The Annals of Applied Probability
1991, Vol. 1, No. 3, 319–378

SPECIAL INVITED PAPER

LOSS NETWORKS

BY F. P. KELLY

University of Cambridge

This paper describes work on the stochastic modelling of loss networks. Such systems have long been of interest to telephone engineers and are becoming increasingly important as models of computer and information systems. Throughout the century problems from this field have provided an impetus to the development of probability theory, pure and applied. This paper provides an introduction to the area and a review of recent work.

1. Introduction. Modern computer and telecommunications networks are able to respond to randomly fluctuating demands and failures by rerouting traffic and by reallocating resources. They are able to do this so well that, in many respects, large-scale networks appear as coherent, almost intelligent, organisms. The design and control of such networks requires an understanding of a variety of fundamental issues and is providing an important stimulus to many areas of mathematics.

In this paper we describe work on a particular class of networks, called loss networks. Our main aim is to present a coherent account of the theory of loss networks, but in passing we hope to indicate connections with, and open questions involving, many more general areas of probability theory. In particular, we shall touch on central limit theorems and weak convergence, interacting particle systems and phase transitions, random graphs and dynamic programming.

1.1. *The historical context.* In 1917 the Danish mathematician A. K. Erlang published his famous formula,

$$E(\nu, C) = \frac{\nu^C}{C!}\left[\sum_{n=0}^{C} \frac{\nu^n}{n!}\right]^{-1}, \tag{1.1}$$

for the loss probability of a telephone system ([6], page 139). The problem considered by Erlang can be phrased as follows. Calls arrive at a link as a Poisson process of rate ν. The link comprises C circuits, and a call is blocked and lost if all C circuits are occupied. Otherwise the call is accepted and occupies a single circuit for the holding period of the call. Call holding periods

Received December 1990.
AMS 1980 *subject classifications.* Primary 60K35; secondary 60K30, 90B15.
Key words and phrases. Erlang's formula, blocking probability, phase transition, network flow, repacking, queueing network.

are independent of each other and of arrival times and are identically distributed with unit mean. Then *Erlang's formula* (1.1) gives the proportion of calls that are lost.

Erlang obtained the formula (1.1) from his development of the concept of statistical equilibrium, and he powerfully demonstrated the concept's ability to deliver exact formulae for many of the important and typical problems of telephony [6]. The concept had been used in other domains: Erlang mentions known applications to Euler's theory on the age distribution of a population, and contemporary work on the ultimate consequences of Mendel's laws of heredity ([6], page 215). Perhaps the most important influence on Erlang was the development through the preceding half-century of statistical mechanics. Certainly he provided an elementary proof of Maxwell's law for the velocity of gas molecules ([6], pages 222–226), and he stressed comparisons between the rôle of probability in such applications and its rôle in the theory of telephone traffic ([6], page 194).

We now identify Erlang's concept of statistical equilibrium with the stationary measure of a Markov process. Thus, if call holding periods are exponentially distributed, then the number of lines occupied is a finite Markov chain and (1.1) gives the stationary probability that all C circuits are busy. If call holding periods are arbitrarily distributed, then the stochastic process describing the number of circuits occupied is more complex. Nevertheless, (1.1) still gives the stationary probability that all C circuits are busy. This was known to Erlang ([6], pages 205–208), but a strict proof was first presented in 1957, by Sevastyanov [64]. Indeed, we now know that the result holds under much weaker independence assumptions; for example, the loss probability remains the same if the successive call holding periods associated with a given circuit are a dependent stationary sequence. This fact motivated and was established by the modern theory of insensitivity [17].

We have seen that Erlang's formula (1.1) gives the stationary probability that all C circuits are busy. Hence, since the arrival stream is Poisson, it also gives the probability that a typical call is lost. The Poisson character of the arrival stream is, of course, important. For example, suppose a call blocked at the link described previously has a second chance: it overflows to a second link comprising C' circuits. The probability that a call is lost from the combined system is $E(\nu, C + C')$, and so the probability that a typical call overflowing the first link is lost is $E(\nu, C + C')/E(\nu, C)$. This differs from $E(\nu E(\nu, C), C')$, the probability that would obtain if the overflow stream, of mean rate $\nu E(\nu, C)$, were Poisson. The second link provides an example of how the probabilistic structure of an arrival stream can affect loss probabilities and motivated the work of Palm [57] and Khintchine [40] on Palm distributions. Palm and Khintchine also gave sufficient conditions for the superposition of independent stationary point processes to approach a Poisson process ([11], Chapter 4). Systems involving both overflow and superposition will figure frequently later in this paper.

In 1925 Erlang described the application of the method of statistical equilibrium to determine loss probabilities in simple networks of links ([6], page 203).

The approach was extended by Beneš, who in 1965 presented his systematic study of the structural aspects of networks and further developed the analogy between traffic theory and statistical mechanics [5]. However, exact results were difficult to obtain, and telecommunications engineers relied increasingly on approximations and simulation ([69, 75]). A recurring theme of this paper will be the search for a mathematical relationship between some of these approximations and exact results.

In recent years there has been a resurgence of interest in the mathematical theory of loss networks and in its application to the design and control of telecommunication systems. This has been prompted in part by the successful development over the last few decades of the theory of queueing networks (see, e.g., [31, 42, 68]) and in part by the challenge of the conceptual issues raised by advances in the technology of computer and communication systems (see, e.g., [24, 63]).

1.2. *A loss network with fixed routing.* Next we define the basic model of a loss network.

Consider a network with J links, labelled $1, 2, \ldots, J$, and suppose that link j comprises C_j circuits. A call on route r uses A_{jr} circuits from link j, where $A_{jr} \in \mathbb{Z}_+$. Let $\mathscr{R}$ be the set of possible routes. In the important special case where each element of the matrix $A = (A_{jr},\ j = 1, 2, \ldots, J,\ r \in \mathscr{R})$ is either 0 or 1, a route r can be identified with a subset of the set of links $\{1, 2, \ldots, J\}$: just let $r = \{j\colon A_{jr} = 1\}$. Calls requesting route r arrive as a Poisson stream of rate ν_r, and as r varies it indexes independent Poisson streams. A call requesting route r is blocked and lost if on any link j, $j = 1, 2, \ldots J$, there are fewer than A_{jr} circuits free. Otherwise the call is connected and simultaneously holds A_{jr} circuits from link j, $j = 1, 2, \ldots, J$, for the holding period of the call. The call holding period is independent of earlier arrival times and holding periods; holding periods of calls on route r are identically distributed with unit mean.

Let $n_r(t)$ be the number of calls in progress at time t on route r, and define the vectors $n(t) = (n_r(t),\ r \in \mathscr{R})$ and $C = (C_1, C_2, \ldots, C_J)$. Then the stochastic process $(n(t),\ t \geq 0)$ has a unique stationary distribution and under this distribution $\pi(n) = \mathbb{P}\{n(t) = n\}$ is given by

$$\pi(n) = G(C)^{-1} \prod_{r \in \mathscr{R}} \frac{\nu_r^{n_r}}{n_r!}, \qquad n \in \mathscr{S}(C), \tag{1.2}$$

where

$$\mathscr{S}(C) = \{n \in \mathbb{Z}_+^{\mathscr{R}}\colon An \leq C\} \tag{1.3}$$

and $G(C)$ is the normalizing constant (or partition function)

$$G(C) = \left(\sum_{n \in \mathscr{S}(C)} \prod_{r \in \mathscr{R}} \frac{\nu_r^{n_r}}{n_r!} \right). \tag{1.4}$$

This result is easy to check in the case where holding times are exponentially

distributed: then $(n(t),\ t \geq 0)$ is a Markov process and the distribution (1.2) satisfies the detailed balance conditions

$$\pi(n) \cdot \nu_r = \pi(n + e_r) \cdot (n_r + 1), \qquad n, n + e_r \in \mathscr{S}(C),$$

where $e_r = (I[r' = r], r' \in \mathscr{R})$ is the unit vector describing just one call in progress on route r. In this form the result has been known for many years [6]; see [9], [17], [31], and [73] for a discussion of the insensitivity of (1.2) to holding time distributions.

The classical example of this model is a telephone network, and hence we couch the definition in terms of calls, links and circuits. The model also arises naturally in the study of local area networks, multiprocessor interconnection architectures, database structures, mobile radio and broadband packet networks (see [24, 32, 45, 53, 59]). In computer communication networks, and increasingly in telephone networks, the circuits are virtual rather than physical: for example, a fixed proportion of the transmission capacity of a communication channel. The term "circuit-switched" is common in some application areas, where it is used to describe systems in which before a request (which may be a call, a task or a customer) is accepted it is first checked that sufficient resources are available to deal with each stage of the request.

Part of the model's attraction is that very many generalizations are readily incorporated. For example, if calls requesting route r arrive at rate ν_r/η_r and have holding periods with mean η_r, then the distribution π associated with the resulting stochastic process is given by the unaltered expression (1.2). If the arrival rate of calls labelled r depends on n_r (the Engset model, for example, assumes a finite source population of each type of call), then the corresponding distribution π is given by a minor variant of the form (1.2). More subtly, if a call can be carried on a number of routes and can be shifted or shared between these routes while the call is in progress, then an equivalent loss network can be defined (Section 3.3). These variations retain the essential features of the model: that a call makes simultaneous use of a number of resources and that blocked calls are lost.

Most quantities of interest can be written in terms of the distribution (1.2) or the partition function (1.4). For example, let L_r be the stationary probability that a call requesting route r is lost. Since the arrival stream of calls requesting route r is Poisson,

$$1 - L_r = \sum_{n \in \mathscr{S}(C - Ae_r)} \pi(n) = G(C)^{-1} G(C - Ae_r). \tag{1.5}$$

Such simple explicit forms might be thought to provide the complete solution. However, this is far from the case. For all but the smallest networks it is impractical to compute G directly—observe that the number of routes $|\mathscr{R}|$ may grow as fast as exponentially with the number of nodes J and that, in the (otherwise trivial) case when $|\mathscr{R}| = J$ and $A = I$, the size of the state space $|\mathscr{S}(C)| = \prod_{j=1}^{J} C_j$ grows rapidly with the capacity limitations $C_1, C_2, \ldots, C_J$. More formally, even in the restricted case where links have capacity 1 and arrival rates are equal, the task of computing the partition function (1.4) is

#P-complete [49]. A theme of much recent work has been to find approaches to the model which avoid these computational problems and which provide deeper insights. Indeed, we shall see that in many respects the properties of the model become simpler the larger or more complex the network.

Observe that the form (1.2) is obtained by truncating $|\mathscr{R}|$ independent Poisson random variables to the polytope (1.3). The natural approximation to the resulting distribution is that obtained by truncating a multivariate normal distribution to a polytope. In Section 2 we develop this approximation, by considering a limiting regime in which the capacities C_j, $j = 1, 2, \ldots, J$, and the offered traffics ν_r, $r \in \mathscr{R}$, are increased together (with ratios C_j/ν_r held fixed). The limiting behaviour of loss probabilities L_r under this regime has a very simple description. There is a parameter $B_j \in [0, 1)$ associated with link j such that

$$1 - L_r \to \prod_{j=1}^{J} (1 - B_j)^{A_{jr}}, \qquad r \in \mathscr{R}. \tag{1.6}$$

For example, if A is a 0–1 matrix, then

$$1 - L_r \to \prod_{j \in r} (1 - B_j), \qquad r \in \mathscr{R}. \tag{1.7}$$

It is *as if* links block independently, link j blocking with probability B_j. The parameters $B = (B_1, B_2, \ldots, B_J)$ emerge as the solution to a straightforward optimization problem involving just J variables: under the identification $B_j = 1 - \exp(-y_j)$, the vector B is just a stationary point of the convex function

$$\sum_{r \in \mathscr{R}} \nu_r \exp\left(- \sum_{j=1}^{J} y_j A_{jr} \right) + \sum_{j=1}^{J} y_j C_j. \tag{1.8}$$

In Section 2 we present the central limit theorem for large capacity networks under fixed routing, from which we deduce the limit (1.6) as a form of the law of large numbers. The central limit theorem allows a number of other, more detailed, conclusions to be drawn. In particular, we shall see that under the limiting regime a link may be classified as overloaded, critically loaded or underloaded, with important consequences for the process describing the number of free circuits at the link.

1.3. *The Erlang fixed point.* Section 3 is concerned with an important approximation procedure. For a loss network with fixed routing and A a 0–1 matrix, let $E_1, E_2, \ldots, E_J$ be a solution to the equations

$$E_j = E(\rho_j, C_j), \qquad j = 1, 2, \ldots, J, \tag{1.9}$$

where

$$\rho_j = \sum_{r:\, j \in r} \nu_r \prod_{i \in r - \{j\}} (1 - E_j) \tag{1.10}$$

and the function E is Erlang's formula (1.1). We shall see that the equations

have a unique solution, which we term the *Erlang fixed point*. Then an approximation for the loss probability on route r is given by

$$1 - L_r \cong \prod_{j \in r} (1 - E_j), \qquad r \in \mathscr{R}. \tag{1.11}$$

The idea underlying the approximation is simple to explain. Suppose that a Poisson stream of rate ν_r is thinned by a factor $1 - E_i$ at each link $i \in r - \{j\}$ before being offered to link j. If these thinnings could be assumed independent both from link to link and over all routes passing through link j (they clearly are not), then the traffic offered to link j would be Poisson at rate (1.10), the blocking probability at link j would be given by (1.9) and the loss probability on route r would satisfy (1.11) exactly. Call expression (1.10) the *reduced load* on link j.

The preceding approximation is worthy of study for a number of reasons. First, it has a long history in the telecommunications literature and is frequently used in practice ([33, 39, 48, 71, 75]). It generalizes straightforwardly to provide a reduced load approximation able to represent important additional features, to be discussed in detail later, such as alternative routing and trunk reservation. Second, the approximation is of some mathematical interest. The similarity between relations (1.7) and (1.11) is clear: At a deeper level we shall establish uniqueness of the Erlang fixed point by showing that it is a stationary point of a strictly convex function strikingly similar in form to the function (1.8). In Section 4 reduced load approximations will appear naturally as limit solutions for a number of models. A further reason for interest is that the relationship between the Erlang fixed point approximation and the underlying stochastic process precisely parallels the relationship in statistical mechanics between the mean field or Bragg–Williams approximation and the stochastic Ising model [8].

How accurate are such approximations? We shall approach this question from a variety of directions. In Section 3 we consider the behaviour of the Erlang fixed point under the limiting regime described earlier, where capacities and offered traffics are increased together. We find that under this limiting regime the error in the approximation (1.11) approaches zero. While this is reassuring, it is a rather crude test of accuracy. More refined tests distinguish between networks containing critically loaded links and those without and indicate the importance of diversity of routing. We study further the issue of diversity in Section 4.

1.4. *Symmetric networks.* The central limit theorems of Section 2 concern networks with increasing link capacities and loads, but fixed network topology. In Section 4 we consider a different form of limiting regime, where link capacities and loads are fixed or bounded and where the numbers of links and routes in the network increase to infinity.

Recall the informal idea underlying the Erlang fixed point approximation given in Section 1.3. The various streams of traffic making up the reduced load on a link are clearly not independent, but we might hope that the approxima-

tion will be more accurate the more diverse the collections of routes passing through any given link. In Section 4 we provide theoretical support for this suggestion by describing asymptotic results for networks exhibiting various symmetries. We consider sequences of networks with $\mathscr{R}$ and J increasing. Symmetry reductions will leave fixed the number of distinct equations in the set (1.9) and the number of distinct approximate forms (1.11), and we shall see that the exact loss probabilities (1.5) converge to the forms (1.11).

The combined simplifications of diverse routing and symmetry reduction allow a number of other important issues to be addressed. In particular, we are able to study networks operating under *alternative routing*, where a call which is blocked on a route may be allowed to try again on another route. We find that the Erlang fixed point again emerges as an asymptotically exact solution, but that it may now have multiple solutions. If alternative routes use more network resources than first-choice routes, then alternative routing can lead to instability and hysteresis, with several modes of behaviour possible. These deleterious consequences of alternative routing can be controlled by allowing a link to reject alternatively routed calls if the link occupancy is above a certain level. This method of giving priority at a link to certain traffic streams is known as *trunk reservation*. For a simple model of a network operating with trunk reservation and multiple alternatives, we find that the reduced load approximation emerges as an asymptotically exact solution, and we discuss the insights into network performance issues provided by the approximation. We also consider a very simple *least busy alternative* scheme, in which a call blocked on its first-choice route selects from among a list of alternative routes that one which is least busy.

Many of the formal limit theorems we review in Section 4 require rather restrictive symmetry assumptions, which often involve ignoring the graph structure naturally associated with a network. Taking proper account of graph structure leads to an interesting class of random graph problems. We describe some of the results in this area due to Hajek ([21, 22]), and use these results to establish the asymptotic form of the optimal admission and routing policy in a network allowing repacking.

1.5. *Lattice models.* In Section 5 we consider loss networks with fixed routing defined on lattices. The regular structure of such networks permits an exact analysis in a number of interesting cases.

One-dimensional networks are of some interest in their own right and provide a contrast to the various asymptotic link independence results of earlier sections. Suppose each route $r \in \mathscr{R}$ is a set of consecutive integers chosen from $\{1, 2, \ldots, J\}$ and that $C_j = C$, $j = 1, 2, \ldots, J$. One could imagine a cable on which are positioned $J + 1$ stations and that communication between two stations uses a fraction C^{-1} of the cable's capacity over the section of cable lying between the two stations. In Section 5 we consider two simple examples. If the offered traffic between two stations decays geometrically with the distance between the stations, then the number of free circuits $(m_1, m_2, \ldots, m_J)$ on the J links of the system is a Markov chain. If $C = 1$

and offered traffic between stations depends arbitrarily on the distance between the stations, then the vector $(m_1, m_2, \ldots, m_J)$ can be described in terms of an alternating renewal process.

There has been substantial progress in the field of interacting particle systems concerning the relationship between macroscopic phenomena, such as the existence of a phase transition, and the microscopic dynamical description of a system [47]. In Section 5 we illustrate connections with this rich field. In particular, we describe loss networks with fixed routing defined on a Bethe lattice (or tree) and on a two-dimensional lattice which exhibit phase transitions analogous to that of the Ising model of ferromagnetism [4]. As loss networks these models are rather special, but they serve to establish that long range influence can emerge from very simple networks involving only nearest-neighbour interactions.

1.6. *Optimization.* A major motivation for the development of loss network models is the hope that these models may help with practical problems concerning how routes should be chosen or capacity allocated. These problems are often quite difficult, owing to the complexity of the various interactions involved. For example, an increase in the offered traffic along a particular route will increase blocking at links along that route; this in turn will affect traffic carried along other routes through these links and also along routes which act as alternatives to these routes. Such knock-on effects will generally propagate throughout the entire network.

A further practical problem concerns the extent to which control can be decentralized. Over a period of time the form of the network or the demands placed on it may change and routings may need to adapt accordingly. A single node could perhaps control this, receiving information from everywhere in the network and making all decisions about routing. But this approach has drawbacks, particularly if links or nodes may fail. Could control be distributed over the nodes of the network, with computations and decisions made locally? A distributed control scheme should be able to react rapidly to a local disturbance at the point of the disturbance, with slower adjustments in the rest of the network as effects propagate outwards.

These practical problems are closely related to the interaction phenomena investigated rigorously in Sections 4 and 5. On the other hand, to be practically useful it is essential that any approach suggested be capable of dealing with asymmetric and irregular network structures, as well as the complications of alternative routing and trunk reservation. In Section 6 we discuss how the simplified analytical model provided by fixed point approximations can be used to provide substantial insight into both the pressing practical problems of routing and capacity allocation and the theoretically challenging issues of interaction and long range influence.

2. Large capacity networks. The distribution (1.2) is that of $|\mathscr{R}|$ independent Poisson random variables conditioned on a collection of linear inequality constraints (1.3). It is natural to look for a limit theorem as the capacities

C_j, $j = 1, 2, \ldots, J$, and the offered traffics ν_r, $r \in \mathscr{R}$, are increased together (with ratios C_j/ν_r held fixed). We would expect the distribution (1.2) to approach that of $|\mathscr{R}|$ independent normal random variables conditioned on a collection of linear inequality constraints. Limit results of this form are familiar when the linear inequalities are replaced by linear equalities and arise naturally in the analysis of contingency tables (see, e.g., [20]). The first step in establishing a result of this form is to find the centering term for the expected central limit theorem. This we do in Section 2.1, by formulating and solving a simple optimization problem which is of some intrinsic interest and which will arise later in a number of other contexts.

2.1. *Finding the mode and the dual problem.* Consider the problem of finding the most likely state n under the probability distribution (1.2). This is equivalent to maximizing

$$\sum_r (n_r \log \nu_r - \log n_r!)$$

over $n \in \mathscr{S}(C)$, a problem which is complicated by the discrete nature of the state space. To simplify things, replace $\log n!$ by $n \log n - n$ [recall that, by Stirling's formula, $\log n! = n \log n - n + O(\log n)$] and replace the integer vector n by a real vector x. The resulting problem is the following.

THE PRIMAL PROBLEM.

$$\begin{aligned} &\text{Maximize} && \sum_r (x_r \log \nu_r - x_r \log x_r + x_r) \\ &\text{subject to} && x \geq 0,\ Ax \leq C. \end{aligned} \tag{2.1}$$

Observe that the objective function in problem (2.1) is differentiable and strictly concave over the cone $x \geq 0$ and tends to $-\infty$ as $\|x\| \to \infty$, and the feasible region is a closed convex set. Hence a maximizing value of x exists and is unique, and can be found by Lagrangian methods [72]. Consider, then, the Lagrangian form

$$\begin{aligned} L(x, z; y) &= \sum_r (x_r \log \nu_r - x_r \log x_r + x_r) + \sum_j y_j \Big(C_j - \sum_r A_{jr} x_r - z_j \Big) \\ &= \sum_r x_r + \sum_r x_r \Big(\log \nu_r - \log x_r - \sum_j y_j A_{jr} \Big) + \sum_j y_j C_j - \sum_j y_j z_j, \end{aligned}$$

where $z = (z_1, z_2, \ldots, z_J)$ is the vector of slack variables $z = C - Ax$ and $y = (y_1, y_2, \ldots, y_J)$ is a vector of Lagrange multipliers. To maximize $L(x, z; y)$ over the cone $x, z \geq 0$, we require that $y \geq 0$, $y \cdot z = 0$ and, differentiating with respect to x_r,

$$\log \nu_r - \log x_r - \sum_j y_j A_{jr} = 0.$$

The maximizing x_r is then

$$\bar{x}_r(y) = \nu_r \exp\Big(- \sum_j y_j A_{jr} \Big) \tag{2.2}$$

and so

$$\begin{aligned} \max_{x, z \geq 0} L(x, z; y) &= \sum_r \bar{x}_r(y) + \sum_j y_j C_j \\ &= \sum_r \nu_r \exp\Big(- \sum_j y_j A_{jr} \Big) + \sum_j y_i C_j. \end{aligned}$$

Hence the Lagrangian dual to the primal problem is the following.

THE DUAL PROBLEM.

$$\begin{aligned} &\text{Minimize} \quad \sum_r \nu_r \exp\Big(- \sum_j y_j A_{jr} \Big) + \sum_j y_j C_j \\ &\text{subject to} \quad y \geq 0. \end{aligned} \tag{2.3}$$

We may solve the primal problem (2.1) by choosing values for the Lagrange multipliers $y = \bar{y}$ so that $\bar{x}(\bar{y}), \bar{y}$ are primal and dual feasible,

$$\bar{x}(\bar{y}) \geq 0, \qquad \bar{z} = C - A\bar{x}(\bar{y}) \geq 0, \qquad \bar{y} \geq 0, \tag{2.4}$$

and satisfy the complementary slackness conditions

$$\bar{y} \cdot \bar{z}(\bar{y}) = 0. \tag{2.5}$$

It is interesting to rewrite these conditions in terms of transformed variables

$$B_j = 1 - \exp(-y_j). \tag{2.6}$$

Under this transformation the conditions (2.4) and (2.5) on $\bar{y}$ become the following.

CONDITIONS ON B.

$$\sum_r A_{jr} \nu_r \prod_i (1 - B_i)^{A_{ir}} \begin{cases} = C_j & \text{if } B_j > 0, \\ \leq C_j & \text{if } B_j = 0, \end{cases} \tag{2.7}$$

$$B_1, B_2, \ldots, B_j \in [0, 1).$$

The convexity properties of the primal problem (2.1) imply that there exist Lagrange multipliers $\bar{y}$ satisfying (2.4) and (2.5), and hence that there exists $B = (B_1, B_2, \ldots, B_J)$ satisfying (2.7). Alternatively, observe that the objective function of the dual problem (2.3) is differentiable and convex over the cone $y \geq 0$ and tends to ∞ as $\|y\| \to \infty$. Hence an optimum $\bar{y}$ exists; differentiation of the dual objective function with respect to y_j establishes a one-to-one correspondence under the transformation (2.6) between optima of the dual problem and solutions B to conditions (2.7). Finally, observe that the mapping $y' \mapsto yA$ is one-to-one from the set $y \geq 0$ if A has rank J. The objective function of the

dual problem (2.3) is thus strictly convex if A has rank J. Hence the optimum $\bar{y}$ is unique if A has rank J.

We collect these observations in the following result [33].

2.8 THEOREM. *There exists a unique optimum to the primal problem* (2.1). *It can be expressed in the form*

$$x_r = \nu_r \prod_j (1 - B_j)^{A_{jr}}, \qquad r \in \mathscr{R},$$

where $B = (B_1, B_2, \ldots, B_J)$ *is any solution to the conditions* (2.7) *on* B. *There always exists a solution to the conditions on* B, *and it is unique if* A *has rank* J. *There is a bijection between solutions to the conditions* (2.7) *on* B *and optima of the dual problem* (2.3), *given by the transformation* (2.6).

Conditions (2.7) have a straightforward interpretation in terms of a continuous, or fluid, flow. Suppose that an offered flow of ν_r on route r is thinned by a factor $(1 - B_i)^{A_{ir}}$ on each link $i = 1, 2, \ldots, J$ so that a flow of

$$\nu_r \prod_i (1 - B_i)^{A_{ir}} \tag{2.9}$$

remains. Assume that one unit of flow on route r uses A_{jr} units of capacity at link j. Then conditions (2.7) state that at any link j for which $B_j > 0$ the total capacity of that link, C_j, must be completely utilized by the superposition over $r \in \mathscr{R}$ of the flows (2.9). Conversely, no thinning of flow is allowed at a link which is not full.

To illustrate the possibility of nonuniqueness consider the matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{2.10}$$

The matrix A has rank 3, and if (B_1, B_2, B_3, B_4) is a solution to condition (2.7) with all entries positive, then so is

$$\big(1 - d(1 - B_1), 1 - d^{-1}(1 - B_2), 1 - d(1 - B_3), 1 - d^{-1}(1 - B_4)\big),$$

for d close enough to 1. Flow can be thinned at either an odd or an even link, and only products of the form $(1 - B_{\text{odd}})(1 - B_{\text{even}})$ are fixed. For a further illustration see Example 2.32.

Networks exhibiting nonuniqueness of the vector $B = (B_1, B_2, \ldots, B_J)$ arise very naturally, and for this reason we are not prepared to exclude them by assuming, say, that A has rank J. We do, however, need to take some additional care with such networks. The set of optima to the dual problem (2.3) is convex, and hence there exists an optimum of maximal support—that is, an optimum y such that if y' is any other optimum, then $y'_j > 0 \Rightarrow y_j > 0$. Thus there exists a solution B to the conditions (2.7) which is of maximal support. For such a solution B let its support be $\mathscr{B} = \{j\colon B_j > 0\}$ and let $A_{\mathscr{B}} = (A_{jr},\ j \in \mathscr{B},\ r \in \mathscr{R})$ be the restriction of the matrix A to the set $\mathscr{B}$.

The representation (2.2) of the unique optimum to the primal problem shows that nonuniqueness of a solution to the conditions (2.7) is equivalent to linear dependence between the rows of $A_{\mathscr{B}}$. Recall the definition of the slack variables $z = C - Ax$, where henceforth $x = (x_r, r \in \mathscr{R})$ is the optimum identified in Theorem 2.8. For $j \in \mathscr{B}$ the slack variable $z_j = 0$. But it is possible that $z_j = 0$ for $j \notin \mathscr{B}$: in the fluid analogy this would correspond to the utilization of link j exactly matching the capacity of link j without the need for any thinning at link j. Define the set of such links $\mathscr{O} = \{j: z_j = 0, j \notin \mathscr{B}\}$ and the corresponding matrix $A_{\mathscr{O}} = (A_{jr}, j \in \mathscr{O}, r \in \mathscr{R})$. Observe that no row of $A_{\mathscr{O}}$ is in the space spanned by the rows of $A_{\mathscr{B}}$, by the maximality of $\mathscr{B}$ and the representation (2.2). Finally, let $\mathscr{F} = \{j: j \notin \mathscr{B} \cup \mathscr{O}\}$. We interpret $\mathscr{B}$ as the set of overloaded (or busy) links, $\mathscr{F}$ as the set of underloaded (or free) links and $\mathscr{O}$ as the set of critically loaded links.

Observe that arbitrarily small perturbations of ν or C are sufficient to render the set $\mathscr{O}$ empty. We shall deal explicitly with the set $\mathscr{O}$, rather than just assume it empty, since in certain circumstances critical loadings may be important. For example, the dimensioning procedures for a network may well lead to critical loadings at some links. We study such links in more detail in Section 2.3.

2.2. *A central limit theorem.* We review the quantities we have defined in Section 2.1. The vector $B = (B_1, B_2, \ldots, B_J)$ is a solution of maximal support to the conditions (2.7) on B, the link subsets $\mathscr{B}$, $\mathscr{O}$ and $\mathscr{F}$ have been identified by the solution B, and the vector $x = (x_r, r \in \mathscr{R})$ is the unique optimum identified in Theorem 2.8.

Now consider a sequence of networks of the form described in Section 1.2, indexed by N. In the Nth network replace C_j by $C_j(N)$ and ν_r by $\nu_r(N)$, where

$$\frac{1}{N}\nu_r(N) \to \nu_r, \qquad \frac{1}{N}C_j(N) \to C_j \tag{2.11}$$

and all limits are as $N \to \infty$. Let $n_r(N)$ be the number of calls in progress using route r: we are interested in the stationary distribution of $n(N) = (n_r(N), r \in \mathscr{R})$. Let $B(N) = (B_1(N), B_2(N), \ldots, B_J(N))$ solve the conditions (2.7) on B with ν_r replaced by $\nu_r(N)$ and with C_j replaced by either $C_j(N)$ for $j \in \mathscr{B}$ or by infinity for $j \notin \mathscr{B}$. Let

$$x_r(N) = \nu_r(N) \prod_j \big(1 - B_j(N)\big)^{A_{jr}}.$$

From the representation of x as the unique optimum of the primal problem (2.1), it follows that $x_r(N)/N \to x_r$, $r \in \mathscr{R}$. Some of our results will need a more precise comparison between offered traffics and capacities than that provided by (2.11): in particular, we shall sometimes require that

$$C_j(N) - \sum_r A_{jr}x_r(N) = o(N^{1/2}), \qquad j \in \mathscr{B}, \tag{2.12}$$

$$C_j(N) - \sum_r A_{jr}x_r(N) = \alpha_j N^{1/2} + o(N^{1/2}), \qquad j \in \mathscr{O}. \tag{2.13}$$

The vector $\alpha_{\mathscr{O}} = (\alpha_j,\ j \in \mathscr{O})$ is thus a corrected second-order measure of capacity at the links in $\mathscr{O}$. Let

$$u_r(N) = N^{-1/2}(n_r(N) - x_r(N)). \tag{2.14}$$

Thus $u_r(N)$ is a normalized version of the number of calls in progress using route r, centred on $x_r(N)$.

Let

$$\mathscr{S}(N) = \{n \in \mathbb{Z}_+^{\mathscr{R}}: An \leq C(N)\}$$

and, for a state $n(N) \in \mathscr{S}(N)$, let

$$m_j(N) = C_j(N) - \sum_j A_{jr} n_r(N). \tag{2.15}$$

Thus the vector $m(N) = (m_j(N),\ j = 1, 2, \ldots, J)$ describes the number of free circuits on each link. It is important to notice the variety of constraints there may be on the values taken by $m(N)$, even when A has full rank. Let

$$\mathscr{M}(N) = \{m: \exists\ n \in \mathscr{S}(N) \text{ with } An + m = C(N)\},$$

the set of possible vectors $m(N)$. For example, consider the matrix

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

Let $C_j(N) = 2N$, for $N = 1, 2, \ldots$, $j = 1, 2, 3$. Then elementary calculations show that

$$\mathscr{M}(N) \to \{(m_1, m_2, m_3): m_j \geq 0,\ m_1 + m_2 + m_3 \text{ even}\}. \tag{2.16}$$

If, instead, $C_j(N) = 2N + 1$, for $j = 1, 2, 3$, then

$$\mathscr{M}(N) \to \{(m_1, m_2, m_3): m_j \geq 0,\ m_1 + m_2 + m_3 \text{ odd}\}. \tag{2.17}$$

If $C_j(N) = N$, for $j = 1, 2, 3$ then the sequence $\mathscr{M}(N)$, $N = 1, 2, \ldots$, oscillates rather than converges. Simplifications may occur if we restrict attention to a subset of the components of m. For $\mathscr{B} \subset \{1, 2, \ldots, J\}$, $m(N) \in \mathscr{M}(N)$, let $m_{\mathscr{B}}(N) = (m_j(N),\ j \in \mathscr{B})$ and let

$$\mathscr{M}_{\mathscr{B}}(N) = \{m_{\mathscr{B}}(N): m(N) \in \mathscr{M}(N)\}.$$

In the preceding example if $\mathscr{B} = \{1, 2\}$, then for each of the three choices $C_j(N) = 2N$, $2N + 1$ or N we have that $\mathscr{M}_{\mathscr{B}}(N) \to \mathbb{Z}_+^2$.

Construct a vector $u = (u_r,\ r \in \mathscr{R})$ by conditioning independent normal random variables, $u_r \sim N(0, x_r)$, $r \in \mathscr{R}$, on $A_{\mathscr{B}} u = 0$ and $A_{\mathscr{O}} u \leq \alpha_{\mathscr{O}}$. Construct an independent vector $m_{\mathscr{B}}$ with distribution

$$\mathbb{P}\{m_{\mathscr{B}}\} = g^{-1} \prod_{j \in \mathscr{B}} B_j (1 - B_j)^{m_j}, \qquad m_{\mathscr{B}} \in \mathscr{M}_{\mathscr{B}}, \tag{2.18}$$

where g is a normalizing constant, chosen so that the distribution sums to unity. Let " $\Rightarrow$ " denote convergence in distribution. The following results are established in [27] and [33].

2.19 THEOREM. *Under the limiting regime* (2.11)–(2.13),

$$u(N) \Rightarrow u.$$

Further, moments of $u(N)$ *converge to the corresponding moments of* u. *If* $\mathscr{M}_{\mathscr{B}}(N) \to \mathscr{M}_{\mathscr{B}}$, *then*

$$(m_{\mathscr{B}}(N), u(N)) \Rightarrow (m_{\mathscr{B}}, u)$$

and moments converge also.

2.20 REMARK. If $\mathscr{M}_{\mathscr{B}} = \mathbb{Z}_+^{\mathscr{B}}$, then the normalizing constant g in (2.18) is unity, and $m_{\mathscr{B}} = (m_j, j \in \mathscr{B})$ are independent random variables, m_j geometrically distributed with parameter $1 - B_j$. A simple sufficient condition for this conclusion is that columns of the matrix $A_{\mathscr{B}}$ include the columns of $I_{\mathscr{B}}$, the $|\mathscr{B}|$-dimensional identity matrix—that is, there is some single link traffic on each link $j \in \mathscr{B}$. In [33] it is shown that a necessary and sufficient condition is that $A_{\mathscr{B}}$ have an integer right inverse, that is, there exists a matrix D with elements from $\mathbb{Z}$ such that $A_{\mathscr{B}}D = I_{\mathscr{B}}$. In general, the form of the set $\mathscr{M}_{\mathscr{B}}$ captures any parity conditions that obtain, such as those illustrated by the limit sets (2.16) and (2.17).

2.3. *Critically loaded links.* For links in $\mathscr{B}$ the number of free circuits is $O(1)$ and well-described by Theorem 2.19 and the distribution (2.18). For links in $\mathscr{F}$ the number of free circuits is $O(N)$: these links have little effect upon loss probabilities or the utilization of other links. For links in $\mathscr{C}$ the number of free circuits is $O(N^{1/2})$: it is these links we now consider. Let

$$v_j(N) = N^{-1/2} m_j(N). \tag{2.21}$$

Thus $v_j(N)$ is a normalized version of the number of free circuits on link j. From (2.14), (2.15) and (2.21) we have that

$$v_j(N) = N^{-1/2}\Big(C_j(N) - \sum_r A_{jr} x_r(N)\Big) - \sum_r A_{jr} u_r(N).$$

Thus, from (2.14),

$$v_j(N) = \alpha_j - \sum_r A_{jr} u_r(N) + o(1), \qquad j \in \mathscr{C}. \tag{2.22}$$

Initially, let u be a multivariate normal random vector, $u \sim N(0, \Sigma)$, where $\Sigma = \operatorname{diag}(x_r)_r$. The distribution of $A_{\mathscr{C}}u$ conditional on the event $A_{\mathscr{B}}u = 0$ is then multivariate normal $N(0, \Theta)$, where Θ is the $|\mathscr{C}|$-dimensional covariance matrix

$$\Theta = A_{\mathscr{C}}\Sigma A_{\mathscr{C}}^T - A_{\mathscr{C}}\Sigma A_{\mathscr{B}}^T\big(A_{\mathscr{B}}\Sigma A_{\mathscr{B}}^T\big)^{-1} A_{\mathscr{B}}\Sigma A_{\mathscr{C}}^T. \tag{2.23}$$

Here $(A_{\mathscr{B}}\Sigma A_{\mathscr{B}}^T)^{-1}$ denotes the generalized inverse [60] in the case where $A_{\mathscr{B}}$ has deficient rank. Thus, from the representation (2.22) and Theorem 2.19, we deduce the following result ([27]; see also [13, 76]).

2.24 COROLLARY. *Under the limiting regime* (2.11)–(2.13), *the distribution of* $v_{\mathscr{O}}(N) = (v_j(N),\ j \in \mathscr{O})$ *converges weakly to the distribution of a vector* $v_{\mathscr{O}} = (v_j,\ j \in \mathscr{O})$ *formed by conditioning a multivariate normal random vector* $v_{\mathscr{O}} \sim N(\alpha_{\mathscr{O}}, \Theta)$ *on* $v_{\mathscr{O}} \geq 0$.

2.25 REMARK. We have seen in Theorem 2.19 and Remark 2.20 that the numbers of free circuits on links in $\mathscr{B}$ are asymptotically independent, or independent modulo the parity conditions captured by the limit set $\mathscr{M}_{\mathscr{B}}$. Corollary 2.24 shows that the numbers of free circuits on critically loaded links are *not* independent. Nevertheless, network structure may lead to approximate independence. Consider, for example, the case where $|\mathscr{O}| = J$, so that all links in the network are critically loaded. Then

$$\Theta_{jk} = \sum_r A_{jr} A_{kr} \nu_r,$$

and so Θ_{jk} is a weighted measure of the volume of traffic going through *both* links j and k. If routing within the network is very diverse, so that Θ_{jk} is small in comparison with Θ_{jj} and Θ_{kk}, for all j and k, then Θ will be nearly diagonal and the components of the conditioned vector $v_{\mathscr{O}}$ will be approximately independent. More generally, in networks containing both busy and critically loaded links, the covariance matrix (2.23) can be used to assess the extent of dependence between the numbers of free circuits on links in $\mathscr{O}$. We return to the study of diverse routing in Section 4.

2.4. *Loss probabilities.* Let $L_r(N)$ be the stationary probability that a call requesting route r is lost. Then, by Little's formula ([61], page 102),

$$(1 - L_r(N))\nu_r(N) = \mathbb{E}(n_r(N)), \qquad r \in \mathscr{R}.$$

Thus, from the definition (2.14),

$$\begin{aligned} 1 - L_r(N) &= \frac{x_r(N)}{\nu_r(N)} + N^{1/2} \frac{\mathbb{E}(u_r(N))}{\nu_r(N)} \\ &= \prod_j \left(1 - B_j(N)\right)^{A_{jr}} + O(N^{-1/2}). \end{aligned} \tag{2.26}$$

From the convergence of $x_r(N)$ to x_r and a consideration of subsequences satisfying (2.12) and (2.13), we deduce the following result.

2.27 COROLLARY. *Under the limiting regime* (2.11),

$$1 - L_r(N) = \prod_j \left(1 - B_j\right)^{A_{jr}} + o(1).$$

2.28 REMARK. Corollary 2.27 is just statement (1.6) of the Introduction: asymptotically it is *as if* requests for circuits are granted or denied independently. Observe that Corollary 2.27 uses only that the error term in expression (2.26) is $o(1)$: in [27] a refinement is obtained. There it is shown that the vector

u, formed by conditioning a multivariate normal random vector $u \sim N(0, \Sigma)$ on $A_{\mathscr{B}} u = 0$ and $A_{\mathscr{O}} u \leq \alpha_{\mathscr{O}}$, satisfies

$$\mathbb{E}\left(\frac{u_r}{x_r}\right) = \sum_j \beta_j A_{jr}, \qquad r \in \mathscr{R}, \tag{2.29}$$

for some $\beta_1, \beta_2, \ldots, \beta_J$. Let

$$b_j(N) = B_j(N) - N^{-1/2}\beta_j\big(1 - B_j(N)\big).$$

2.30 THEOREM. *Under the limiting regime* (2.11)–(2.13),

$$1 - L_r(N) = \prod_j \big(1 - b_j(N)\big)^{A_{jr}} + o(N^{-1/2}), \qquad r \in \mathscr{R}.$$

2.31 REMARK. If there is a route which uses just a single circuit from link j, then we can deduce from Theorem 2.30 that the probability that link j is full is $b_j(N) + o(N^{-1/2})$, since this is the loss probability for that route. In any event we have refined the implication of Corollary 2.27, that limiting loss probabilities are *as if* links block independently.

It is possible to obtain still more precise error bounds on $L_r(N)$, provided we have more precise information about capacities than is contained in relations (2.12) and (2.13). We illustrate this point with a brief account of an example considered in detail by Mitra [52].

2.32 EXAMPLE. There are $K + 1$ links, and

$$\mathscr{R} = \{\{j, K+1\} : j = 1, 2, \ldots, K\}.$$

Capacities and offered traffics in network N satisfy the limiting regime (2.11) and, in addition,

$$\sum_{j=1}^{K} C_j(N) - C_{K+1}(N) = I,$$

where $I \geq 1$, for each value of N. Thus there is not quite sufficient capacity on the common link $K + 1$ to be able to deal with all the traffic that could be carried on links $1, 2, \ldots, K$. Assume $\nu_{\{j,K+1\}} > C_j$, and let $1 - B_j = C_j/\nu_{\{j,K+1\}}$. Then, from Theorem 2.19, the limiting distribution for the number of idle circuits on links $1, 2, \ldots, K$ is

$$P\{m_1, m_2, \ldots, m_K\} = g^{-1} \sum_{j=1}^{K} B_j(1 - B_j)^{m_j}, \quad \text{for } \sum_{j=1}^{K} m_j \geq I. \tag{2.33}$$

The normalizing constant is thus

$$g = 1 - h \prod_{j=1}^{K} B_j, \quad \text{where } h = \sum_{n \in \mathscr{S}} \prod_{j=1}^{K} (1 - B_j)^{n_j}$$

and

$$\mathscr{S} = \left\{ (n_1, n_2, \ldots, n_{K+1}) \in \mathbb{Z}_+^{K+1} : \sum_{j=1}^{K+1} n_j = I \right\}.$$

The reader may recognize h as the normalizing constant for a closed queueing network with I customers and $K+1$ single server queues ([31, 68]), and part of the attraction of this example is the relationship it exhibits between loss networks and queueing networks. The stationary distribution for the queueing network is

$$P(n_1, n_2, \ldots, n_{K+1}) = h^{-1} \prod_{j=1}^{K} (1 - B_j)^{n_j}, \quad \text{over } \mathscr{S}. \tag{2.34}$$

An elementary calculation now shows that

$$\mathbb{E}m_j = g^{-1}\left[\frac{1 - B_j}{B_j} - h\left(\prod_{j=1}^{K} B_j \right) \mathbb{E}n_j \right], \tag{2.35}$$

where $\mathbb{E}m_j, \mathbb{E}n_j$ are calculated with respect to the distributions (2.33) and (2.34), respectively. From Theorem 2.19, moments of $m_j(N)$ converge to moments of m_j, and hence expression (2.35) gives $\mathbb{E}(m_j(N))$ to within an error of $o(1)$. Thus,

$$1 - L_{\{j, K+1\}}(N) = \frac{\left[C_j(N) - \mathbb{E}m_j \right]}{\nu_{\{j, K+1\}}(N)} + o(N^{-1}),$$

and so we have the loss probability to within an error of $o(N^{-1})$. Mitra [52] gives a complete analysis of this example, establishing a full expansion for loss probabilities in inverse powers of N, without the restriction that $\nu_{\{j, K+1\}} > C_j$. The analysis provides an efficient recursive formula for calculating the general term of the expansion and provides tight upper and lower bounds.

2.5. *Time evolution.* Until now we have been concerned entirely with the distribution of the stochastic process $(n(t), t \geq 0)$ as a fixed point in time. Here we provide a brief heuristic discussion of the time evolution of the process, intended to illuminate the results of Section 2 and to display a range of open questions. Observe that all of the results of Section 2 have been derived from the exact stationary distribution (1.2): progress with these open questions might provide techniques able to handle networks where a product-form solution is not available. Throughout assume the limiting regime (2.11)–(2.13).

Write $n(N, t), m(N, t)$ for $n(N), m(N)$, to emphasize their dependence on the time parameter t. Observe that $n_r(N, t)$ moves to $n_r(t) \pm 1$ at rate $O(N)$, and so, if the network is stationary, we would expect it to traverse a distance $O(N^{1/2})$ in time of order $O(1)$. A relaxation time of order $O(1)$ also seems reasonable since call holding periods are of order $O(1)$. Put more precisely, we would expect to obtain a nondegenerate limit process $(u(t), t \geq 0)$ from the

sequence

$$N^{-1/2}[n(N,t) - x(N)], \qquad N \to \infty. \tag{2.36}$$

If holding times are exponentially distributed, then we would expect $(u(t), t \geq 0)$ to be an Ornstein–Uhlenbeck diffusion within the region

$$\{u: A_{\mathscr{B}}u = 0,\ A_{\mathscr{O}}u \leq \alpha_{\mathscr{O}}\},$$

reflected at the boundary $A_{\mathscr{O}}u = \alpha_{\mathscr{O}}$. The reflection must be such that when stationary the process $(u(t), t \geq 0)$ is reversible, with stationary distribution the conditioned multivariate normal distribution identified in Section 2.2.

Consider now links $j \in \mathscr{B}$, and assume that $\mathscr{M}_{\mathscr{B}}(N) \to \mathscr{M}_{\mathscr{B}}$. By Theorem 2.19 the fluctuations of the number of spare circuits $m_j(N,t)$ are of order $O(1)$. But $m_j(N,t)$ moves to $m_j(N,t) \pm 1$ at rate $O(N)$, indicating that the relaxation time of $m_j(N,t)$ is of order $O(N^{-1})$. We would then expect to obtain a nondegenerate limit process from the sequence

$$m_{\mathscr{B}}(N, tN), \qquad N \to \infty. \tag{2.37}$$

This can be checked readily in special cases such as Example 2.32, where $A_{\mathscr{B}}$ has rank $|\mathscr{R}|$.

Note especially the different time-scale normalizations appearing in (2.36) and (2.37): they suggest that $m(N,t)$ might quickly reach a quasistationary distribution even when the network is not stationary. More precisely we present the following conjecture of Kurtz [44] and Hunt [26]. Let $\lambda = (\lambda_r, r \in \mathscr{R})$ and let $\mathscr{D} = \{j: \Sigma_r A_{jr}\lambda_r = C_j\}$. Consider a Markov chain $m_{\mathscr{D}} = (m_j, j \in \mathscr{D})$ with state space $\mathbb{Z}_+^{|\mathscr{D}|}$ and nonzero transition rates

$$q(m_{\mathscr{D}}, m_{\mathscr{D}} + A_{\mathscr{D}}e_r) = \lambda, \qquad q(m_{\mathscr{D}} + A_{\mathscr{D}}e_r, m_{\mathscr{D}}) = \nu_r.$$

Let $\pi_r(\lambda)$ be the limiting probability under these transition rates that $m_j \geq A_{jr}$ for all $j \in \mathscr{D}$. Consider now the sequence

$$N^{-1}n(N,t), \qquad N \to \infty, \tag{2.38}$$

with initial condition $N^{-1}n(N,0) \to X(0)$. The conjecture is that the sequence (2.38) converges to a limit $(X(t), t \geq 0)$ which is determined as a unique solution of the integral equations

$$X_r(t) = X_r(0) + \int_0^t [\nu_r \pi_r(X(s)) - X_r(s)]\, ds, \qquad r \in \mathscr{R}. \tag{2.39}$$

The form of this equation is easy to explain. From the vector $X(t)$ determine the set $\mathscr{D}$ of links which are nearly full. From the Markov chain describing the number of free circuits on links in $\mathscr{D}$, calculate the acceptance probability $\pi_r(X(t))$ on route r: the net drift upwards in $X_r(t)$ is then just the integrand of (2.39). Note the interesting interplay: The current vector $X(t)$ determines the transition rates of the Markov chain and hence the limiting probabilities $\pi_r(X(t))$; conversely, these limiting probabilities determine the rate of change of the vector $X(t)$. Hunt [26] has shown that a separation of time scales allows the limiting probability $\pi(\cdot)$ to be used to give the rates of change of $X(t)$, but has also shown that in more complex networks with various forms of routing

control it is possible for interesting and nonuniquely determined behaviour to occur following points in time when the set $\mathscr{D}$ changes. It remains open whether such behaviour can occur in networks with fixed routing.

3. Fixed point approximations. It will be convenient to consider the following generalization of the Erlang fixed point, (1.9) and (1.10). Let $E_1, E_2, \ldots, E_J$ be a solution to the equations

$$E_j = E(\rho_j, C_j), \qquad j = 1, 2, \ldots, J, \tag{3.1}$$

where

$$\rho_j = (1 - E_j)^{-1} \sum_r A_{jr} \nu_r \prod_i (1 - E_i)^{A_{ir}} \tag{3.2}$$

and the function E is Erlang's formula (1.1). Then an approximation for the loss probability on route r is given by

$$1 - L_r \cong \prod_j (1 - E_j)^{A_{jr}}. \tag{3.3}$$

Again the underlying idea is simple to explain. If a request for a circuit from link i is denied with probability E_i and if we make the approximation that all such requests are granted or denied independently, then the traffic offered to link j will comprise independent Poisson streams and the level of carried traffic on link j will be $\sum_r A_{jr}\nu_r \prod_i (1 - E_i)^{A_{ir}}$. Equations (3.1) and (3.2) simply state that the blocking probability on link j should be consistent with this level of carried traffic, under the Erlang model of a single link offered Poisson single-circuit traffic. Observe that (3.1) and (3.2) reduce to (1.9) and (1.10) when A is a 0–1 matrix.

Under the limiting regime considered in Section 2, we have seen that the vector $B = (B_1, B_2, \ldots, B_J)$ of blocking probabilities emerges as a solution to the dual problem (2.3) or, equivalently, as a solution to the conditions (2.7) on B. In Section 3.1 we show that a natural relaxation of the dual problem and of the conditions on B provides us with the Erlang fixed point equations (3.1) and (3.2). This characterization of the Erlang fixed point provides a simple proof of its uniqueness under fixed routing. In Section 3.2 we consider the accuracy of the approximation (3.3) under the limiting regime of Section 2. In Section 3.3 we take the opportunity to define a loss network with repacking, and we illustrate the application of the Erlang fixed point. Finally, in Section 3.4 we discuss the Erlang fixed point as a special case of a general reduced load approximation. This prepares the way for Section 4, where reduced load approximations will appear as limit solutions for a number of models.

3.1. *Uniqueness and the revised dual.* Our starting point is the dual problem (2.3); recall that this arose as the dual of the problem of finding the mode of the stationary distribution (1.2) for a loss network with fixed routing. We shall see that by amending the final term in the objective function of (2.3) we can establish the connection with the Erlang fixed point.

Define a utilization function $U(y, C)$ by the implicit relation

$$U(-\log(1 - E(\nu, C)), C) = \nu(1 - E(\nu, C)). \tag{3.4}$$

Observe that as ν increases from 0 to ∞ the first argument of U increases from 0 to ∞ and so this implicit relation defines a function $U: \mathbb{R}_+ \times \mathbb{Z}_+ \to \mathbb{R}_+$. Indeed, under the Erlang model of a single link (Section 1.1) the quantity $U(y, C)$ is just the mean number of circuits in use (the *utilization*) when the blocking probability is $1 - \exp(-y)$. Thus $U(y, C)$ is a strictly increasing function of y.

Consider now the following problem.

THE REVISED DUAL PROBLEM.

$$\begin{aligned} &\text{Minimize} \quad \sum_r \nu_r \exp\Big(-\sum_j y_j A_{jr}\Big) + \sum_j \int_0^{y_j} U(z, C_j)\, dz \\ &\text{subject to} \quad y \geq 0. \end{aligned} \tag{3.5}$$

Since $U(y, C)$ is a strictly increasing function of y, $\int_0^y U(z, C)\, dz$ is a strictly convex function of y. Hence the objective function of problem (3.5) is strictly convex: it thus has a unique minimum. The objective function is also differentiable, hence the stationarity conditions

$$\sum_r A_{jr}\nu_r \exp\Big(-\sum_i y_i A_{ir}\Big) = U(y_j, C_j), \qquad j = 1, 2, \ldots, J, \tag{3.6}$$

obtained by differentiating the objective function with respect to y_j, $j = 1, 2, \ldots, J$, locate a unique vector $y \geq 0$. Now suppose that $(E_1, E_2, \ldots, E_J) \in [0, 1]^J$ is a solution to (3.1) and (3.2). Under the one-to-one transformation $E_j = 1 - \exp(-y_j)$ and using the definition (3.4), these equations become precisely (3.6). Hence we obtain the following result [33].

3.7 THEOREM. *Equations* (3.1) *and* (3.2) *have a unique solution,* $(E_1, E_2, \ldots, E_J)$, *given in terms of the optimum* y *of the revised dual problem* (3.5) *by* $E_j = 1 - \exp(-y_j)$.

3.8 REMARK. The conditions (2.7) insist that the carried traffic on a link must equal capacity before the blocking probability on that link can be positive. Conditions (3.6) are a natural relaxation: as carried traffic approaches capacity, blocking increases in a manner corresponding to Erlang's formula (1.1). Similarly, replacing the function U by a function

$$U_f(z, C) = C, \qquad z > 0,$$

reduces the revised problem (3.5) to the dual problem (2.3). The utilization function U_f would be natural for fluid flow, where if there is any blocking, then all capacity is in use. For C large there is not much difference between U

and U_f: we can show [33] that

$$U(z,C) = C - (e^z - 1)^{-1} + o(1), \quad \text{as } C \to \infty, \tag{3.9}$$

uniformly over z in any compact subset of $(0,\infty)$.

3.10 REMARK. Whittle [74] has shown that the revised dual problem (3.5), with $U(z,C)$ replaced by

$$U_g(z,C) = \left(\sum_{n=0}^{C} e^{nz} \right)^{-1} \left(\sum_{n=0}^{C} ne^{nz} \right),$$

emerges naturally from a saddlepoint approximation to a contour integral representation of the expected value of n_r. Observe that $U_g(z,C)$ is the utilization of a link if the number of free circuits on the link has a geometric distribution with parameter $1 - B = \exp(-z)$, truncated to the range $\{0,1,\ldots,C\}$. Thus U_g also corresponds to the asymptotic geometric characterization (2.18) established in Theorem 2.19. Recall that the definition (3.4) of the function U corresponds to the Erlang model of a single link, where the number of circuits in use has a truncated Poisson distribution. We can thus view the utilization functions U_f, U_g and U as determined by increasingly refined models of the behaviour of a single link. The saddlepoint approximation and the Erlang fixed point approximation behave similarly under the limiting regime of Section 2, to be considered further in Section 3.2; we shall see that reduced load approximations such as the Erlang fixed point are preferred under the asymptotic regimes to be considered in Section 4, where the capacity C of a link is held fixed.

3.2. *Limiting loss probabilities.* We consider now a sequence of networks, with ν_r, C_j replaced by $\nu_r(N), C_j(N)$, respectively, satisfying the limiting regime (2.11). For network N the revised dual objective function satisfies, using relation (3.9),

$$\sum_r \nu_r(N) \exp\left(- \sum_j y_j A_{jr} \right) + \sum_j \int_0^{y_j} U\big(z, C_j(N)\big)\, dz$$

$$= N\left\{ \sum_r \nu_r \exp\left(- \sum_j y_j A_{jr} \right) + \sum_j y_j C_j \right\} + o(N), \quad \text{as } N \to \infty,$$

uniformly over y in any compact subset of $(0,\infty)^J$. Thus, under the limiting regime (2.11), the revised dual objective function approaches a scaled version of the original dual objective function. Let $(E_1(N), E_2(N), \ldots, E_J(N))$ be the Erlang fixed point for the Nth network. The following result [33] then follows from Corollary 2.27.

3.11 THEOREM. *Under the limiting regime* (2.11),

$$1 - L_r(N) = \prod_j \big(1 - E_j(N)\big)^{A_{jr}} + o(1), \qquad r \in \mathscr{R}.$$

We can establish more when the network contains no critically loaded links.

3.12 THEOREM. *Under the limiting regime* (2.11) *and* (2.12), *and provided* $\mathscr{O}$ *is empty*,

$$1 - L_r(N) = \prod_j \big(1 - E_j(N)\big)^{A_{jr}} + o(N^{-1/2}), \qquad r \in \mathscr{R}.$$

PROOF. If the set $\mathscr{O}$ is empty, then the vector u of Theorem 2.19, formed by conditioning $u \sim N(0, \Sigma)$ on $A_{\mathscr{B}} u = 0$, has distribution $u \sim N(0, \Sigma - \Sigma A_{\mathscr{B}}^T (A_{\mathscr{B}} \Sigma A_{\mathscr{B}}^T)^{-1} A_{\mathscr{B}} \Sigma)$. In particular, $\mathbb{E}(u_r) = 0$, $r \in \mathscr{R}$, and so a solution to (2.29) is $\beta_j = 0$, $j = 1, 2, \ldots, J$. Thus, from Theorem 2.30,

$$1 - L_r(N) = \prod_j \big(1 - B_j(N)\big)^{A_{jr}} + o(N^{-1/2}), \qquad r \in \mathscr{R}.$$

The result now follows, since by the characterizations of $(B_j(N),\ j = 1, 2, \ldots, J)$ and $(E_j(N),\ j = 1, 2, \ldots, J)$ in terms of solutions to equations of the respective forms (2.7) and (3.6),

$$\prod_j \big(1 - B_j(N)\big)^{A_{jr}} = \prod_j \big(1 - E_j(N)\big)^{A_{jr}} + o(N^{-1/2}), \qquad r \in \mathscr{R}. \qquad \square$$

3.13 REMARK. The restriction in Theorem 3.12, that $\mathscr{O}$ be empty, is important. In [27] an example is given which shows that the Erlang fixed point approximation may *not* be accurate to order $o(N^{-1/2})$ in a network containing critically loaded links. In view of Theorem 2.30, this is an interesting observation. In networks containing critically loaded links, there is a level of accuracy at which acceptance probabilities have a limiting product-form decomposition, and yet at this same level of accuracy the Erlang fixed point approximation fails. This observation provides insight into the informal argument leading to the Erlang fixed point: at the $N^{-1/2}$ level of refinement, mean carried traffics *are* given by multiplicative thinnings; however, the rate of arrivals at a link varies sufficiently to invalidate the Poisson approximation, and hence the use of Erlang's formula.

Zachary [76] has developed a refinement to the Erlang fixed point which, provided $\mathscr{O}$ is empty, reduces the error in estimating loss probabilities to order $o(N^{-3/2+\varepsilon})$.

3.3. *Repacking*. We have interpreted our earlier results primarily in terms of a loss network with fixed routing, as introduced in Section 1.2. Here we show that a network operating under a certain form of dynamic routing, termed *repacking*, is equivalent to a transformed network operating under

fixed routing: in particular, the central limit theorem of Section 2 and our results on the Erlang fixed point apply to networks with repacking.

Let the label q of a call arriving at the network identify not a single route, but a set of routes, any of which could serve the call. Set $D_{qr} = 1$ if a call labelled q can be carried on the route r, and set $D_{qr} = 0$ otherwise. This defines a 0–1 matrix $D = (D_{qr},\ q \in \mathscr{Q},\ r \in \mathscr{R})$. As before assume that a call carried on route r uses A_{jr} circuits from link j and that link j comprises C_j circuits. Assume now, however, that a call labelled q can be shifted or shared between the routes $\{r:\ D_{qr} = 1\}$ while the call is in progress and that an arriving call is accepted provided the calls already in progress can be repacked to provide space for the arriving call. More formally, suppose that calls labelled q arrive at the network as a Poisson stream of rate ν_q, and let n_q be the number of calls in progress labelled q. Then an arriving call is accepted provided the vector $n = (n_q,\ q \in \mathscr{Q})$ remains within the set

$$\mathscr{S} = \left\{n \in \mathbb{Z}_+^{\mathscr{Q}}\colon \exists\ x \in \mathbb{R}_+^{\mathscr{R}} \text{ such that } Dx = n,\ Ax \leq C\right\};$$

otherwise, the call is lost. Each call holding period is independent of earlier arrival times and holding periods and has unit mean.

Let

$$\mathscr{T} = \{x \geq 0\colon Ax \leq C\}.$$

Thus $\mathscr{T}$ is the intersection of the half-spaces $\{x\colon (Ax)_j \leq C_j\}$, $j = 1, 2, \ldots, J$, and the nonnegative orthant. With no loss of generality assume no column of A is null; hence the set $\mathscr{T}$ is bounded, and is the convex hull of a finite number of extreme points. Hence $D\mathscr{T}$ is the convex hull of a finite number of extreme points, or, equivalently, the bounded intersection of a finite set of half-spaces. But $\mathscr{S} = D\mathscr{T} \cap \mathbb{Z}_+^{\mathscr{Q}}$ and hence there exists a representation

$$\mathscr{S} = \left\{n \in \mathbb{Z}_+^{\mathscr{Q}}\colon \bar{A}n \leq \bar{C}\right\},$$

for some choice of $\bar{A}$ and $\bar{C}$. Moreover, $\bar{A}, \bar{C}$ can be chosen to have nonnegative elements, since $0 \in \mathscr{S}$, and $(n \in \mathscr{S},\ n' \in \mathbb{Z}_+^{\mathscr{Q}},\ n' \leq n)$, implies $n' \in \mathscr{S}$.

Hence, under repacking the vector $n = (n_q,\ q \in \mathscr{Q})$ has stationary distribution

$$\pi(n) = G^{-1} \prod_{q \in \mathscr{Q}} \frac{\nu_q^{n_q}}{n_q!}, \qquad n \in \mathscr{S},$$

where G is the normalizing constant chosen so that $\pi(n)$ sums to unity over the set $\mathscr{S}$. We can thus apply any of our earlier results concerning the truncation of independent Poisson random variables to a polytope.

As a simple example, consider a three-node network in which a call between nodes α and β can be routed via the direct link of capacity $C_{\alpha\beta}$, or on the two-link alternative route through node γ, for $(\alpha, \beta, \gamma) = (1, 2, 3),\ (2, 3, 1),\ (3, 1, 2)$. Let $\nu_{\alpha\beta}$ be the arrival rate of calls between nodes α and β, and let $n_{\alpha\beta}$ be the number of calls in progress between nodes α and β, for $(\alpha, \beta) =$

$(1,2),(2,3),(3,1)$. Then it is easy to check [29] that

$$\mathscr{S} = \{(n_{12}, n_{23}, n_{31}) \in \mathbb{Z}_+^3 : n_{\alpha\beta} + n_{\beta\gamma} \leq C_{\alpha\beta} + C_{\beta\gamma}$$
$$\text{for } (\alpha,\beta,\gamma) = (1,2,3), (2,3,1) \text{ and } (3,1,2)\}.$$

From Section 2 we can deduce a central limit theorem for the network, or from (1.9) and (1.10) we can develop a fixed point approximation. Observe that a call between nodes α and β can be blocked either because there are no free circuits out of node α or because there are no free circuits out of node β: under the limiting regime of Section 2 it is as if these events are independent. The Erlang fixed point estimates the blocking probability out of node β by E_β, where

$$E_\beta = E(\rho_\beta, C_{\alpha\beta} + C_{\beta\gamma}), \qquad \rho_\beta = \nu_{\alpha\beta}(1 - E_\alpha) + \nu_{\beta\gamma}(1 - E_\gamma)$$

for $(\alpha,\beta,\gamma) = (1,2,3),(2,3,1),(3,1,2)$.

Similarly, if the network comprises a complete graph on four nodes with arbitrary repacking, the constraints defining $\mathscr{S}$ can be written in the form [65]

$$\sum_{\alpha\in\Delta}\sum_{\beta\notin\Delta} n_{\alpha\beta} \leq \sum_{\alpha\in\Delta}\sum_{\beta\notin\Delta} C_{\alpha\beta}, \qquad \Delta \subset \{1,2,3,4\},$$

and the Erlang fixed point attaches a blocking probability to each of these natural cut constraints. For a complete graph on five nodes with arbitrary repacking, there is a further class of linear constraints defining the polytope $\mathscr{S}$ in addition to the natural cut constraints [29]. For networks with arbitrary repacking, the results of Papernov (described in [29]) provide necessary and sufficient conditions on graph topology for cut constraints alone to define the polytope $\mathscr{S}$.

We consider in detail a further example of repacking in Section 4.6.

3.4. *Reduced load approximations.* The Erlang fixed point (1.9), (1.10) is a special case of a general *reduced load approximation*, constructed as follows. Model link j, $j = 1,2,\ldots,J$, as if calls on routes passing through link j arrive as independent Poisson streams at given reduced loads, calls on route r require A_{jr} circuits and call holding times are exponential. The resulting finite state Markov chain provides blocking probabilities at link j for each route passing through link j, as a function of the reduced loads on link j. Finally, calculate the reduced load on a link by assuming that blocking events are independent from link to link along each route. This procedure produces a set of fixed point equations, which are just (1.9) and (1.10) in the case of a network with fixed routing and A a 0–1 matrix. More generally, the procedure provides a straightforward and canonical approximation scheme for networks involving features such as alternative routing and trunk reservation, as we shall illustrate in Section 4.

When A is not a 0–1 matrix, the Erlang fixed point (3.1), (3.2) contains a further approximation over and above that contained in the general reduced load approximation: it treats the combined arrival process at a link as a Poisson process, rather than a compound Poisson process, and ignores the

feature that multiple circuits on a link occupied by a single call are all released together. The distinction disappears under the limiting regime of Section 2, where, despite its crude method of dealing with calls requiring multiple circuits from a single link, the Erlang fixed point gets the loss probabilities for such calls asymptotically correct (Theorem 3.10). In general, we shall use the term "Erlang fixed point" for reduced load approximations where the finite state Markov chain modelling a single link is the Erlang model of single link.

The fixed point equations produced by the general reduced load approximation are guaranteed to have a solution, by the Brouwer fixed point theorem. We shall see that there can be multiple solutions, even in the apparently simple case of fixed routing (Example 4.21). In practice the equations are usually solved by the method of successive approximation, often with some additional damping. See Whitt [71] for some results on the convergence of successive approximation schemes.

4. Symmetric networks. In this section we consider a limiting regime where link capacities and loads are fixed or bounded and where the numbers of links and routes in the network increases to infinity. We shall find that reduced load approximations emerge as asymptotically exact provided routing within networks is sufficiently diverse. Throughout this section, assume call holding periods are exponentially distributed with unit mean.

4.1. *Fixed routing in star networks.* Consider a network with $\mathscr{R}$ the set of all subsets of size w of $\{1, 2, \ldots, J\}$. Set $\nu_r = \nu$, $r \in \mathscr{R}$, and $C_j = C$, $j = 1, 2, \ldots, J$. For $w = 2$, this network might model a star network, where there are J stations linked through a common hub and a call is equally likely to connect any pair of stations. The model also arises in a study by Mitra and Weinberger [53] of database locking: "links" become items in the database, and a "call" is a transaction that involves several items in the database. A transaction locks the items it needs, and the case $C = 1$ corresponds to one copy of each item. (For closely related queueing models, see [12, 30].)

Let

$$\lambda = \nu \binom{J-1}{w-1},$$

so that λ is the rate of offered traffic involving any single station. Let $Q_{Jn}(t)$ be the numbers of links with n busy circuits at time t. Let $X_{Jn}(t) = J^{-1}Q_{Jn}(t)$ and let $X_J(t) = (X_{Jn}(t),\ n = 0, 1, \ldots, C)$. Note that $X_J(t)$ lies in the simplex

$$\Delta = \left\{(x_0, x_1, \ldots, x_C) \in \mathbb{R}_+^{C+1} \colon \sum_0^C x_n = 1\right\}.$$

Write X_J, or occasionally $X_J(\cdot)$, for $(X_J(t),\ t \geq 0)$. Let "$\Rightarrow$" denote convergence in distribution of random elements in the state space Δ or the space of all sample paths $D_\Delta[0, \infty)$; for background see Ethier and Kurtz [14]. The following functional law of large numbers is due to Whitt [71].

4.1 THEOREM. *If* $X_J(0) \Rightarrow X(0)$ *in* Δ, *then* $X_J \Rightarrow X$ *in* $D_\Delta[0,\infty)$, *where* $X(\cdot) = (x_0(\cdot), x_1(\cdot), \ldots, x_C(\cdot))$ *is the unique solution to the equations*

$$\frac{d}{dt}\left(\sum_{m=0}^{n} x_m(t)\right) = (n+1)x_{n+1}(t) - \gamma(t)x_n(t), \quad n = 0,1,\ldots,C-1, \tag{4.2}$$

where

$$\gamma(t) = \lambda(1 - x_C(t))^{w-1}.$$

PROOF. We sketch a proof based on Whitt [71] and Hunt [26]. The first step, which we omit, establishes from tightness arguments that the sequence $(X_J)_J$ is relatively compact in $D_\Delta[0,\infty)$ and that the limit of any convergent subsequence has continuous sample paths [71]. The second step, which we outline, uses a martingale convergence argument to characterize the limit [26]. Let

$$d(X_J(t)) = \lim_{h \downarrow 0} \mathbb{E}\left[\frac{X_J(t+h) - X_J(t)}{h} \middle| X_J(t)\right], \tag{4.3}$$

$$M_J(t) = X_J(t) - X_J(0) - \int_0^t d(X_J(s))\,ds. \tag{4.4}$$

Then M_J is an $\{\mathcal{F}_t^{X_J}\}$-martingale and the cross variation $[M_{Jm}, M_{Jn}](t) \to 0$ as $J \to \infty$. Hence ([14], page 339, Theorem 1.4) $M_J \Rightarrow 0$. Next combine the preceding two steps. Along any convergent subsequence of $(X_J)_J$ the continuous mapping theorem (see Whitt [70]) implies that M_J converges, and by the second step, the limit must be 0. Thus the differential equations (4.2), derived from the integral representation (4.4) and an explicit calculation of the drift (4.3), are satisfied by the limit of every convergent subsequence. Hence the sequence $(X_J)_J$ converges to the solution of (4.2). □

Thus $x = (x_0, x_1, \ldots, x_C) \in \Delta$ is a fixed point of the flow (4.2) if and only if

$$(n+1)x_{n+1} = \gamma x_n, \qquad n = 0,1,\ldots,C-1, \tag{4.5}$$

where

$$\gamma = \lambda(1 - x_C)^{w-1}. \tag{4.6}$$

But equations (4.5) are just the equilibrium equations for the stationary distribution of a single link of capacity C circuits offered Poisson traffic at rate γ. Thus $x_C = E(\gamma, C)$ and (4.6) becomes

$$\gamma = \lambda(1 - E(\gamma, C))^{w-1}. \tag{4.7}$$

Now $E(\gamma, C)$ is an increasing function of γ. Hence (4.7) provides a unique solution for γ. The vector x given by

$$x_n = \frac{\gamma^n}{n!}\left(\sum_{m=0}^{C} \frac{\gamma^m}{m!}\right)^{-1}, \qquad n = 0,1,\ldots,C, \tag{4.8}$$

is thus the unique fixed point in Δ of the flow (4.2). Whitt [71] goes on to establish convergence of any solution of the flow (4.2) to this fixed point and to deduce convergence in probability of X_{Jn}.

4.9 COROLLARY. *For any initial vector* $X(0)$, $X(t) \to x$ *as* $t \to \infty$.

4.10 THEOREM. *Let* X_{Jn} *be the proportion of links with* n *busy circuits in a stationary star network with* J *links. Then* $X_{Jn} \to_p x_n$ *as* $J \to \infty$, *for each* n, *where* x_n *is given by* (4.8) *with* γ *the unique solution to* (4.7).

4.11 REMARK. Let Γ be the one-dimensional submanifold of Δ defined by

$$\Gamma = \left\{ (x_0, x_1, \ldots, x_C) : x_n = \frac{\gamma^n}{n!} \left(\sum_{m=0}^{C} \frac{\gamma^m}{m!} \right)^{-1}, \right. \tag{4.12}$$
$$\left. n = 0, 1, \ldots, C, \text{ for some } \gamma \in (0, \infty) \right\}.$$

The submanifold Γ is a natural space to consider: as λ varies over the range $(0, \infty)$, the fixed point x traces out Γ. Also, if $\gamma(t)$ is replaced by a constant γ in (4.2), then a solution to the resulting linear differential equations converges exponentially quickly to the submanifold Γ, to the point parametrized by γ. This is clear, since with $\gamma(t)$ replaced by γ, equations (4.2) are Kolmogorov's forward equations for the Markov process describing a single link of capacity C circuits offered Poisson traffic at rate γ (see, e.g., [15], page 461).

Under the identification $x_C(t) = 1 - \exp(-y(t))$, the function $U(y, C)$ defined by (3.4) satisfies

$$U(y(t), C) = \sum_{m=0}^{C} m x_m(t), \quad \text{for } x(t) \in \Gamma.$$

Now, from Theorem 4.1,

$$\begin{aligned} \frac{d}{dt} U(y(t), C) &= \lambda (1 - x_C(t))^w - \sum_{m=0}^{C} m x_m(t) \\ &= \lambda \exp(-wy(t)) - U(y(t), C) \quad [\text{for } x(t) \in \Gamma] \\ &= -J^{-1} \frac{d}{dy} \Phi(y) \Big|_{y=y(t)}, \end{aligned}$$

where $\Phi(y)$ is the objective function of the revised dual problem (3.5). Thus Φ is a form of potential function on the manifold Γ. This establishes an interesting connection between the revised dual problem (3.5) and the differential equations (4.2). It would be nice to be able to deduce directly the convergence of any solution to (4.2) to the unique minimum of the convex function Φ, but this seems difficult: in particular, the manifold Γ is not closed under the flow (4.2).

4.13 REMARK. Whitt [71] has conjectured the form of a functional central limit theorem to accompany his functional law of large numbers (4.1). Define

$$V_{Jn}(t) = J^{-1/2}(Q_{Jn}(t) - Jx_n),$$

and let $V_J(t) = (V_{Jn}(t),\ n = 1, 2, \ldots, C)$. For each value of J, let $V_J = (V_J(t),\ t \geq 0)$ be the stationary version of this normalized stochastic process. Then Whitt's conjecture is that $V_J \Rightarrow V$ in $D_{\mathbb{R}^C}[0, \infty)$, where V is a stationary multivariate Ornstein–Uhlenbeck diffusion process. Whitt [71] establishes the convergence of the drift of V_J as $J \to \infty$, but the infinitesimal covariance is more difficult. Hunt [26] has established the conjecture in the case $C = 1$: then $Q_{J1}(t)$ is Markov, and the drift and infinitesimal variance of the limiting process $V_1(\cdot)$ are $-(1 + w\lambda(1 - x_1)^{w-1})V_1(t)$ and $2wx_1$, respectively.

4.2. *Poisson convergence.* For the network considered in Section 4.1, the Erlang fixed point equations (1.9) and (1.10) reduce to the single equation

$$B = E\big(\lambda(1 - B)^{w-1}, C\big) \tag{4.14}$$

for every value of J. This is just a rewritten version of (4.7), with $B = E(\gamma, C)$, and so, from Theorem 4.10,

$$1 - L_r \to (1 - B)^w, \qquad r \in \mathscr{R}, \tag{4.15}$$

as $J \to \infty$. Thus the Erlang fixed point approximation is asymptotically exact.

Variants of this result can be established for many other forms of symmetric network. We describe two further examples. Consider a network with $\mathscr{R}$ the set of all subsets of size less than or equal to W, and let

$$\nu_r = \lambda_{|r|}\binom{J-1}{|r|-1}^{-1},$$

so that the total offered traffic involving link j and $w - 1$ other links is λ_w. Let all links have capacity C. Then ([71, 78]) as $J \to \infty$,

$$1 - L_r \to (1 - B)^{|r|}, \qquad r \in \mathscr{R}, \tag{4.16}$$

where B is the solution to

$$B = E\left(\sum_{w=1}^{W} \lambda_w (1 - B)^{w-1}, C\right). \tag{4.17}$$

Next consider an unbalanced star network, where

$$\mathscr{R} = \big\{\{j, k + J_\alpha\} : j = 1, 2, \ldots, J_\alpha,\ k = 1, 2, \ldots, J_\beta\big\}$$

and $C_j = C_\alpha$, $j = 1, 2, \ldots, J_\alpha$, $C_j = C_\beta$, $j = J_\alpha + 1, \ldots, J_\alpha + J_\beta$. This network might model a system where stations are of two types linked through a common hub and a call connects two randomly chosen stations of distinct types. This is perhaps the simplest variant of the star network to allow distinct blocking probabilities to emerge from (1.9) and (1.10). For this network, if

$J_\alpha = Jp_\alpha$, $J_\beta = Jp_\beta$, with $p_\alpha + p_\beta = 1$, and $\nu_r = \lambda/J$, then [78] as $J \to \infty$

$$1 - L_r \to (1 - B_\alpha)(1 - B_\beta), \tag{4.18}$$

where (B_α, B_β) is the solution to

$$B_\alpha = E\big(\lambda p_\beta(1 - B_\beta), C_\alpha\big), \qquad B_\beta = E\big(\lambda p_\alpha(1 - B_\alpha), C_\beta\big). \tag{4.19}$$

Let $\xi_j(t)$ be the number of calls offered to link j in the interval $[0, t]$, excluding calls blocked at other links, in a stationary network. Then for the preceding three examples, the process $(\xi_j(t), t \geq 0)$ converges in distribution to a Poisson process as $J \to \infty$ ([71, 78]; see also [7]). Moreover, the rate of the Poisson process is just the reduced load on link j under the Erlang fixed point approximation. This observation provides an intuitively appealing explanation for the limiting results (4.14)–(4.19) and leads us to formulate a more general conjecture.

Consider the network described in Section 1.2, with A a 0–1 matrix. Let

$$\delta_j = \sum_k \left(\sum_{r:\, j,k \in r} \nu_r \right)^2, \qquad \delta = \max_j \delta_j.$$

If δ_j is small, then two calls through link j are unlikely to share another link elsewhere in the network; thus the network measure δ assesses the diversity of routing. Consider now a sequence of such networks, with J, $|\mathscr{R}| \to \infty$. Suppose that the reduced load on link j, given by (1.10), is ρ_j for each network in the sequence.

4.20 CONJECTURE. *If $\delta \to 0$, then $(\xi_j(t), t \geq 0)$ converges in distribution to a Poisson process of rate ρ_j.*

The conjecture is based on the belief that the key property of the examples considered earlier is diversity of routing (appropriately formulated) rather than any particular network symmetries. The conjecture is made plausible by the results of Palm and Khintchine on Poisson convergence for the superposition of independent point processes (Section 1.1; [11], Chapter 4). The difficulty, of course, is that in the systems we consider the superimposed streams are not quite independent.

4.21 EXAMPLE. This example, taken from [78], illustrates circumstances in which the various arrival streams at a link approximate to independent Poisson processes conditional on their rates, but where the rates themselves are random variables. It also shows that the general reduced load approximation may have multiple solutions, even in the case of fixed routing.

Links $1, 2, \ldots, 2K$ each have C circuits. An arriving call of type α requires all C circuits from a link chosen randomly from the set $\{1, 2, \ldots, K\}$ and a single circuit from each link in a randomly chosen subset of size C from the set $\{K + 1, \ldots, 2K\}$. Similarly, an arriving call of type β requires all C circuits from a link chosen randomly from the set $\{K + 1, \ldots, 2K\}$ and a single circuit

from each link in a randomly chosen subset of size C from the set $\{1, 2, \ldots, K\}$. For any given link, let ν be the total arrival rate at the system of calls requiring the C circuits from that link. Observe that the model differs from those considered earlier in that a call may require more than one circuit from a single link. Let

$$F(\rho, \bar{\rho}) = \left(\bar{\rho} + \frac{\rho^C}{C!}\right)\left(\bar{\rho} + \sum_{n=0}^{C} \frac{\rho^n}{n!}\right)^{-1}$$

$$\overline{F}(\rho, \bar{\rho}) = 1 - \left(\bar{\rho} + \sum_{n=0}^{C} \frac{\rho^n}{n!}\right)^{-1}.$$

These expressions correspond to Erlang's formula (1.1), but for a link offered a Poisson stream at rate ρ of calls requiring a single circuit and an independent Poisson stream at rate $\bar{\rho}$ of calls requiring C circuits. The expression $F(\rho, \bar{\rho})$ gives the loss probability for calls requiring a single circuit, and the expression $\overline{F}(\rho, \bar{\rho})$ gives the loss probability for calls requiring C circuits. We can use the expressions to construct a reduced load approximation: let

$$B_k = F(\rho_k, \bar{\rho}_k), \qquad \overline{B}_k = \overline{F}(\rho_k, \bar{\rho}_k), \qquad k = 1, 2, \ldots, 2K,$$

where $\rho_k, \bar{\rho}_k$ are the reduced loads of the two types of call at link k, calculated by assuming a stream of traffic at link i is thinned by a factor $(1 - B_i)$ or $(1 - \overline{B}_i)$, according as it requires one or C circuits from link i. Thus a symmetric solution, that is one in which $(\rho_k, \bar{\rho}_k) = (\rho, \bar{\rho})$, for $k = 1, 2, \ldots, 2K$, satisfies

$$B = F(\rho, \bar{\rho}), \qquad \overline{B} = \overline{F}(\rho, \bar{\rho}),$$

where

$$\rho = \nu C(1 - \overline{B})(1 - B)^{C-1}, \qquad \bar{\rho} = \nu(1 - B)^C.$$

By the Brouwer fixed point theorem there must exist a solution to these equations, and hence a symmetric solution to the reduced load approximation. But now this solution to the reduced load approximation may not be unique, and, in particular, there may exist asymmetric solutions. Indeed, as ν increases above a critical level, there emerges an asymmetric solution in which $(B_k, \overline{B}_k) = (B_\beta, \overline{B}_\alpha)$ or $(B_\alpha, \overline{B}_\beta)$, according as $k \le K$ or $k > K$, where $(B_\alpha, \overline{B}_\alpha; B_\beta, \overline{B}_\beta)$ satisfy the equations

$$B_\alpha = F(\rho_\alpha, \bar{\rho}_\beta), \qquad \overline{B}_\alpha = \overline{F}(\rho_\beta, \bar{\rho}_\alpha),$$

$$B_\beta = F(\rho_\beta, \bar{\rho}_\alpha), \qquad \overline{B}_\beta = \overline{F}(\rho_\alpha, \bar{\rho}_\beta),$$

with

$$\rho_\alpha = \nu C\big(1 - \overline{B}_\alpha\big)(1 - B_\alpha)^{C-1}, \qquad \bar{\rho}_\alpha = \nu(1 - B_\alpha)^C,$$

$$\rho_\beta = \nu C\big(1 - \overline{B}_\beta\big)(1 - B_\beta)^{C-1}, \qquad \bar{\rho}_\beta = \nu(1 - B_\beta)^C.$$

For example, if $(\nu, C) = (0.2, 10)$, then these equations have a solution

$$\left(B_\alpha, \bar{B}_\alpha; B_\beta, \bar{B}_\beta\right) = (0.01, 0.25; 0.13, 0.74).$$

Further, the asymmetric solution can be a good approximation for the behavior of a network with K large over a finite period. Observe that if the network has a preponderance of type α calls in progress, then arriving calls are more likely to be accepted if they too are of type α rather than β. Offered traffic at each of links $1, 2, \ldots, K$ will be approximately Poisson at rates $\rho_\beta, \bar{\rho}_\alpha$. Similarly, offered traffic at each of links $K+1, \ldots, 2K$ will be approximately Poisson at rates $\rho_\alpha, \bar{\rho}_\beta$, where $\rho_\alpha > \rho_\beta$ and $\bar{\rho}_\beta < \bar{\rho}_\alpha$. But if the network is observed over a long enough period, it eventually will flip to the opposite regime, where there is a preponderance of type β calls in progress, with the symmetric solution corresponding to an unstable intermediate regime.

We expect the preceding reduced load approximation and the associated bistable behaviour to emerge from a limiting regime where $K \to \infty$ with ν and C held fixed. We do not pursue this here: instead, we explicitly demonstrate bistable behaviour in the much simpler case when $C = K$. Let n_α, n_β be the number of calls in progress of types α and β, respectively. Then (n_α, n_β) is Markov: indeed, either n_α or n_β is zero and so $(n_\alpha - n_\beta)$ is a birth and death process. The stationary distribution for $n = (n_\alpha - n_\beta)$ is

$$\pi(n) = \left[2(\nu+1)^K - 1\right]^{-1}\binom{K}{n}\nu^n, \qquad n = -K, -K+1, \ldots, K.$$

The stationary distribution for (n_α, n_β) thus has two modes, at $(n^*, 0)$ and $(0, n^*)$ where $n^* = [K\nu(\nu+1)]^{-1} + 1$ provided $\nu \in (K^{-1}, K)$. Type α calls are offered to links in the set $\{1, 2, \ldots, K\}$ at rate $K\nu$ during the periods while $n_\alpha \geq 0$, and at rate zero during the periods while $n_\beta > 0$. These periods have mean lengths $(\nu+1)^K(K\nu)^{-1}$ and $[(\nu+1)^K - 1](K\nu)^{-1}$, respectively.

REMARK. Hajek and Krishna [23] recently have established a result concerning networks in which each route is two links long and each link can carry at most one call at a time. Let $\nu = \max_{j,k} \nu_{\{j,k\}}$, $\tau = \max_j \sum_k \nu_{\{j,k\}}$. They show that, as $\nu \to 0$ with τ fixed, the error of the reduced load approximation tends to zero, uniformly over all such networks.

4.3. *Alternative routing.* We describe a model of alternative routing in a fully connected network. Suppose that K nodes are linked to form a complete graph. Between any pair of nodes, calls arise at rate ν and there is a link of capacity C. If there is a spare circuit on the link joining the end points of a call, then the call is accepted and carried by that circuit. Otherwise the call chooses at random a two-link path joining its end points: the call is accepted on that path if both links have a spare circuit, and is lost otherwise.

The preceding model of a fully connected network is difficult to analyse, and we consider instead the following simpler version. There are $J = \frac{1}{2}K(K-1)$ links, each link comprising C circuits. Calls requiring link j arrive as a Poisson process of rate ν. If the call is blocked on its first choice link, it tries

two other links chosen at random from the $J-1$ other links. If neither of these links is full, a circuit is held from each; otherwise the call is lost. Call holding periods are exponentially distributed with unit mean. Observe that the original model is invariant under permutations of nodes, a group of order $K!$. The simpler model is invariant under permutations of links, a much larger group of order $J!$. We call it the *exchangeable model*, since links in this model can be permuted arbitrarily without affecting the stationary behaviour.

Let $Q_{Jn}(t)$ be the number of links with n busy circuits at time t. As in Section 4.1 let $X_{Jn}(t) = J^{-1}Q_{Jn}(t)$ and let $X_J(t) = (X_{Jn},\ n = 0, 1, \ldots, C)$. Again, $X_J(t)$ lies in the simplex Δ. The following result can be established by the techniques used to prove Theorem 4.1 ([18, 26]).

4.22 THEOREM. *If $X_J(0) \Rightarrow X(0)$ in Δ, then $X_J \Rightarrow X$ in $D_\Delta[0,\infty)$, where $X(\cdot) = (x_0(\cdot), x_1(\cdot), \ldots, x_C(\cdot))$ is the unique solution to the equations*

$$\frac{d}{dt}\left(\sum_{m=0}^{n} x_m(t)\right) = (n+1)x_{n+1}(t) - (\nu + \sigma(t))x_n(t), \tag{4.23}$$

$$n = 0, 1, \ldots, C-1,$$

where

$$\sigma(t) = 2\nu x_C(t)(1 - x_C(t)).$$

Thus $x = (x_0, x_1, \ldots, x_C) \in \Delta$ is a fixed point of the flow (4.23) if and only if

$$(n+1)x_{n+1} = (\nu + \sigma)x_n, \qquad n = 0, 1, \ldots, C-1, \tag{4.24}$$

where

$$\sigma = 2\nu x_C(1 - x_C). \tag{4.25}$$

A fixed point x is thus of the form (4.8), where γ solves

$$\gamma = \nu + 2\nu E(\gamma, C)[1 - E(\gamma, C)].$$

This equation for γ is equivalent to the equation

$$B = E(\nu + 2\nu B(1 - B), C) \tag{4.26}$$

for B, under the transformation $B = E(\gamma, C)$. Equation (4.26) is, of course, just a reduced load approximation. Suppose that links block independently, each with probability B. The probability that a call overflows is B and the probability it can be accepted at the other link of a two-link path is $1 - B$; the arrival rate of overflowing calls at a link is then $2\nu B(1 - B)$.

The locus of points satisfying (4.26) is illustrated in Figure 1. Observe the possibility of multiple solutions for B, for C large enough and for a narrow range of the ratio ν/C. The upper and lower solutions correspond to stable fixed points for the flow (4.23), while the middle solution corresponds to an unstable fixed point.

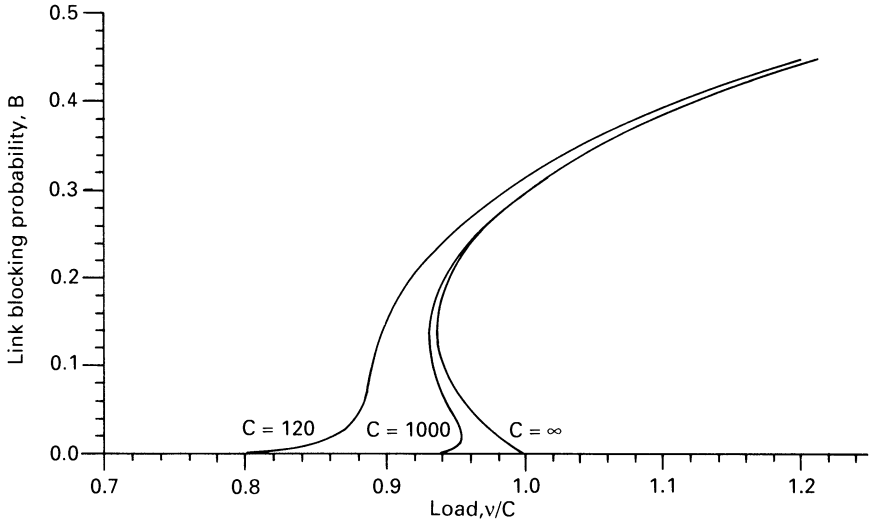

FIG. 1. *Instability of blocking probability.*

4.27 REMARK. If Γ is the one-dimensional submanifold (4.12) and if $y(t), U(y, C)$ are as in Remark 4.11, then

$$\begin{aligned}\frac{d}{dt}U(y(t),C) &= \nu(1 - x_C(t)) + 2\nu x_C(t)(1 - x_C(t))^2 - \sum_{m=0}^{C} m x_m(t) \\ &= \nu e^{-y(t)} + 2\nu e^{-2y(t)}(1 - e^{-y(t)}) - U(y(t),C) \quad [\text{for } x(t) \in \Gamma] \\ &= -\frac{d}{dy}\Phi(y)\Big|_{y=y(t)},\end{aligned}$$

where

$$\Phi(y) = \nu e^{-y} + \nu e^{-2y}\left(1 - \tfrac{2}{3}e^{-y}\right) + \int_0^y U(z,C)\,dz. \tag{4.28}$$

Thus Φ is a form of potential function on the manifold Γ.

The function Φ allows the bistability exhibited in Figure 1 to be described in the language of catastrophe theory, as we now briefly indicate. Equation (4.26) locates the stationary points of the potential function $\Phi(y)$ under the equivalence $B = 1 - e^{-y}$. Regard ν/C as the normal variable and C as the splitting variable. Then Figure 1 illustrates three cross-sections of the cusp catastrophe. Comparing (4.28) and the objective function of the revised dual problem (3.5), we see that alternative routing has led to the introduction of the second term in (4.28) and hence to nonconvexity and multiple minima.

Simulations of similar fully connected networks with tens or scores of nodes ([1, 2, 18, 43]) indicate that the upper and lower solutions for B in Figure 1

correspond to distinct locally stable modes of a stationary distribution. Simulations also indicate a hysteresis effect: if ν is varied slowly, the mode which obtains may depend not just on the current value of ν but also upon whether ν approached this value from above or below. This is just what we would expect from the preceding analysis, and an intuitive explanation is easy to provide. The lower solution corresponds to a mode in which blocking is low, calls are mainly routed directly and relatively few calls are carried on two-link paths. The upper solution corresponds to a mode in which blocking is high and many calls are carried over two-link paths. Such calls use two circuits each, and this additional demand on network resources may cause a substantial number of subsequent calls also to attempt two-link paths. Thus a form of positive feedback may keep the system in the high blocking mode.

4.29 REMARK. In [18] (see also [55]), a one-dimensional diffusion model is developed, based on the very crude approximation that $X_J(t)$ lives on the submanifold Γ. The model takes into account the number of nodes J in the system and is able to estimate the probability mass attached to each of the two modes of the stationary distribution, as well as the expected time to tunnel from one to the other.

4.4. *Trunk reservation and multiple alternatives.* For the network considered in Section 4.3, the network loss probability for the parameter choice $(\nu, C) = (115, 120)$ is about 0.12. If alternative routing is not allowed, so that a call blocked on its direct link is lost, then the network loss probability is given by Erlang's formula (1.1) to be 0.05. Thus, allowing a blocked call to attempt a two-link alternative route may *increase* the loss probability of a network, and we might expect this effect to become even more pronounced if a blocked call can attempt a sequence of alternatives routes. Recall that if a link accepts an alternatively routed call, it may later have to block a directly routed call which will then attempt to find two circuits elsewhere in the network. A natural response is to allow a link to reject alternatively routed calls if the link occupancy is above a certain level. Suppose then that in a fully connected network a call attempting a two-link alternative route is only accepted if on each of the two links the number of circuits occupied is less than $C - s$. This method of giving priority at a link to certain traffic streams is known as *trunk reservation* and the constant s is known as the trunk reservation parameter for the link.

The preceding model for a fully connected network of K nodes is difficult to analyse, and instead we consider a simpler exchangeable model. Suppose there are $J = \frac{1}{2}K(K-1)$ links and that a call blocked on its first choice link tries two other links chosen at random from among the $J - 1$ remaining links. If the number of circuits occupied on each of the two links is less than $C - s$, then the call is routed via that pair of links. If not, the call can try another pair of links chosen at random from among the $J - 3$ remaining links. On each link a trunk reservation parameter of s acts against alternatively routed calls, and a call is lost after it has tried v pairs.

Define $Q_{Jn}(t)$, $X_{Jn}(t)$ and $X_J(t)$ as in Sections 4.1 and 4.3. Thus, $X_{Jn}(t)$ is the proportion of links with n busy circuits at time t. The proof of the following result parallels that of Theorem 4.1 (cf. [18, 51]).

4.30 THEOREM. *If* $X_J(0) \Rightarrow X(0)$ *in* Δ, *then* $X_J \Rightarrow X$ *in* $D_\Delta[0,\infty)$, *where* $X(\cdot) = (x_0(\cdot), x_1(\cdot), \ldots, x_C(\cdot))$ *is the unique solution to the equations*

$$\frac{d}{dt}\left(\sum_{m=0}^{n} x_m(t)\right) = (n+1)x_{n+1}(t) - (\nu + \sigma(t))x_n(t), \qquad n = 0,1,\ldots,C-s-1$$

$$= (n+1)x_n(t) - \nu x_n(t), \qquad n = C-s,\ldots,C-1,$$

where

$$\sigma(t) = 2\nu x_C(t)\left(\sum_{m=0}^{C-s-1} x_m(t)\right)^{-1}\left\{1 - \left[1 - \left(\sum_{m=0}^{C-s-1} x_m(t)\right)^2\right]^v\right\}.$$

A fixed point $x = (x_0, x_1, \ldots, x_C) \in \Delta$ of the preceding system of differential equations satisfies

$$\begin{aligned} (n+1)x_{n+1} &= (\nu + \sigma)x_n, && n = 0,1,\ldots,C-s-1, \\ (n+1)x_{n+1} &= \nu x_n, && n = C-s,\ldots,C-1, \end{aligned} \tag{4.31}$$

where

$$\begin{aligned} \sigma &= 2\nu B_1(1-B_2)^{-1}\left\{1 - \left[1 - (1-B_2)^2\right]^v\right\}, \\ B_1 &= x_C, \qquad B_2 = \sum_{m=C-s}^{C} x_m. \end{aligned} \tag{4.32}$$

The network loss probability corresponding to a solution to (4.31) and (4.32) is

$$L = B_1\left[1 - (1-B_2)^2\right]^v. \tag{4.33}$$

We can interpret this form as follows: A call is lost if it is blocked on its first choice route, which happens with probability B_1, and if it is then blocked on each of v alternatives. It is blocked on an alternative route with probability $1 - (1-B_2)^2$, where B_2 is the probability that a link has s or fewer free circuits. Observe that equations (4.31) are the equilibrium equations for a single link of capacity C circuits and a trunk reservation parameter s offered independent Poisson streams of priority traffic at rate ν and nonpriority traffic at rate σ. Equations (4.31)–(4.33) are just the reduced load approximation for a symmetric network with trunk reservation and multiple alternatives.

The effect of varying the trunk reservation parameter s and the number of alternatives allowed v can be assessed from the above reduced load approximation. It is found that as v increases the hysteresis effect noted in Section 4.3 occurs at a lower capacity C. As s increases the hysteresis effect is diminished

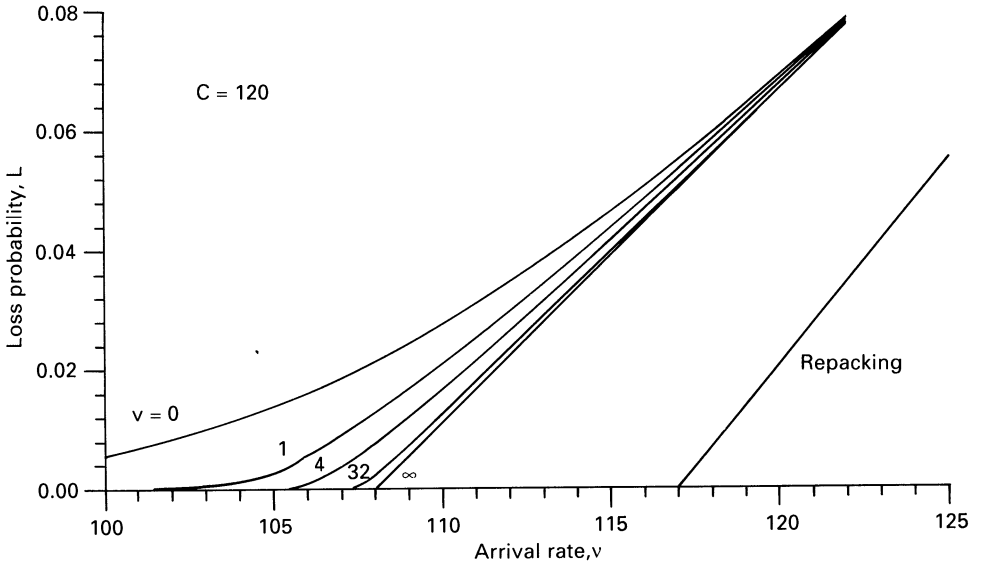

FIG. 2. *Minimal loss probability under alternative routing and repacking.*

or disappears, and the loss probability is lowered for larger values of the offered traffic ν. The minimal loss probability, using the best (integer) choice of trunk reservation parameter, is shown in Figure 2 for various values of v, the number of alternative routes. The case $v = 0$ corresponds to Erlang's formula. Note the diminishing benefits that occur from increasing v (cf. [19]).

Figure 2 includes the limiting case $v = \infty$. This case has an interesting alternative derivation, which we describe next.

4.5. *Least busy alternative schemes.* Suppose that a call blocked on its direct route in a fully connected network looks at each of the $K - 2$ two-link alternatives routes and selects, in some sense, the alternative route which is least busy. If m_j, m_k give the number of circuits *occupied* on links j and k, then we measure the business of a route through links j and k by $c(m_j, m_k)$, where $c(\cdot, \cdot)$ is a symmetric positive function, increasing in both its arguments. As an important special case, if $c(m_j, m_k) = c(m_j) + c(m_k)$ and $c(m + 1) > 2c(m)$, for all m, then the alternative chosen is a path j, k which minimizes $\max\{m_j, m_k\}$. Again, we analyse this model not for a fully connected network but instead for an exchangeable model consisting of $J = \frac{1}{2}K(K - 1)$ parallel links. We suppose a call blocked on its first choice link selects at random $K - 2$ disjoint pairs of other links and chooses the pair $\{j, k\}$ which minimizes $c(m_j, m_k)\}$. If $m_j, m_k < C$, the call holds a circuit each from links j and k for an exponentially distributed period with unit mean. If either of links j and k are full, the call is lost.

Define $Q_{J_n}(t)$, $X_{J_n}(t)$, $X_J(t)$ as in earlier sections. Thus, $X_{J_n}(t)$ is the proportion of links with n busy circuits at time t. The following result can be established ([26], cf. [50]).

4.34 THEOREM. *If $X_J(0) \Rightarrow X(0)$ in Δ, then $X_J \Rightarrow X$ in $D_\Delta[0,\infty)$, where $X(\cdot) = (x_0(\cdot), x_1(\cdot), \ldots, x_C(\cdot))$ is the unique solution to the equations*

$$\begin{aligned}
&\frac{d}{dt}\left(\sum_{m=0}^{n} x_m(t)\right) \\
&\quad = 0, \qquad n = 0, 1, \ldots, u-2 \\
&\quad = \left[u x_u(t) - 2\nu x_C(t)\right]^+, \quad n = u-1 \\
&\quad = (u+1)x_{u+1}(t) - \nu x_u(t) - \left[2\nu x_C(t) - u x_u(t)\right]^+, \quad n = u \\
&\quad = (n+1)x_{n+1}(t) - \nu x_n(t), \quad n = u+1, \ldots, C-1,
\end{aligned}$$

where

$$u(X(t)) = \min\{n : x_n(t) > 0\}.$$

A fixed point $x = (x_0, x_1, \ldots, x_C) \in \Delta$ of the preceding system of differential equations satisfies

$$\begin{aligned}
x_n &= 0, \qquad n = 0, 1, \ldots, u-1, \\
u x_u &= 2\nu(1-\alpha)x_C, \\
(u+1)x_{u+1} &= \nu x_u + 2\nu\alpha x_C, \\
(n+1)x_{n+1} &= \nu x_n, \qquad n = u+1, \ldots, C-1,
\end{aligned} \tag{4.35}$$

for some $\alpha \in [0,1]$. These equations imply that

$$\frac{\nu^{C-u}}{C(C-1)\cdots(u+1)}\left[\alpha + (1-\alpha)\frac{\nu}{u}\right] = \frac{1}{2}. \tag{4.36}$$

If $\nu \geq \frac{1}{2}C$, then $x = (0, 0, \ldots, 0, 1)$ is a stable fixed point; further, if

$$\frac{\nu^{C-n+1}}{C(C-1)\cdots n} > \frac{1}{2}, \qquad n = 1, 2, \ldots, C, \tag{4.37}$$

then this is the only fixed point. The minimum of (4.37) over n occurs when $n = [\nu]$; thus (4.37) is satisfied when $\nu > \nu^*$, where ν^* solves

$$\frac{\nu^{*C-[\nu^*]+1}}{C(C-1)\cdots[\nu^*]} = \frac{1}{2}.$$

For $\nu \leq \nu^*$, we can set

$$u = \min\left\{n : \frac{\nu^{C-n+1}}{C(C-1)\cdots n} \leq \frac{1}{2}\right\}$$

and choose α so that (4.36) is satisfied; this choice of u, α provides a solution

to (4.35) and a stable fixed point. It corresponds to a regime where no calls are lost and where, when an alternatively routed call is set up through a link, the link's occupancy rises to either u or $u + 1$. If the system is operated with a trunk reservation parameter of, for example, $s = C - u - 1$, then this regime becomes the only one.

If $\nu > \nu^*$, the only stable fixed point is $x_C = 1$. If a trunk reservation parameter of s is imposed, the network's loss probability becomes

$$L = \left(1 - \frac{1}{2}\frac{C(C-1)\cdots(C-s)}{\nu^{s+1}}\right) \times \left[\sum_{m=0}^{s} \frac{C(C-1)\cdots(C-m+1)}{\nu^m}\right]^{-1}. \tag{4.38}$$

Minimizing this with respect to s provides the curve labelled $v = \infty$ in Figure 2.

4.39 REMARK. Nakagomi and Mori [54] first observed that an independent link approximation leads to bistability in networks with alternative routing. Starting from an independent links hypothesis, Marbukh ([50, 51]) derived differential equations analogous to those presented in Theorems 4.30 and 4.34. He analyzes the stability of the corresponding fixed points and discusses the possibility of hysteresis. In particular, Marbukh [50] obtains the expression (4.38) for the loss probability of a network operating a form of least busy alternative routing with trunk reservation. Anantharam [3] considers spatially distributed versions of these models of alternative routing and obtains a hydrodynamic equation for a lattice caricature.

4.6. *Results respecting graph structure.* Our earlier discussion of alternative routing was motivated by a fully connected network, but in fact all our results have been for systems of exchangeable links. In this section we present results which respect the graph structure underlying the fully connected network. In particular, we find the asymptotic form of the optimal admission and routing policy for a fully connected network operating under repacking. The results show that, at least for this case, the graph structure does not, in fact, matter: the form of the policy is that which would have been found from an exchangeable model.

We begin with a simple static problem. Choose $p \in (0, 1)$ and suppose the edges of a complete graph on K nodes are independently coloured red with probability p and white with probability $1 - p$. The graph represents a set of K stations which are fully interconnected. At any given time only certain pairs of stations need to communicate: these pairs are identified by the red edges. The communication resource needed by a communicating pair is fixed and is exactly twice the capacity provided by a single edge. A communicating pair uses the entire capacity of the red edge joining them and can divide arbitrarily the excess traffic between them over two-edge paths through tandem nodes

which are connected to each of the pair by white edges. Let $P(K)$ be the probability that it is possible to do this simultaneously for each red edge without exceeding the capacity of any white edge. The following result is due to Hajek [21].

4.40 THEOREM. *For $p < \frac{1}{3}$, $P(K) \to 1$ as $K \to \infty$.*

PROOF. Choose a red edge. Look through the $K-2$ two-link alternatives and identify those composed of two white edges. Randomly select $(2/3)^2(K-2)$ of these and divide the unit excess traffic between them equally. Repeat for each red edge. This procedure will work provided each red edge has enough two-link alternatives composed of white edges and provided each white edge is not part of too many such alternatives.

For a given red edge, let Z_1 be the number of two-link alternatives to this edge which are composed of two white edges. Then Z_1 has a binomial distribution, $B(K-2,(1-p)^2)$. Thus,

$$\begin{aligned}
\mathbb{P}\{Z_1 < \tfrac{4}{9}(K-2)\} &\le \mathbb{E}z^{(4/9)(K-2)-Z_1} \quad (\text{for } z > 1) \\
&= \left(1 - \left(1 - \frac{1}{z}\right)(1-p)^2\right)^{K-2} z^{(4/9)(K-2)} \\
&\le \exp\left[(K-2)\left(\tfrac{4}{9}\log z - \left(1 - \frac{1}{z}(1-p)^2\right)\right)\right] \\
&= \exp\left[-(K-2)I_1(p)\right].
\end{aligned} \tag{4.41}$$

The choice $z = (9/4)(1-p)^2 > 1$ gives $I_1(p) > 0$.

For a given white edge, let Z_2 be the number of nodes such that the triangle formed by the edge and node has two white edges and one red edge. Then Z_2 has a binomial distribution, $B(K-2, 2p(1-p))$. Thus,

$$\begin{aligned}
\mathbb{P}\{Z_2 > \tfrac{4}{9}(K-2)\} &\le \mathbb{E}z^{Z_2-(4/9)(K-2)} \quad (\text{for } z > 1) \\
&= (1 + 2p(1-p)(z-1))^{K-2} z^{-(4/9)(K-2)} \\
&\le \exp\left[-(K-2)\left(\tfrac{4}{9}\log z - 2p(1-p)(z-1)\right)\right] \\
&= \exp\left[-(K-2)I_2(p)\right],
\end{aligned} \tag{4.42}$$

where since $p < \frac{1}{3}$ we can choose $z > 1$ to force $I_2(p) > 0$.

Since there are just $\frac{1}{2}K(K-1)$ edges in the network, the bounds (4.41) and (4.42) imply that $P(K) \to 1$ as $K \to \infty$. □

The method of routing used in the proof of Theorem 4.40 will divide the excess traffic from a communicating pair over a large number of alternative routes. Let $Q(K)$ be the probability that this excess traffic can be carried on a *single* alternative route, simultaneously for each communicating pair. It is helpful to rephrase this in graph-theoretic terms. Let a *triangle* be a set of

three edges joining each pair from a set of three nodes. Call a triangle *good* if it contains one red and two white edges. Then $Q(K)$ is the probability that there exists a set of disjoint good triangles such that each red edge is contained in a triangle. Clearly $Q(K) \leq P(K)$. The following result is due to Hajek [22].

4.43 THEOREM. *For* $p < \frac{1}{3}$, $Q(K) \to 1$ *as* $K \to \infty$.

4.44 A GREEDY ALGORITHM. Hajek's methods are informative about the performance of algorithms as well as the structure of random graphs, and we outline one aspect of his proof. Consider the following very simple greedy algorithm. Suppose disjoint good triangles $T_1, T_2, \ldots, T_k$ have been found already and the algorithm has not yet stopped. If there are no remaining red edges, declare the algorithm successful and stop. Otherwise, call a triangle available (after k steps) if it is a good triangle which is disjoint from $T_1, T_2, \ldots, T_k$. Choose a red edge e at random from the remaining red edges. If no available triangle contains e, declare the algorithm unsuccessful and stop. Otherwise, choose T_{k+1} at random from among the available triangles containing e. Hajek conjectures that this algorithm is successful with probability approaching 1. He proves Theorem 4.43 by consideration of a modified algorithm. Choose ε with $0 < \varepsilon < \frac{1}{3} - p$. Independently consider each white edge and delete it with probability ε. Now run the original algorithm (although now only available triangles with nondeleted edges are counted). Then this modified algorithm is successful with probability approaching 1. Hajek's proof [22] proceeds by comparing the evolution of the algorithm on a complete graph and on an exchangeable model.

Next we extend Theorem 4.40 in a different direction. Suppose the offered load between a pair of nodes is a nonnegative real valued random variable X and suppose the offered loads between different pairs of nodes are independent and identically distributed. Let the capacity of each edge be C. Let $P(K)$ be the probability that all offered loads can be carried over direct and two-edge routes, with no edge in the network required to carry more than its capacity.

4.45 THEOREM. *If* $\mathbb{E}(e^{\lambda X}) < \infty$, *for some* $\lambda > 0$, *and if*

$$2\mathbb{E}(X - C)^+ < \mathbb{E}(C - X)^+, \tag{4.46}$$

the $P(K) \to 1$ *as* $K \to \infty$.

4.47 COMMENT. If $C = 1$, $\mathbb{P}\{X = 2\} = p$, $\mathbb{P}\{X = 0\} = 1 - p$, the result reduces to Theorem 4.40. The proof is a natural extension of Hajek's proof of Theorem 4.40.

PROOF OF THEOREM 4.45. Select three nodes, label the edges joining these nodes 1, 2, 3 and let X_1, X_2, X_3 be the respective offered loads on these edges. Thus, X_1, X_2, X_3 are independent random variables, each distributed as X.

Let link 1 reserve capacity through each of edges 2 and 3 for a flow of

$$\frac{(X_1 - C)^+ (C - X_2)^+ (C - X_3)^+}{(K-2)\left\{\left[\mathbb{E}(C - X)^+\right]^2 - \varepsilon\right\}},$$

for small $\varepsilon > 0$. Repeat this step with the edge labels $1, 2, 3$ rotated cyclically. Observe that positive capacity will be reserved at most once, when the labels are such that $X_1 > C$ and $X_2, X_3 < C$, so that edge 1 has excess flow and edges 2 and 3 have excess capacity. Repeat the entire procedure with every subset of three nodes.

Select an edge with excess flow. The capacity reserved through two-link alternatives will fail to cope with this excess flow with probability

$$P_1 = \mathbb{P}\left\{\sum_{i=1}^{K-2} (K-2)^{-1} Y_i < 1\right\},$$

where Y_i, $i = 1, 2, \ldots, K-2$, are independent random variables each distributed as

$$Y = \frac{(C - X_2)^+ (C - X_3)^+}{\left[\mathbb{E}(C - X)^+\right]^2 - \varepsilon}.$$

But the expectation of Y is greater than 1, and so [10]

$$P_1 \leq \exp\left[-(K-2) I_1\right], \tag{4.48}$$

for some $I_1 > 0$.

Next select an edge with excess capacity. The capacity reserved through this link will be greater than the excess capacity with probability

$$P_2 = \mathbb{P}\left\{\sum_{i=1}^{K-2} (K-2)^{-1} Z_i > 1\right\},$$

where Z_i, $i = 1, 2, \ldots, K-2$, are independent random variables each distributed as

$$Z = \frac{(X_1 - C)^+ (C - X_2)^+ + (X_2 - C)^+ (C - X_1)^+}{\left[\mathbb{E}(C - X)^+\right]^2 - \varepsilon}.$$

But the expectation of Z is less than 1 for ε sufficiently small, by (4.46). Now X (and hence Z) has a moment generating function in a neighborhood of the origin, and so [10]

$$P_2 \leq \exp\left[-(K-2) I_2\right], \tag{4.49}$$

for some $I_2 > 0$.

Since there are just $\frac{1}{2}K(K-1)$ edges in the network, the bounds (4.48) and (4.49) imply that $P(K) \to 1$ as $K \to \infty$. □

4.50 REMARK. The bound (4.46) is clearly the best possible; it corresponds to the statement that the network load generated per edge must be less than edge capacity. The requirement that X have a moment generating function in a neighborhood of the origin can be weakened to the condition that $\mathbb{E}[X^{3+\varepsilon}] < \infty$, for some $\varepsilon > 0$ (cf. [28], page 258).

So far the results of this section have concerned systems observed at a fixed point in time. Now we return to consider a network evolving over time, driven by Poisson arrival streams. Consider the fully connected network: let each link have capacity C circuits and suppose calls arrive as independent Poisson streams at rate ν between each pair of nodes. Let call holding times be exponentially distributed with unit mean. Assume that the network is operated with repacking. At any time the route carrying a call between two nodes can be changed, and indeed that call can be divided over a collection of routes connecting the two nodes, provided the collection together provide unit capacity for the call.

We describe next one possible strategy for operation of this network. Suppose that a link counts how many calls n are in progress between its end points: an arriving call between the link's end points is provisionally accepted by the network if $n < C + s$, is rejected if $n > C + s$ and is provisionally accepted with probability α if $n = C + s$. Call this the link admission policy. The network attempts to route calls as follows. If n calls are in progress between two nodes and if $n \leq C$, then the network's aim is to route all these calls directly. If $C < n \leq C + s + 1$, the network's aim is to route C calls directly and to route the extra $n - C$ calls via underutilized two-link paths. A call provisionally accepted by a link's admission policy is accepted by the network if the network can succeed with its routing aim.

Suppose for the moment that the network's routing aim is always achieved and that all calls accepted by the link admission policies are carried. Then the number of calls in progress between two nodes is a birth and death process, with stationary distribution

$$\begin{aligned} \pi(n) &= b\frac{\nu^n}{n!}, \qquad n = 0, 1, \ldots, C + s \\ &= b\alpha\frac{\nu^n}{n!}, \qquad n = C + s + 1, \end{aligned} \tag{4.51}$$

where b is a normalizing constant chosen so that π sums to unity. The expected number of calls between two nodes which are routed via two-link paths and the expected number of circuits on a link not utilized by directly routed calls are, respectively, $G_\nu(s, \alpha) = \mathbb{E}_\pi(n - C)^+$ and $H_\nu(s, \alpha) = \mathbb{E}_\pi(C - n)^+$. The proportion of calls rejected is $L_\nu(s, \alpha) = 1 - \nu^{-1}\mathbb{E}_\pi(n)$, which reduces to Erlang's formula $E(\nu, C + s)$ when $\alpha = 0$. Observe that G_ν, H_ν and L_ν are readily calculated from the definition (4.51) of π. Define $\bar{s}, \bar{\alpha}$ by

$$\bar{s} + \bar{\alpha} = \sup\{s + \alpha: s \in \mathbb{N}, \alpha \in [0, 1), H_\nu(s, \alpha) > 2G_\nu(s, \alpha)\}, \tag{4.52}$$

and let $\overline{L}_\nu = L_\nu(\bar{s}, \bar{\alpha})$. Let ν^* be the unique solution to the equation

$$H_{\nu^*}(\infty, 0) = 2G_{\nu^*}(\infty, 0).$$

Then for $\nu < \nu^*$ the supremum in (4.52) is unbounded, and we set $\bar{s} = \infty$ and $\overline{L}_\nu = 0$.

The parameter s is rather like a *negative* trunk reservation parameter, and we shall call the strategy defined previously a trunk reservation strategy with parameters (s, α). It is just one strategy for the operation of a fully connected network on K nodes with arrival streams of rate ν and with repacking. Let $L_\nu(K)$ be the minimal network loss probability over all nonanticipating stationary strategies. (Note that the network loss probability is clearly defined for stationary strategies; also the theory of Markov decision processes with finite state space [62] ensures that we lose no essential generality by restricting attention to strategies which are stationary.)

4.53 THEOREM. *In a fully connected network on K nodes in which repacking is allowed, the minimal network loss probability satisfies*

$$L_\nu(K) > \overline{L}_\nu. \tag{4.54}$$

Further,

$$L_\nu(K) \to \overline{L}_\nu \quad \text{as } K \to \infty, \tag{4.55}$$

and hence a trunk reservation strategy with parameters $(\bar{s}, \bar{\alpha})$ defined by (4.52) *is asymptotically optimal.*

PROOF. Consider the following sequential optimization problem. We have a single link of infinite capacity offered Poisson traffic at rate ν, where accepted calls have independent exponentially distributed holding times with unit mean. Let n be the number of calls in progress. The control action allowed is to admit or to reject an arriving call, and the objective is to maximize $\mathbb{E}(n)$ over stationary policies which satisfy $2\mathbb{E}(C - n)^+ \leq \mathbb{E}(n - C)^+$, where expectations are taken with respect to the induced stationary distribution for n. Then the optimal policy is just the link admission policy described previously, namely, a trunk reservation strategy with parameters $(\bar{s}, \bar{\alpha})$. Call this the optimal single link policy. The inequality $L_\nu(K) \geq \overline{L}_\nu$ follows since otherwise a randomly chosen link in the fully connected network could be used to define a policy which improved upon the optimal single link policy. We make the inequality strict, and obtain (4.54), by observing that in a network with finitely many nodes the links are not quite able to operate the optimal single link policy, since there is a positive probability that the network will not be able to route all the calls provisionally accepted by links.

We now turn to the limit (4.55). Suppose that each link operates as its link admission policy a trunk reservation strategy with parameters (s, α), where $s + \alpha < \bar{s} + \bar{\alpha}$. Then

$$2\mathbb{E}_\pi(C - n)^+ < \mathbb{E}_\pi(n - C)^+,$$

where n is a random variable with the distribution π defined by (4.51). Consider the network at the points in time at which a call is provisionally accepted by a link. Can network routing cope with the additional call? Including the additional call, the numbers of calls in progress between pairs of nodes are stochastically dominated by a collection of independent random variables, each with the distribution π. Hence, by Theorem 4.45, the probability that the additional call can be accommodated approaches 1 as $K \to \infty$. The network's loss probability thus approaches $L_\nu(s, \alpha)$ as $K \to \infty$. Since (s, α) was chosen arbitrarily subject to $s + \alpha < \bar{s} + \bar{\alpha}$, we have the result (4.55). □

The bound $\bar{L}_\nu$ is illustrated in Figure 2, under the label repacking.

5. Lattice models. In this section we consider loss networks with fixed routing where the structure of the system is sufficiently regular that an exact analysis is possible.

5.1. *One-dimensional networks.* The results of Sections 2 and 4 indicate that under diverse routing conditions we should expect the approximation procedures introduced in Section 3 to perform well. But there are, of course, circumstances where the approximation procedures may be inadequate. For example, if a number of small capacity links are arranged one after another in a line, then we might expect considerable dependence between the number of free circuits on adjacent links. In fact it is not difficult to analyze exactly systems with an essentially one-dimensional structure, as we illustrate in this section (for a further illustration see [32], Section 3).

Suppose each route $r \in \mathscr{R}$ is a set of consecutive integers chosen from $\{1, 2, \ldots, J\}$ and that $C_j = C$, $j = 1, 2, \ldots, J$. This model arises naturally in the study of local area networks [45]; one can imagine a cable on which are positioned $J + 1$ stations and that communication between two stations uses a fraction C^{-1} of the cable's capacity over the section of cable lying between the two stations. Let the arrival rate of calls between stations i and j be

$$\nu_r = \kappa \mu^{j-i}, \qquad r = \{i + 1, i + 2, \ldots, j\}.$$

Thus calls between stations a distance v apart are attempted at rate $\kappa \mu^v$. If $\mu = 1$ distance has no effect on calling rates. Let

$$m_j = \sum_{r:\, j \in r} n_r, \qquad j = 1, 2, \ldots, J, \tag{5.1}$$

and let

$$M_J = \big\{ m = (m_j,\ j = 0, 1, \ldots, J + 1) \colon m_0 = m_{J+1} = 0, \\ m_j \in \{0, 1, \ldots, C\},\ j = 1, 2, \ldots, J \big\}. \tag{5.2}$$

Thus $m \in M_J$ describes the number of circuits in use on links $1, 2, \ldots, J$, with m_0 and m_{J+1} held identically zero. Recall that $\pi(n)$, defined by (1.2), gives the stationary distribution of n; through (5.1) the distribution π induces a

distribution over M_J for m. Call this distribution $\sigma_J(m)$. The following result is established in [34].

5.3 THEOREM. *There exists a transition matrix* $p(\cdot, \cdot)$ *over* $\{0, 1, \ldots, C\}^2$ *such that*

$$\sigma_J(m) = \frac{\prod_{j=0}^{J} p(m_j, m_{j+1})}{p^{J+1}(0,0)}, \qquad m \in M_J.$$

5.4 REMARK. Thus $m = (m_0, m_1, \ldots, m_{J+1})$ is distributed as the sample path of a Markov chain with transition matrix $p(\cdot, \cdot)$ conditioned on the end effects $m_0 = m_{J+1} = 0$.

Next we consider a slightly different one-dimensional system. Now let $C = 1$ and suppose the arrival rate of calls between stations i and j is

$$\nu_r = \kappa f(j - i), \qquad r = \{i + 1, i + 2, \ldots, j\}, \tag{5.5}$$

for f an arbitrary nonnegative function. Calls are thus attempted at a rate depending arbitrarily on the distance between stations. Assume that

$$1 - \alpha = \kappa \sum_{v=1}^{\infty} \alpha^{v+1} f(v) \tag{5.6}$$

has a solution $\alpha \in (0, 1)$ and that

$$\sum_{v=1}^{\infty} v\alpha^{v+1} f(v) < \infty. \tag{5.7}$$

Again define m_j by (5.1) and M_J by (5.2), and again let $\sigma_J(m)$ be the distribution over M_J induced by the stationary distribution (1.2). Let $X = \{0, 1\}^{\mathbb{Z}}$, with the product topology and with measurable structure given by the σ-algebra of Borel sets. Construct the probability measure σ on X corresponding to a stationary alternating renewal process, the lengths of successive blocks of ones having distribution

$$g_1(v) = \kappa(1 - \alpha)^{-1}\alpha^{v+1} f(v), \qquad v = 1, 2, \ldots, \tag{5.8}$$

and the lengths of the intervening blocks of zeros having the geometric distribution

$$g_0(u) = (1 - \alpha)\alpha^u, \qquad u = 0, 1, \ldots. \tag{5.9}$$

Let ϕ be the projection mapping on X which sends $x \in X$ to $(x_j,\ j = 0, 1, \ldots, J + 1)$.

5.10 THEOREM.

$$\sigma_J(m) = \frac{\sigma(x\colon \phi(x) = m)}{\sigma(x\colon \phi(x) \in M_J)}, \qquad m \in M_J.$$

5.11 REMARK. Let $(x_j,\ j \in \mathbb{Z})$ be the stationary alternating renewal process with renewal measure σ. Then Theorem 5.10, proved in [34], shows that the distribution of the sequence $(m_1, m_2, \ldots, m_J)$ is the same as the conditional distribution of $(x_1, x_2, \ldots, x_J)$ given that $x_0 = x_{J+1} = 0$.

5.12 EXAMPLE. If $f(v) = 1,\ v = 1, 2, \ldots,$ then the solution to (5.6) is $\alpha = (1 + \kappa^{1/2})^{-1}$, and the distributions (5.8) and (5.9) are both geometric in form. The proportion of occupied links under the stationary renewal measure σ is $1 - \alpha$. If $f(v)$ increases with v, it is quite possible to construct examples [34] where *increasing* the arrival rate parameter κ has the effect of *decreasing* the proportion of occupied links under the measures σ and σ_J.

5.13 REMARK. It is interesting to consider a continuous unbounded version of the preceding models. Imagine that users are arranged along an infinitely long cable and that a call between two points on the cable $s_1, s_2 \in \mathbb{R}$ involves just that section of cable between s_1 and s_2. Past any point along its length the cable has the capacity to carry simultaneously up to C calls: a call attempt between $s_1, s_2 \in \mathbb{R}$, $s_1 < s_2$, is lost if past any point of the interval $[s_1, s_2]$ the cable is already carrying C calls. The statistics of call attempts are most easily defined using a space–time diagram. Let a rectangle $\{(s, t)\colon s_1 \le s \le s_2,\ t_1 \le t \le t_2\}$ represent a call attempt between points s_1 and s_2 made at time t_1. If accepted, this call will last until time t_2. Assume the northeast corners of rectangles are distributed as a Poisson process of rate κ (with respect to Lebesgue measure on $(\mathbb{R})^2$). Assume that heights have unit mean, that widths have a distribution G with finite mean μ^{-1} and that heights and widths are independent of each other and of the positions of northeast corners. Informally, the probability that at time t a call attempt arises connecting a point s to a point $s + z \in (s, \infty)$ is $\kappa\, dt\, ds\, dG(z)$. Let $m(s, t)$ be the number of calls in progress past point s on the cable at time t. It is possible to show that, from an initial configuration of calls in progress at time $t = 0$, the space–time diagram defines the stochastic process $((m(s, t),\ s \in (\mathbb{R}),\ t \ge 0)$. It seems plausible (but has not been proved rigorously) that this process has a unique invariant measure, constructed as follows: let $X(s)$ be the number of customers at time s in a stationary $M/G/\infty$ queue with arrival rate κ and with service time distribution G. Then $(X(s),\ s \in (\mathbb{R})$ is a stochastic process, stationary with respect to the parameter s. Now condition $(X(s),\ s \in [L, -L])$ on the event $\{X(s) \le C,\ s \in [L, -L]\}$, and let L tend to infinity. We expect the limit process to be the unique invariant measure for $((m(s, t),\ s \in \mathbb{R}),\ t \ge 0)$. The structure of the limit process is considered in detail by Ziedins [77], in the case where G is exponential. In this case $X(s)$ is a Markov chain with transition rates $q(x - 1, x) = \kappa$, $q(x, x - 1) = x\mu$, $x = 1, 2 \ldots$.

5.2. A TREE NETWORK. In sympathy with historical developments in statistical mechanics [4] and interacting particle systems [66], we consider next a loss network defined on a Bethe lattice or tree [32].

Let T be the infinite tree with d edges from each vertex, and let T_M be the finite subgraph consisting of a distinguished vertex 0 and each vertex not more than M steps from it. Associate the edges of T_M with the links of the loss network, and suppose each link has capacity $C = 1$. Associate a call with a vertex of T_M, and suppose a call arriving at a vertex requires the use of each edge out of the vertex. Thus, a call may require the d edges out of an internal vertex, such calls arriving at rate ν for each internal vertex, or a call may require the single edge out of an external vertex, such calls arriving at rate ν_e for each external vertex. Let $n_v = 1$ or 0, according as there is a call in progress associated with vertex v or not. Thus, $n_u n_v = 0$ if u and v are neighbours. The stationary distribution of $n = (n_v, v \in T_M)$ is given by (1.2) and can be constructed directly as follows.

Let $s(v)$ be the minimum number of steps from vertex v to an external vertex and let $u(v)$ be the neighbour of v which is one step nearer to vertex 0 than v. Let $P_m = (P_m(i, j); i, j = 0, 1)$ be the transition matrix defined by

$$P_m = \begin{pmatrix} a_m & 1 - a_m \\ 1 & 0 \end{pmatrix},$$

for $a_1, a_2, \ldots, a_m$ to be determined, and set

$$\pi(n) = \pi(n_0) \prod_{v \neq 0} P_{s(u(v))}(n_{u(v)}, n_v),$$

for $\pi(n_0)$ to be determined. Under this distribution, the values observed along a path of length M from vertex 0 to an external vertex are generated by a nonhomogeneous Markov chain, with transition matrices $P_M, P_{M-1}, \ldots, P_1$. Let π_m be the induced probability that $n_v = 1$, for a vertex v with $s(v) = m$. Then $\pi_M = \pi(x_0)$ and

$$\pi_{m-1} = (1 - \pi_m)(1 - a_m), \qquad m = 1, 2, \ldots, M. \tag{5.14}$$

For $\pi(n)$ to be the stationary distribution of n the following detailed balance conditions must be satisfied:

$$(1 - \pi_1)a_1\nu_e = (1 - \pi_1)(1 - a_1), \tag{5.15}$$

$$(1 - \pi_m)a_m a_{m-1}^{d-1}\nu = (1 - \pi_m)(1 - a_m), \qquad m = 2, 3, \ldots, M, \tag{5.16}$$

$$(1 - \pi_M)a_M^d\nu = \pi_M. \tag{5.17}$$

Equation (5.17), for example, arises by considering an arrival at, or departure from, vertex 0 of a call centred there. Equation (5.15) determines a_1 in terms of ν_e. Equation (5.16) becomes

$$a_m = \frac{1}{1 + \nu a_{m-1}^{d-1}}, \qquad m = 2, 3, \ldots, M, \tag{5.18}$$

determining $a_2, a_3, \ldots, a_M$. The probability π_M is then given by (5.17), and $\pi_0, \pi_1, \ldots, \pi_{M-1}$ by the recursion (5.14).

The recursion $a_m = f(a_{m-1})$ given by (5.18) has one fixed point, a, the positive root of $a + \nu a^d = 1$. There is an associated value $\nu_e(a) = (1 - a)/a$

which generates a solution $a_1 = \cdots = a_M = a$ and $\pi_0 = \pi_1 = \cdots = \pi_M = (1 - a)/(2 - a)$. This is an appealing solution, since under it the stationary distribution over a vertex and its neighbours is identical at each internal vertex. For example, the probability of acceptance of a call centred at any internal vertex is $a^d(1 - a)/(2 - a)$. However, if

$$\nu > \frac{1}{d-1}\left(\frac{d-1}{d-2}\right)^d, \tag{5.19}$$

then $f'(a) < -1$ and so the fixed point a is unstable: a value of ν_e arbitrarily close to $\nu_e(a)$ gives rise to a sequence $a_1, a_2, \ldots, a_M$ which oscillates away from a. The stationary distribution over a vertex v and its neighbours, and, in particular, the derived acceptance probability, will then depend upon the location of v and markedly upon whether $s(v)$ is even or odd. Edge effects will predominate, no matter how large the value of M.

5.20 REMARK. A loss network defined on the infinite tree T may well have more than one invariant measure: for example, there is certainly more than one invariant measure when condition (5.19) is satisfied.

5.3. *A two-dimensional network.* For $d > 2$, most of the vertices of the tree T_M are external vertices, and so the unstable behaviour described in Section 5.2 is perhaps not unexpected. However, related phenomena can occur when the underlying graph is a section of the two-dimensional lattice.

Let V_M be the graph with vertices $\{(i, j) \in \mathbb{Z}^2: |i|, |j| \leq M\}$ and with edges between vertices unit distance apart. Associate the edges of V_M with the links of a loss network, and suppose each link has capacity $C = 1$. Associate a call with a vertex of V_M, and suppose a call arriving at a vertex requires the use of each edge out of the vertex. Thus, a call arriving at an internal vertex requires the four edges out of that vertex: assume such calls arrive at the same rate ν for each internal vertex. A call arriving at a boundary vertex requires the two or three edges out of that vertex: let the vector $\nu_b(M) = (\nu_{(i,j)}, |i| \text{ or } |j| = M)$ describe the arrival rates around the boundary. Call the vertex (i, j) odd or even, according as $i + j$ is odd or even. Two boundary conditions will be of interest. Let the odd boundary condition $\nu_b^{\text{odd}}(M)$ be defined by $\nu_{(i,j)} = \nu$ or 0, according as (i, j) is an odd or even boundary node. Similarly, let the even boundary condition $\nu_b^{\text{even}}(M)$ be defined by $\nu_{(i,j)} = 0$ or ν, according as (i, j) is an odd or even boundary node. The stationary distribution of the network is, as usual, given by the form (1.2). Let $\pi_0(\nu_b(M))$ be the induced probability that a call is in progress centred at the origin with boundary condition $\nu_b(M)$. Louth [49] has established the following result.

5.21 THEOREM. *For ν sufficiently large*

$$\pi_0\big(\nu_b^{\text{odd}}(M)\big) < \tfrac{1}{3}, \qquad \pi_0\big(\nu_b^{\text{even}}(M)\big) > \tfrac{2}{3}, \qquad \textit{for all } M \in \mathbb{Z}.$$

5.22 REMARK. Thus, for ν large enough, the effect of the boundary is felt at the centre no matter how far away it is. Louth's proof of Theorem 5.21 is an adaptation of the contour method of Peierls ([41, 58]), used to establish phase transition in the two-dimensional Ising model of a ferromagnet. In some respects the preceding network resembles an antiferromagnet—calls repel each other—although note that there is a lack of symmetry between occupied and unoccupied vertices, and configurations with adjacent vertices occupied are infeasible [49].

5.23 REMARK. Both the odd and the even boundary conditions tend to produce configurations with a chequerboard pattern for ν large enough. Calls tend to be centred at *either* the odd *or* the even vertices, depending on which boundary condition is in force. By allowing the boundary to move off to infinity, it is possible to construct two distinct invariant measures for the process defined on the infinite lattice. More pertinent to our later discussion is the following related observation: wrap the graph V_M on a torus, by identifying pairs of vertices (i, M), $(i, -M)$, for $|i| \leq M$, and (M, j), $(-M, j)$, for $|j| < M$, so that each vertex is internal with four edges incident. For ν sufficiently large, the stationary distribution π, given by (1.2), is bimodal, placing mass on configurations which are close to a chequerboard pattern centred on *either* odd vertices *or* even vertices. (See [41] for an illuminating discussion of the relationships between this form of bistability, phase transition, and the existence of multiple invariant measures for the infinite lattice.) If we condition the distribution π on the presence of a call centred at the vertex 0, we shall increase substantially the probability that a call is present at any given even vertex, no matter how far away. In Section 6 we shall be interested in the effect of capacity changes. But the stationary distribution conditioned on the presence of a call centred at the origin is just the unconditioned stationary distribution of a network which has had the capacity associated with the edges from vertex 0 removed. Thus the effects of capacity change can make themselves felt at arbitrary distances from the point of change.

5.24 REMARK. Consider an arbitrary loss network with fixed routing, with route set $\mathscr{R}$ and link-route incidence matrix A. Louth ([49]; see also [38]) defines the route interaction graph $I(\mathscr{R}, A) = (\mathscr{R}, \mathscr{E})$ to be the graph with node set $\mathscr{R}$ and with an edge $\{r_1, r_2\} \in \mathscr{E}$ if there exists a link j of positive capacity with $A_{jr_1}, A_{jr_2} > 0$. The route interaction graph of the loss network considered in this section is thus isomorphic to V_M, the graph underlying the loss network itself: routes are mapped to the vertices of V_M and two routes are adjacent in the route interaction graph if and only if the vertices are adjacent in V_M. In this example the route interaction graph is bipartite: routes can be labelled even or odd, and overlapping routes have different parity. Thus the various chains of influence between routes are in phase and reinforce one another. Louth [49] terms a loss network *frustrated* if the route interaction graph is *not* bipartite: some of the chains of influence between routes are out of phase and compete with one another. Frustration is the crucial property

that lies at the heart of the spin glass problem of statistical mechanics [16]. We return to this point in Section 6.5.

6. Optimization of routing and capacity. How should calls be routed or capacity allocated in a loss network so as to improve the performance of the network? For example, for the model of a network with fixed routing described in Section 1.2 there may be a number of routes r that carry traffic between the same two end points, and we might be interested in varying the amounts of traffic ν_r offered to each of these routes. Or we might be interested in how to allocate additional capacity over the links of the network. What is the effect on the performance of the network of changes in the parameters ν or C?

In Sections 6.1–6.3 we address these issues for a network with fixed routing, comparing exact answers obtained from the stationary distribution (1.2) with approximate answers obtained from the Erlang fixed point. In Sections 6.4 and 6.5 we describe how the latter can be used to provide substantial insights into the issues of decentralization and long range influence.

6.1. *An exact result under fixed routing.* Consider the basic model of a loss network with fixed routing, introduced in Section 1.2. Let $n = (n_r,\ r \in \mathscr{R})$ have the distribution (1.2), and write $\nu = (\nu_r,\ r \in \mathscr{R})$ and $C = (C_1, C_2, \ldots, C_J)$ for the vectors of offered traffics and capacities, respectively. Suppose that a call carried out on route r generates an expected revenue w_r (or, equivalently, interpret w_r as the cost of losing a call on route r). Then the rate of return from the network will be

$$W^\pi(\nu; C) = \mathbb{E}^\pi\Big(\sum_r w_r n_r\Big),$$

where the expectation is taken with respect to the stationary distribution (1.2). Throughout this section we shall use the superscript π to mark quantities calculated from this exact distribution. Thus, for example, L_r^π will denote the loss probability on route r, given by (1.5). The following result [35] is readily established by an explicit differentiation.

6.1 THEOREM.

$$\frac{d}{d\nu_r} W^\pi(\nu; C) = (1 - L_r^\pi)(w_r - c_r^\pi), \tag{6.2}$$

where

$$c_r^\pi = W^\pi(\nu; C) - W^\pi(\nu; C - Ae_r). \tag{6.3}$$

Equation (6.2) shows that the effect of increasing traffic on route r can be assessed by the following rule of thumb: An additional call offered to route r will be accepted with probability $1 - L_r^\pi$; if accepted, it will earn w_r directly, but at an implied cost of c_r^π to other calls arriving at the network later. The

relation (6.3) shows that the implied cost c_r^π also has an interpretation as the shadow price for a decrease in the capacity of the network from C to $C - Ae_r$.

Hunt [25] has established the following alternative representation of c_r^π by an explicit calculation from the representation (6.3).

6.4 THEOREM.

$$c_r^\pi = w_r - \frac{\sum_{r'} w_{r'} \operatorname{cov}^\pi(n_r, n_{r'})}{\mathbb{E}^\pi(n_r)}.$$

6.2. *Differentiation of an approximation.* The exact forms (6.2) and (6.3) are awkward to use, because of their dependence on L_r^π and W^π and hence on the partition function. They are, however, extremely suggestive and have more tractable analogues obtained from the Erlang fixed point approximation.

Let $E = (E_j,\ j = 1, 2, \ldots, J)$ be the unique solution to the fixed point equations (3.1) and (3.2). To emphasize its dependence on the parameter vectors ν and C, write $E = E(\nu; C)$. Under the Erlang fixed point approximation, the rate of return from the network is given by

$$W(\nu; C) = \sum_r w_r \lambda_r, \quad \text{where } \lambda_r = \nu_r(1 - L_r),\ 1 - L_r = \prod_j (1 - E_j)^{A_{jr}},$$

and $E = E(\nu; C)$. Thus L_r, λ_r are the loss probability and carried traffic, respectively, on route r, as calculated from the approximation. Let

$$\delta_j = \rho_j\big(E(\rho_j, C_j - 1) - E(\rho_j, C_j)\big). \tag{6.5}$$

Extend the definition (1.1) to nonintegral values of scalar C by linear interpolation, and at integer values of C_j define the derivative of $W(\nu; C)$ with respect to C_j to be the left derivative.

6.6 THEOREM.

$$\frac{d}{d\nu_r} W(\nu; C) = (1 - L_r) s_r \tag{6.7}$$

and

$$\frac{d}{dC_j} W(\nu; C) = c_j, \tag{6.8}$$

where $s = (s_r, r \in R)$, $c = (c_1, c_2, \ldots, c_J)$ *are the unique solution to the linear equations*

$$s_r = w_r - \sum_j c_j A_{jr}, \tag{6.9}$$

$$c_j = \delta_j \frac{\sum_r A_{jr} \lambda_r (s_r + c_j)}{\sum_r A_{jr} \lambda_r}. \tag{6.10}$$

6.11 REMARK. It is interesting to compare Theorems 6.1 and 6.6. Note that using the Erlang fixed point approximation has the effect of replacing the quantity c_r^π appearing in the derivative (6.2) by the additive form $\sum_j c_j A_{jr}$ appearing, through (6.9), in the derivative (6.7). The vector $c^\pi = (c_r^\pi, r \in \mathscr{R})$ is difficult to evaluate. The vector $c = (c_1, c_2, \ldots, c_J)$ has potentially much smaller dimension and is defined through (6.9) and (6.10) as the solution to a set of just J linear equations.

We can interpret s_r as the *surplus value* of a call on route r: if such a call is accepted, it will earn w_r directly but at an *implied cost* of c_j for each circuit used from link j. The implied costs c measure the expected knock-on effects of accepting a call upon later arrivals at the network. From (6.8) it follows that c_j is also a *shadow price*, measuring the sensitivity of the rate of return to the capacity C_j of link j. The local character of (6.9) and (6.10) is striking. The right-hand side of (6.9) involves costs c_j only for links j on the route r, while (6.10) exhibits c_j in terms of an average, weighted over just those routes through link j, of $s_r + c_j$.

Expression (6.5) for δ_j is called Erlang's improvement formula, and its use in capacity expansion decisions is known as Moe's principle ([6], pages 216–221, [67]). Observe that δ_j is simply the increase in the rate at which calls are blocked if a single link offered Poisson traffic at rate ρ_j has its capacity reduced by one circuit, and that δ_j increases from zero to one as ρ_j increases from zero to infinity.

The formal mathematical derivation of the relationships (6.7)–(6.10) is, in a certain sense, elementary. These are, after all, simply relationships between the derivatives of an implicitly defined function. The elementary approach is, however, tedious. It is illustrated in [35] where a frontal assault is made on (1.9) and (1.10), involving calculation of partial and total derivatives of $B_1, B_2, \ldots, B_J$ with respect to ν and C and subsequent reduction of the equations obtained. An elegant alternative approach is suggested by the work of Whittle [74]. The fixed point $B_1, B_2, \ldots, B_J$ locates a stationary point of a potential function, and so derivatives of W can be deduced from derivatives of the potential function (note that Whittle [74] focusses on the saddlepoint approximation, but his approach applies in the present context also). Unfortunately, this approach does not appear capable of extension to more complex models, involving, for example, trunk reservation: these models lack the required characterization of fixed points as stationary points of a potential function. A third approach [36] is based on the differentiation of W on various carefully constructed manifolds around the point $(\nu, C) \in \mathbb{R}^{\mathscr{P}} \times \mathbb{R}^J$. Currently, this approach seems to be the most widely applicable; it also seems to be the most direct, in that equations possessing the local character of (6.9) and (6.10) emerge naturally.

We leave to Sections 6.4 and 6.5 a discussion of some of the implications of (6.9) and (6.10); next we consider how accurately the derivative (6.7) estimates the derivative (6.2).

6.3. *Accuracy of approximation.* We have seen in Sections 3 and 4 that the Erlang fixed point is asymptotically accurate under various limiting regimes. The derivatives (6.7) and (6.8) may, in some circumstances, inherit the asymptotic accuracy. More precisely, consider the limiting regime defined in Section 2, where capacities and offered traffics increase to infinity together. Let $c_r^\pi(N), c_j(N)$ be the exact and approximate quantities c_r^π, c_j, respectively, evaluated for the Nth network. Then Hunt [25] has established the following two theorems.

6.12 THEOREM. *Under the limiting regime* (2.11)–(2.13),

$$\lim_{N\to\infty} c_r^\pi(N) = w_r - x_r^{-1} \sum_{r'} w_{r'} \operatorname{cov}(u_r, u_{r'}), \qquad r \in \mathscr{R},$$

where $u = (u_r, r \in \mathscr{R})$ *is the truncated multivariate normal vector of Theorem* 2.19. *Further, there exists a vector* $\bar{c} = (\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_J)$ *such that*

$$\lim_{N\to\infty} \left(c_r^\pi(N), r \in \mathscr{R}\right) = \bar{c}A.$$

6.13 THEOREM. *Under the limiting regime* (2.11) *and provided the set of critically loaded links* $\mathscr{O}$ *is empty,*

$$\lim_{N\to\infty} \left(c_r^\pi(N), r \in \mathscr{R}\right) = \lim_{N\to\infty} \left(\sum_j c_j(N) A_{jr}, r \in \mathscr{R} \right)$$

$$= w\Sigma A_{\mathscr{B}}^T \left(A_{\mathscr{B}} \Sigma A_{\mathscr{B}}^T\right)^{-1} A_{\mathscr{B}}.$$

6.14 REMARK. Recall, from Remark 6.11, that the Erlang fixed point approximation has the effect of replacing the exact quantity c^π with the additive form cA. Theorem 6.13 shows that, under the limiting regime considered and provided $\mathscr{O}$ is empty, this replacement is asymptotically accurate. Theorem 6.12 shows that, whether or not $\mathscr{O}$ is empty, the replacement of c^π by an additive form $\bar{c}A$ is asymptotically accurate. Hunt [25] provides an example which shows that in the presence of critically loaded links the limiting value of $c(N)A$ may not equal $\bar{c}A$. Recall, from Remark 3.13, that in the presence of critically loaded links the Erlang fixed point approximation may not be accurate to order $o(N^{-1/2})$, even though a product-form decomposition for loss probabilities is. This corresponds to our present observation that in the presence of critically loaded links implied costs calculated from the Erlang fixed point approximation may not be accurate to order $o(1)$, even though an additive form is. The extent of the discrepancy between limiting implied costs calculated exactly and from the Erlang fixed point approximation is related to the diversity of routing in the network. In particular, it is related to the extent of dependence between numbers of free circuits on critically loaded links, which can be assessed using the covariance matrix (2.23)—recall Remark 2.25.

We can pursue the issue of routing diversity from another angle. Consider the star network of Section 4.1, under the limiting regime defined there. Let $w_r = 1$, $r \in \mathscr{R}$. Then c_r^π and c_j are independent of r and j, respectively. Hunt [25] deduces from Theorem 6.4 that

$$c_r^\pi = 1 - \frac{\mathrm{var}^\pi(\Sigma_r n_r)}{\mathbb{E}^\pi(\Sigma_r n_r)}.$$

The variance-to-mean ratio involved here can be calculated, in the case $C = 1$, from the stationary distribution of the diffusion limit obtained for this network (Remark 4.13). Hunt [25] establishes that, when $C = 1$, $c_r^\pi - mc_j \to 0$ as $J \to \infty$, where J is the number of links in the network and m is the number of links involved in a single call. Thus, the derivatives (6.7) and (6.8) are asymptotically accurate for the star network of Section 4.1 with $C = 1$ under the limiting regime defined there. It seems likely that these derivatives are asymptotically accurate under much more general circumstances, but current proof techniques seem inadequate. For the star network with $C > 1$, the conjectured diffusion limit of Remark 4.13 would decide the issue. But the star network is a very special case: we expect the implied cost c_j to be asymptotically accurate under the conditions of Conjecture 4.20. It is possible to define implied costs for networks with alternative routing and trunk reservation [36]: in circumstances where reduced load approximations for loss probabilities are asymptotically accurate (e.g., Sections 4.3–4.6) are implied costs accurate also?

6.4. *Decentralization.* The implied cost equations (6.9) and (6.10) are linear in the costs c, and so there are a large number of methods available for finding a solution. For example, the most direct method would involve explicitly inverting a matrix of dimension J. In this section we describe some simple iterative methods which have the property that the calculations involved can be carried out in a distributed fashion, using only local information.

As motivation it is helpful to think in terms of the following model of a distributed computation. Suppose there is a limited intelligence in the form of arithmetical processing ability available for each link j and for each route r. This intelligence may be located centrally or it may be distributed over the nodes of the network; for example, the processing for route r might be carried out at the source node for calls on route r. We will require the possibility of limited communication between the intelligences of link j and route r provided A_{jr} is nonzero. Suppose for the moment that the value of δ_j is fixed and known to the intelligence of link j, while the value of λ_r is fixed and it, together with w_r, is known to the intelligence of route r.

Consider now (6.9) and (6.10). One method for attempting a solution to these equations is repeated substitution. Choose a vector c; substitute it into (6.9) to obtain a vector s; substitute these into (6.10) to obtain a revised vector c, and repeat. This computation can be distributed over the intelligences of links and routes, since (6.9) for s_r involves implied costs c_j only for links j on

route r, while (6.10) for c_j involves only surplus values s_r for routes r passing through link j.

Will repeated substitution converge? Define a linear mapping $f\colon \mathbb{R}^J \to \mathbb{R}^J$ by $f = (f_1, f_2, \ldots, f_J)$,

$$f_j(x) = \delta_j\Big(\sum_r A_{jr}\lambda_r\Big)^{-1} \sum_r A_{jr}\lambda_r\Big(w_r + c_j - \sum_k x_k A_{kr}\Big).$$

Thus, repeated substitution calculates the sequence $f^m(x)$, $m = 1, 2, \ldots$. For $a \in (0, 1)$, define $f_{(a)}\colon \mathbb{R}^J \to \mathbb{R}^J$ by $f_{(a)}(x) = (1 - a)x + af(x)$. Thus, $f_{(a)}(\cdot)$ is simply a damped version of $f(\cdot)$. Define a norm on $\mathbb{R}^J$ by

$$\|x\|_A = \max_{j,r:\, A_{jr}>0} \Big\{\sum_k |x_k|A_{kr} - |x_j|\Big\}.$$

The following result can be established (cf. [35]).

6.15 THEOREM. (i) *Suppose that* $\|\delta\|_A < 1$. *Then the mapping* $f\colon \mathbb{R}^J \to \mathbb{R}^J$ *is a contraction under the norm* $\|\cdot\|_A$, *and so the sequence* $f^m(x)$, $m = 1, 2, \ldots$, *converges to* c *for any* $x \in \mathbb{R}^J$.

(ii) *If* $a < J^{-1}$, *then the sequence* $f_{(a)}^m(x)$, $m = 1, 2, \ldots$, *converges to* c *for any* $x \in \mathbb{R}^J$.

6.16 REMARK. The condition $\|\delta_A\| < 1$ is a form of light traffic condition. It may well be violated in networks with long routes or heavily loaded links. In these circumstances an attempt to solve (6.9) and (6.10) by repeated substitution may fail: it may, for example, produce a sequence which oscillates away from the solution. Part (ii) of Theorem 6.15 shows that a sufficient damping of the function f can guarantee convergence. The damping can be implemented by a distributed computation, but individual intelligences now require some knowledge of the network beyond that locally available. In fact, it is sufficient for the intelligences to know just one item of global information, namely, J, the total number of links in the network.

The quantities δ_j and λ_r appearing in (6.9) and (6.10) are not fixed and known. However, they can be estimated by intelligences of links and routes from, for example, local measurements of carried loads. The estimates can then by used in a distributed computation of the vector s. Finally, the derivatives (6.7) can be used to implement a decentralized hill-climbing search procedure able to vary routing patterns in response to changes in the demands on the network (see [35] for a fuller discussion). Thus the structural form of the derivatives exposed by Theorem 6.6 directly suggests a decentralized routing algorithm.

6.5. *Long range influence.* How sensitive is the network to perturbations in capacities or offered traffics? To explore this question further, it is helpful to imagine a network with a very large number of links and nodes, but where

routes are short and each route overlaps with only a small number of other routes. Define the matrices $\lambda = \text{diag}(\lambda_r)_r$ and $\kappa = \text{diag}(\delta_j(\Sigma_r A_{jr}\lambda_r)^{-1})_j$. Then (6.9) and (6.10) can be written in the form $c = w\lambda A^T\kappa - c\kappa^{-1/2}\Lambda\kappa^{1/2}$, where $\Lambda = \kappa^{1/2}A\lambda A^T\kappa^{1/2} - \text{diag}(\delta_j)_j$. Thus c has the representation

$$c = w\lambda A^T\kappa^{1/2}\left\{\sum_{n=0}^{\infty}(-\Lambda)^n\right\}\kappa^{1/2}, \tag{6.17}$$

provided the summation converges, a condition which is implied by the condition $\|\delta\|_A < 1$. Observe that $\Lambda_{jk} = 0$ if there is no route through both links j and k. Similarly, $(\Lambda^n)_{jk} = 0$ if it is not possible to reach link k from link j by a concatenation of n overlapping routes. Thus the higher powers of $(-\Lambda)$ in the representation (6.17) provide the linkage through which changes in capacity at link j can affect traffic on routes r widely separated from j. For example, if $w_r = I[r = r']$, $r \in \mathscr{R}$, then c_j assesses, through the derivative (6.8), the consequence for carried traffic on route r' of a change in the capacity of link j.

The consequences of changes in offered traffic can be assessed similarly. Let $\gamma = \text{diag}(\delta_j(1-\delta_j)^{-1}(\Sigma_r A_{jr}\lambda_r)^{-1})_j$, and let $\Delta = \lambda^{1/2}A^T\gamma A\lambda^{1/2}$. Then cA has the representation

$$cA = -w\lambda^{1/2}\left\{\sum_{n=1}^{\infty}(-\Delta)^n\right\}\lambda^{-1/2}, \tag{6.18}$$

provided the summation converges. Note that $\Delta_{rr'} = 0$ if there is no link common to routes r and r', and $(\Delta^n)_{rr'} = 0$ if it is not possible to reach route r from route r' by a concatenation of n or fewer overlapping routes. Through the derivative (6.7) and the surplus value (6.9), the relation (6.18) assesses the effects on other routes of an increased offered traffic on route r. The alternating nature of the series (6.17) and (6.18) reflects the potential for frustration (Remark 5.34), where chains of influence along different paths compete with one another.

If Λ^n or Δ^n do not decay with n, then the underlying approximation suggests that perturbations will have influence over arbitrarily great distances. Note that the existence of such effects is *not* an artifact of the approximation: Sections 5.2 and 5.3 have described examples where long range influence can be deduced from the exact distribution (1.2). It is possible to define implied costs and surplus values for fixed point approximations of alternative routing and trunk reservation and to show that they solve linear relations generalizing (6.9) and (6.10) (see [36, 37, 39]). The relationship between this approach to dynamic routing and the Markov decision theory approach of [46] and [56] is discussed in detail in [36]. The potential for long range influence and instability is more pronounced in networks with alternative routing: when the network becomes overloaded, the chains of influence along different paths tend to reinforce one another. Again, the existence of such effects is not an artifact of the approximation: recall that we have observed phase transition in the symmetric network of Section 4.3.

Additional references. Research in the area of loss networks continues apace, and the following recent reports bear on the material discussed in this paper. See Crametz ([79]) for Theorem 4.45 with *integer* rerouting; Crametz and Hunt ([80]) for a proof of Theorem 4.22 *without* the exchangeable assumption; Hunt and Kurtz ([81]) for work touched on in Section 2.5; Mitra and Gibbens ([82]) for an analysis of fixed point models of least busy alternative schemes generalizing those considered in Section 4.5; and Ross and Tsang ([83]) for computational schemes for the normalizing constant (1.4).

Acknowledgments. This paper is based on talks given at the NSF-funded Workshop on Stochastic Networks held in Madison, Wisconsin, in June 1987, and on a Special Invited Lecture given at the IMS meeting in Honolulu, Hawaii, in June 1988. I am grateful to all those who have commented on presentations and drafts of the paper, and especially to Richard Gibbens, Phil Hunt, Tom Kurtz, Neil Laws and Ward Whitt.

REFERENCES

[1] ACKERLEY, R. G. (1987). Hysteresis-type behavior in networks with extensive overflow. *British Telecom Technol. J.* **5** 42–50.

[2] AKINPELU, J. M. (1984). The overload performance of engineered networks with nonhierarchical and hierarchical routing. *AT & T Tech. J.* **63** 1261–1281.

[3] ANANTHARAM, V. (1989). Metastability and phase transitions associated to dynamic routing in networks. In *28th IEEE Conference on Decision and Control.* IEEE, New York.

[4] BAXTER, R. J. (1982). *Exactly Solved Models in Statistical Mechanics.* Academic, London.

[5] BENEŠ, V. E. (1965). *Mathematical Theory of Connecting Networks and Telephone Traffic.* Academic, New York.

[6] BROCKMEYER, E., HALSTROM, H. L. and JENSEN, A. (1948). *The Life and Works of A. K. Erlang.* Academy of Technical Sciences, Copenhagen.

[7] BROWN, T. C. and POLLETT, P. K. (1982). Some distributional approximations in Markovian queueing networks. *Adv. in Appl. Probab.* **14** 654–671.

[8] BURLEY, D. M. (1972). Closed form approximations for lattice systems. In *Phase Transitions and Critical Phenomena* **2** (C. Domb and M. S. Green, eds.) 329–374. Academic, London.

[9] BURMAN, D. Y., LEHOCZKY, J. P. and LIM, Y. (1984). Insensitivity of blocking probabilities in a circuit switching network. *J. Appl. Probab.* **21** 850–859.

[10] CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23** 493–507.

[11] COX, D. R. and ISHAM, V. (1980). *Point Processes.* Chapman and Hall, London.

[12] DOBRUSHIN, R. L. and SUKHOV, YU. M. (1976). Asymptotic investigation of star-shaped message switching networks with a large number of rays. *Problemy Peradači Informacii* **12** 70–94.

[13] DZIONG, Z. and ROBERTS, J. W. (1987). Congestion probabilities in a circuit-switched integrated services network. *Performance Eval.* **7** 267–284.

[14] ETHIER, S. N. and KURTZ, T. G. (1986). *Markov Processes: Characterization and Convergence.* Wiley, New York.

[15] FELLER, W. (1968). *An Introduction to Probability Theory and its Applications* **1**, 3rd ed. Wiley, New York.

[16] FISHER, M. E. and SINGH, R. R. P. (1990). Critical points, large-dimensionality expansions, and the Ising spin glass. In *Disorder in Physical Systems* (G. R. Grimmett and D. J. A. Welsh, eds.) 87–111. Oxford Univ. Press.

[17] FRANKEN, P., KÖNIG, D., ARNDT, U. and SCHMIDT, V. (1981). *Queues and Point Processes*. Akademie-Verlag, Berlin.
[18] GIBBENS, R. J., HUNT, P. J. and KELLY, F. P. (1990). Bistability in communication networks. In *Disorder in Physical Systems*. (G. R. Grimmett and D. J. A. Welsh, eds.) 113–127. Oxford Univ. Press.
[19] GIBBENS, R. J. and KELLY, F. P. (1990). Dynamic routing in fully connected networks. *IMA J. Math. Control Inform.* **7** 77–111.
[20] HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. Univ. Chicago Press.
[21] HAJEK, B. (1986). Personal communication.
[22] HAJEK, B. (1987). Average case analysis of greedy algorithms for Kelly's triangle problem and the independent set problem. In *26th IEEE Conference on Decision and Control*. IEEE, New York.
[23] HAJEK, B. and KRISHNA, A. (1990). Bounds on the accuracy of the reduced-load blocking formula in some simple circuit-switched networks. In *Proc. Internat. Conf. on New Trends in Communication, Control and Signal Processing, Ankara, Turkey*.
[24] HUI, J. Y. (1990). *Switching and Traffic Theory for Integrated Broadband Networks*. Kluwer, Boston.
[25] HUNT, P. J. (1989). Implied costs in loss networks. *Adv. in Appl. Probab.* **21** 661–680.
[26] HUNT, P. J. (1990). Limit theorems for stochastic loss networks. Ph.D. dissertation, Univ. Cambridge.
[27] HUNT, P. J. and KELLY, F. P. (1989). On critically loaded loss networks. *Adv. in Appl. Probab.* **21** 831–841.
[28] IBRAGIMOV, I. A. and LINNIK, YU. V. (1971). *Independent and Stationary Sequences of Random Variables*. Walters-Nordhoff, Groningen.
[29] KARZANOV, A. V. (1987). Half-integral five-terminus flows. *Discrete Appl. Math.* **18** 263–278.
[30] KELBERT, M. YA. and SUKHOV, YU. M. (1989). Poissonian limit theorem for hybrid star-like networks: A mean-field approximation. *Problemy Peradači Informacii* **25** 78–87.
[31] KELLY, F. P. (1979). *Reversibility and Stochastic Networks*. Wiley, Chichester.
[32] KELLY, F. P. (1985). Stochastic models of computer communication systems. *J. Roy. Statist. Soc. Ser. B* **47** 379–395.
[33] KELLY, F. P. (1986). Blocking probabilities in large circuit-switched networks. *Adv. in Appl. Probab.* **18** 473–505.
[34] KELLY, F. P. (1987). One-dimensional circuit-switched networks. *Ann. Probab.* **15** 1166–1179.
[35] KELLY, F. P. (1988). Routing in circuit-switched networks: Optimization, shadow prices and decentralization. *Adv. in Appl. Probab.* **20** 112–144.
[36] KELLY, F. P. (1990). Routing and capacity allocation in networks with trunk reservation. *Math. Oper. Res.* **15** 771–793.
[37] KEY, P. B. (1988). Implied cost methodology and software tools for a fully connected network with DAR and trunk reservation. *British Telecom Technol. J.* **6** 52–65.
[38] KEY, P. B. (1990). Optimal control and trunk reservation in loss networks. *Prob. Engrg. Inf. Sci.* **4** 203–242.
[39] KEY, P. B. and WHITEHEAD, M. J. (1988). Cost-effective use of networks employing Dynamic Alternative Routing. In *Proc. 12th Internat. Teletraffic Congress, Turin*. North-Holland, Amsterdam.
[40] KHINTCHINE, A. YA. (1955). Mathematical methods in the theory of queueing. *Trudy Mat. Inst. Steklov* **49**. [English translation (1960), Griffin, London.]
[41] KINDERMAN, R. and SNELL, J. L. (1980). *Markov Random Fields and Their Applications*. Amer. Math. Soc., Providence, R.I.
[42] KLEINROCK, L. (1975). *Queueing Systems, Vol II. Computer Applications*. Wiley, New York.
[43] KRUPP, R. S. (1982). Stabilization of alternate routing networks. In *IEEE Internat. Communications Conference*. IEEE, New York.
[44] KURTZ, T. G. (1987). Personal communication.
[45] LAGARIAS, J. C., ODLYZKO, A. M. and ZAGIER, D. B. (1985). Realizable traffic patterns and capacity of disjointly shared networks. *Comput. Networks* **10** 275–285.

[46] LAZAREV, W. E. and STAROBINETS, S. M. (1977). The use of dynamic programming for optimization of control in networks of commutation of channels. *Engrg. Cybernetics* **15** 107–117.
[47] LIGGETT, T. M. (1985). *Interacting Particle Systems*. Springer, New York.
[48] LIN, P. M., LEON, B. J. and STEWART, C. R. (1978). Analysis of circuit-switched networks employing originating-office control with spill-forward. *IEEE Trans. Comm.* **26** 754–765.
[49] LOUTH, G. M. (1990). Stochastic networks: Complexity, dependence and routing. Ph.D. dissertation, Univ. Cambridge.
[50] MARBUKH, V. V. (1981). Asymptotic investigation of a complete communications network with a large number of points and bypass routes. *Problemy Peradači Informacii* **16** 89–95.
[51] MARBUKH, V. V. (1983). Investigations of a fully connected channel switching network with many nodes and alternative routes. *Avtomat. i Telemekh.* **12** 86–94.
[52] MITRA, D. (1987). Asymptotic analysis and computational methods for a class of simple, circuit-switched networks with blocking. *Adv. in Appl. Probab.* **19** 219–239.
[53] MITRA, D. and WEINBERGER, P. J. (1984). Probabilistic models of database locking: Solutions, computational algorithms and asymptotics. *J. Assoc. Comput. Mach.* **31** 855–878.
[54] NAKAGOME, Y. and MORI, H. (1973). Flexible routing in the global communication network. In *Proc. 7th Internat. Teletraffic Congress*. North-Holland, Amsterdam.
[55] NELSON, R. (1986). How to control those unruly tape measures; or, computer performance modeling using stochastic catastrophe theory. *Math. Intelligencer* **8** 50–56.
[56] OTT, T. J. and KRISHNAN, K. R. (1985). State dependent routing of telephone traffic and the use of separable routing schemes. In *Proc. 11th Internat. Teletraffic Congress* (M. Akiyama, ed.) North-Holland, Amsterdam.
[57] PALM, C. (1943). Intensitätsschwankungen im Fernsprechverkehr. *Ericsson Tech.* **44** 1–189.
[58] PEIERLS, R. (1936). On Ising's model of ferromagnetism. *Proc. Cambridge Philos. Soc.* **36** 477–481.
[59] PINSKY, E. and YEMENI, Y. (1984). A statistical mechanics of some interconnection networks. In *Performance '84* (E. Gelenbe, ed.) North-Holland, Amsterdam.
[60] RAO, C. R. (1965). *Linear Statistical Inference and its Applications*. Wiley, New York.
[61] ROSS, S. M. (1970). *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.
[62] ROSS, S. M. (1983). *Introduction to Stochastic Dynamic Programming*. Academic, New York.
[63] SCHWARTZ, M. (1987). *Telecommunication Networks: Protocols, Modelling and Analysis*. Addison-Wesley, Reading, Mass.
[64] SEVASTYANOV, B. A. (1957). Limit theorems for Markov processes and their application to telephone loss systems *Theory Probab. Appl.* **2** 104–112.
[65] SEYMOUR, P. D. (1980). Four-terminus flows. *Networks* **10** 79–86.
[66] SPITZER, F. (1975). Markov random fields on an infinite tree. *Ann. Probab.* **3** 387–398.
[67] SYSKI, R. (1960). *Introduction to Congestion Theory in Telephone Systems*. Oliver and Boyd, Edinburgh.
[68] WALRAND, J. (1988). *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, N. J.
[69] WEBER, J. H. (1964). A simulation study of routing control in communication networks. *Bell System Tech. J.* **43** 2639–2676.
[70] WHITT, W. (1980). Some useful functions for functional limit theorems. *Math. Oper. Res.* **5** 67–85.
[71] WHITT, W. (1985). Blocking when service is required from several facilities simultaneously. *AT & T Tech. J.* **64** 1807–1856.
[72] WHITTLE, P. (1971). *Optimization under Constraints*. Wiley, Chichester.
[73] WHITTLE, P. (1986). *Systems in Stochastic Equilibrium*. Wiley, Chichester.
[74] WHITTLE, P. (1988). Approximation in large-scale circuit-switched networks. *Prob. Engrg. Inf. Sci.* **2** 279–291.

[75] WILKINSON, R. I. (1956). Theory for toll traffic engineering in the USA. *Bell System Tech. J.* **35** 421–513.
[76] ZACHARY, S. (1991). On blocking in loss networks. *Adv. in Appl. Probab.* **23**.
[77] ZIEDINS, I. B. (1987). Quasi-stationary distributions and one-dimensional circuit-switched networks. *J. Appl. Probab.* **24** 965–977.
[78] ZIEDINS, I. B. and KELLY, F. P. (1989). Limit theorems for loss networks with diverse routing. *Adv. in Appl. Probab.* **21** 804–830.

ADDITIONAL REFERENCES

[79] CRAMETZ, J.-P. (1990). Graph structure in large-scale communication networks. Départment de Mathématiques Appliquées, Ecole Polytechnique.
[80] CRAMETZ, J.-P. and HUNT, P. J. (1991). A limit result respecting graph structure for a fully connected loss network with alternative routing. *Ann. Appl. Probab.* **1** 436–444.
[81] HUNT, P. J. and KURTZ, T. G. (1991). Unpublished manuscript.
[82] MITRA, D. and GIBBENS, R. J. (1991). Analysis and optimal design of aggregated-least-busy-alternative routing on symmetric loss networks with trunk reservations. In *Proc. Internat. Teletraffic Congress*. North-Holland, Amsterdam.
[83] ROSS, K. W. and TSANG, D. (1990). Teletraffic engineering for product-form circuit-switched networks. *Adv. in Appl. Probab.* **22** 657–675.

STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
16 MILL LANE
CAMBRIDGE CB2 1SB
ENGLAND