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], p.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 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], p.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], pp.222-226); and he stressed comparisons between the rôle of probability in such applications and its rôle in the theory of telephone traffic ([6], p.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 equation (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 equation (1.1) still gives the stationary probability that all C circuits are busy. This was known to Erlang ([6], pp.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 above has a second chance: it overflows to a second link comprising C' circuits. The probability 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 Khinchin [40] on Palm distributions. Palm and Khinchin 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], p.203). The approach was extended by Benes, 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, for example, [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, for example, [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 $\mathcal{R}$ be the set of possible routes. In the important special case where each element of the matrix $A = (A_{jr}, j = 1, 2, \ldots, J; r \in \mathcal{R})$ 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 $r = \{j : A_{jr} = 1\}$. 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 less 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 $n(t) = (n_r(t), r \in \mathcal{R})$ 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 $\pi(n) = P\{n(t) = n\}$ is given by

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

where

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

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

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

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

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

where $e_r = (\mathrm{I}[r' = r], r' \epsilon \mathcal{R})$ 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],[73] for a discussion of the insensitivity of expression (1.2) to holding time distributions.

The classical example of the above 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, multi-processor 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, or 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 above 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,

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

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 $|\mathcal{R}\,|$ may grow as fast as exponentially with the number of nodes J, and that in the (otherwise trivial) case when $|\mathcal{R}\,| = J$ and $A = \mathrm{I}$ the size of the state space $|\mathcal{S}(C)| = \prod_{j=1}^{J} C_j$ grows rapidly with the capacity limitations $C_1, C_2, \ldots, C_J$. More formally, even in the restricted case where links have capacity one 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 $|\mathcal{R}\,|$ 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 Part 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 $\nu_r, r \in \mathcal{R}$ 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

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

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

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

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

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

In Part 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 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.* Part 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

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

where

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

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

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

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 above 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. Secondly 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 Part 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 Part 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 Part 4.

1.4. *Symmetric networks.* The central limit theorems of Part 2 concern networks with increasing link capacities and loads, but fixed network topology. In Part 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 approximation will be more accurate the more diverse the collections of routes passing through any given link. In Part 4 we provide theoretical support for this suggestion, by describing asymptotic results for networks exhibiting various symmetries. We consider sequences of networks with $\mathcal{R}$ 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 resource 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 obtain in Part 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 Part 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 Parts. Suppose each route $r \in \mathcal{R}$ is a set of consecutive integers chosen from the $\{1, 2, \ldots, J\}$, 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 Part 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 Part 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 Parts 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 Part 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.