the transition rate matrix A or z_t

by replacing the constant λ by the function $\lambda(\cdot)$. Then for each t

$$A(t)q(t) = 0$$

i.e.,

$$\{|\, x \,| + \lambda(t) \,|\, A_x \,|\}q_x(t) = \sum_{y \epsilon A_x} q_y(t) + \lambda(t) \sum_{y \epsilon B_x} q_y(t).$$

Thus an analog of the statistical equilibrium equation holds at each point in time, and in this sense, a system described by $\{q(t), t_1 \leqq t \leqq t_2\}$ may be said to be locally in equilibrium throughout the interval (t_1, t_2).

Let us now redefine the process z_t to be the time-dependent Markov process corresponding to the (time-dependent) transition rate matrix $A(\cdot)$. We know that if $\lambda(\cdot)$ were a constant function with the particular value $\lambda(u)$, then the process z_t would have a stationary or equilibrium distribution over S given by

$$q_x(u) \frac{[\lambda(u)]^{|x|}}{\Phi(\lambda(u))}.$$

We may therefore expect that if $\lambda(\cdot)$ is not constant, but changes only *slowly* with time, and if z_0 has the absolute distribution (vector) $q(0)$, then z_t for $t > 0$ has a distribution *approximately* given by $q(t)$. Let us see in detail how this occurs.

The transition probability matrix

$$Q(t_1, t_2) = (q_{xy}(t_1, t_2)),$$

$$q_{xy}(t_1, t_2) = \Pr\{z_{t_2} = y \mid z_{t_1} = x\},$$

is now indexed by two time parameters instead of one, because of the time-dependence of z_t. The forward Kolmogorov equation for $Q(\cdot, \cdot)$ is

$$(\partial/\partial t)Q(u,t) = Q(u,t)A(t), \qquad u < t,$$

or

$$(\partial/\partial t)q_{xy}(u,t) = -[|\, y \,| + \lambda(t)s(y)]q_{xy}(u,t) + \sum_{z \epsilon A_y} q_{xz}(t) + \lambda(t) \sum_{z \epsilon B_y} q_{xz}(t)$$

with $Q(t,t) = I$. It is easily seen that

$$Q(u,t) = \exp \int_u^t Q(w)\, dw,$$

the exponential of a matrix being defined by the usual series in powers of the matrix. Therefore if

$$\Pr\{z_0 = x\} = q_x(0),$$

then

$$\Pr\{z_t = x\} = \sum_{y \in S} \Pr\{z_0 = y\} q_{yx}(t),$$

and the absolute distribution of z_t is given by the vector

$$Q(0,t)q(0).$$

We now write

$$Q(0,t) = \exp tA(t) + \left\{\exp \int_0^t A(u)\, du - \exp tA(t)\right\},$$

observe that

$$\int_0^t A(u)\, du - tA(t) = \int_0^t u\, dA(u)$$

and make this heuristic argument: Since dA/du is small, $Q(0,t)$ is approximately $\exp tA(t)$; however, for t fixed and u varying

$$\exp uA(t)$$

is the transition probability matrix of a process z_t obtained by setting

$$\lambda(u) \equiv \lambda(t).$$

This matrix approaches, as $u \to \infty$, the matrix each of whose rows is $q(t)$. If t is large compared to the time it takes this to happen, we may expect, by Theorem 3, that

$$\exp tA(t)$$

have rows all approximately equal to $q(t)$, so that

$$\text{distr}\{z_t\} \approx q(t).$$

The informal argument just given can be made precise. For the purposes of this last section, we again introduce the vector norm

$$\| f \| = \sum_{x \in S} | f_x |$$

and the matrix norm

$$\| M \| = \sum_{x,y \in S} | m_{xy} |.$$

Also, we use the following result:

Lemma 6: Let M,N be $| S |$-dimensional square matrices, with

$$c = \tfrac{1}{2} \sup_{x,y \epsilon S} | m_{xy} | + \tfrac{1}{2} \sup_{x,y \epsilon S} | n_{xy} |$$

Then for integers $k \geqq 1$

$$\| M^k - N^k \| \leqq (2c \mid S \mid)^{k-1} \| M - N \|.$$

Proof: There obtains the identity

$$M^k - N^k = \tfrac{1}{2}(M + N)(M^{k-1} - N^{k-1}) + \tfrac{1}{2}(M - N)M^{k-1} + N^{k-1}).$$

If (b_{xy}) are the elements of B and $(b_{xy}^{(k)})$ are those of B^k, then

$$\sup_{x,y} | b_{xy}^{(k)} | \leqq | S | \cdot \sup_{x,y} | b_{xy}^{(k-1)} | \cdot \sup_{x,y} | b_{xy} |.$$

Hence, with $k > 1$

$$\sup_{x,y} | m_{xy}^{(k-1)} + n_{xy}^{(k-1)} | \leqq | S |^{k-2} (2c)^{k-1},$$

and

$$\| (M - N)(M^{k-1} + N^{k-1}) \| \leqq (2c \mid S \mid)^{k-1} \cdot \| M - N \|.$$

Also,

$$\| (M + N)(M^{k-1} - N^{k-1}) \| \leqq c \mid S \mid \cdot \| M^{k-1} - N^{k-1} \|.$$

Thus

$$\begin{aligned} \| M^k - N^k \| &\leqq c \mid S \mid \cdot \| M^{k-1} - N^{k-1} \| + \tfrac{1}{2} \| M - N \| (2c \mid S \mid)^{k-1} \\ &\leqq \| M - N \| (2c \mid S \mid)^{k-1} \\ &\qquad \cdot \{ (\tfrac{1}{2})^{k-1} + \frac{1}{2} \sum_{j=0}^{k-2} (c \mid S \mid)^j (2c \mid S \mid)^{-j} \} \\ &\leqq \| M - N \| (2c \mid S \mid)^{k-1}. \end{aligned}$$

Using the lemma we find that the norm of

$$\exp \int_0^t A(u)\, du - \exp tA(t)$$

is at most

$$\sum_{n=1}^{\infty} \frac{\left\| \left(\int_0^t A(u)\, du \right)^n - t^n A^r(t) \right\|}{n!} \leqq \frac{e^b - 1}{f} \left\| \int_0^t u\, dA(u) \right\|$$

where

$$b = 2 \mid S \mid \sup_{x,y} \left| \int_0^t a_{xy}(u)\, du \right| + t \sup_{x,y} | a_{xy}(t) |.$$

It can be verified that

$$\left\| \int_0^t u \, dA(u) \right\| = 2 \sum_{\substack{x,y \\ y \in A_x}} \left| \int_0^t u \, d\lambda(u) \right|$$

$$\leqq 2t^2 \Phi'(1) \cdot \sup_{u \in [0,t]} |\lambda'(u)| .$$

Thus if $\lambda'(\cdot)$ is small on $[0,t]$ the distribution of z_t is nearly

$$(\exp tA(t))q(0)$$

(in the sense of the vector norm of this section). By Theorem 3, however, this will be nearly $q(t)$ (in the sense of the norm defined by $q(t)$) if t is large compared to the time it takes $\exp uA(t)$ to approach its limit as $u \to \infty$.

XVI. ACKNOWLEDGMENTS

The author expresses his appreciation of the help and encouragement given him in the preparation of this study by A. Descloux, J. R. Pierce, H. O. Pollak, J. Riordan, and E. Wolman.

REFERENCES

1. Beneš, V. E., Heuristic Remarks and Mathematical Problems Regarding the Theory of Connecting Systems, B.S.T.J., **41,** July, 1962, pp. 1201–1247.
2. Beneš, V. E., Algebraic and Topological Properties of Connecting Networks, B.S.T.J., **41,** July, 1962, pp. 1249–1274.
3. Lee, C. Y., Analysis of Switching Networks, B.S.T.J., **34,** November, 1955, pp. 1287–1315.
4. LeGall, P., Methode de Calcul de L'encombrement dans les Systèmes Téléphoniques Automatiques à Marquage, Ann. des Télécom., **12** (1957), pp. 374–386.
5. Jensen, A., An Elucidation of Erlang's Statistical Works Through the Theory of Stochastic Processes, in *The Life and Works of A. K. Erlang*, Trans. Danish Acad. Sciences, 1948, pp. 23–100.
6. Khinchin, A. I., *Mathematical Foundations of Statistical Mechanics*, Dover, New York, 1949.
7. Birkhoff, G., Lattice Theory, Amer. Math. Soc. Colloq. Publ. XXV, rev. ed., 1948.
8. Doob, J. L., *Stochastic Processes*, John Wiley and Sons, New York, 1953.
9. Tolman, R. C., *The Principles of Statistical Mechanics*, Oxford Univ. Press, London, 1955.
10. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London, 1960.
11. Beneš, V. E., Markov Processes Representing Traffic in Connecting Networks, to appear.
12. Halmos, P. R., *Finite-Dimensional Vector Spaces*, second ed., Van Nostrand, Princeton, 1958.
13. Beneš, V. E., The Covariance Function of a Simple Trunk Group, with Applications to Traffic Measurement, B.S.T.J., **40,** January, 1961, pp. 117–148.
14. Kramer, H. P., Symmetrizable Markov Matrices, Ann. Math. Stat. **30** (1959), p. 149.