Exactly Solvable Telephone Switching Problems

89-17

Nicholas Pippenger

Department of Computer Science
The University of British Columbia
Vancouver, British Columbia V6T 1W5
CANADA

For a certain class of telephone switching problems, much of our understanding arises from an analogy with statistical mechanics that was proposed by Beneš in 1963. This analogy has lead to the exact solution of a number of idealised problems, which we survey in this paper.

1. Introduction

This paper aims to survey a certain class of results about a certain class of communication problems. The class of problems may be described as follows.

Streams of requests, of one or more different types, arrive at a system for service. The streams are independent Poisson random processes with stationary mean interarrival times that depend only on the type of request. A request of a given type can be satisfied by one or more alternative resources or combinations of resources within the system. At each moment in time, each resource is either idle or busy. If a suitable combination of idle resources is available when a request arrives, such a combination is engaged by the request, and the constituent resources become busy for the interval of time required to service the request. The durations of these service intervals are independent exponentially distributed random variables, with means that depend only on the type of request. At the end of the service interval, the engaged resources become idle, and are once again available to serve other requests. If no suitable combination of resources is idle when a request arrives, the request is lost (not served) and has no further effect on the system.

Various problems in this class arise according to various specifications of the types of requests and their arrival rates and service times, the types of resources in the system and the combinations suitable to the different types of requests, and the policy according to which a particular suitable combination of idle resources is chosen when more than one is available for a request. Variations that do *not* fall within the scope of this class are (1) "refusals", whereby a request may be denied service despite the availability of a suitable combination of idle resources (presumably in the hope of preventing even less desirable denial of service in the future), (2) "deferrals", whereby the service of a request may be delayed until a suitable combination of idle resources is available, and (3) "rearrangements", whereby the combination of resources serving a request may be changed during the interval of service.

Numerous problems from the class described above are met in the design and analysis of telephone switching systems, local- and wide-area computer communication networks, and time- and frequency-slot assignment systems. All of these problems reduce to the analysis of a discrete-state, continuous-time, Markov jump process which, if the number of states is not too large, can be carried out by standard methods. For many problems of interest, however, the number of states is too large to allow such a solution by brute force, and a number of expedients have been devised to deal with such problems. Firstly, there are some qualitative results that apply to all problems in the class (and thus of necessity give rather little information about any particular problem). For example, a

theorem of Beneš [Be66] says that, if we seek to minimise the rate at which requests are lost, we may confine our search to policies that are "deterministic", in the sense that, to a given type of request arriving in a given state, they always assign the same combination of resources (rather than randomising among the available combinations). Secondly, there are methods of approximation by which one may hope to obtain rough quantitative information about a complicated situation. For example, there is a vast literature devoted to the analysis of "channel graphs", which at its outset replaces the probability distribution on the states by one in which each resource in idle or busy independently of the others. (For a survey of this literature as of 1979, see Hwang [Hw79]. For a more general account of approximate methods, see Syski [S60].) Finally, there may be problems that, by virtue of their special structure, can be solved analytically, either in each instance, or asymptotically in some limiting situation. Such problems give qualitative understanding of the possible phenomena that may occur in our class of systems. (One must be warned, however, by the situation in classical mechanics whereby problems, by virtue of being solvable analytically ("integrable"), are prevented from possessing certain traits (such as "chaotic behaviour") common to more general problems.) Analytically solvable problems may also be consulted for quantitative information about similar systems (though usually without any assurance as to the degree of approximation involved). These analytically solvable problems are the subject of our survey in the present paper. We shall begin with some examples that have been fundamental to the development of the theory of stochastic service systems in general.

2. Classical Systems

Consider a stream of requests of a single type arriving at a system with rate ν. Let the system contain r resources called "servers", each of which is capable of serving a request with mean service time 1. (We may arrange that the mean service time is 1 by appropriate choice of the unit of time.) Since each of the servers is capable of serving any request, it is immaterial which is chosen when more than one is idle.

A state of the system corresponds to a set of busy servers, but states with the same number of busy servers are equivalent, so instead of analysing a system with 2^r states, we can analyse one with just $r+1$ states. Let us call these states $S_0, \ldots, S_r$, where j servers are busy in S_j. If we are in state S_j, and $j < r$, then we make a transition to state S_{j+1} at rate ν (the rate at which a new request arrives to engage another server). If we are in state S_j, and $j > 0$, then we make a transition to state S_{j-1} at rate j (the rate at which one of j requests completes service and releases a server).

Since the state transition diagram is finite and strongly connected, the process is ergodic and there is a unique equilibrium distribution, which we can discover in the following way. Let P_j denote the probability of being in state S_j. Transitions from the set $\{S_0, \ldots, S_{j-1}\}$ to the set $\{S_j, \ldots, S_r\}$ alternate with transitions in the reverse direction, so in equilibrium both types of transition must occur at equal rates. Transitions from $\{S_0, \ldots, S_{j-1}\}$ to $\{S_j, \ldots, S_r\}$ are transitions from S_{j-1} to S_j, however, and similarly in the reverse direction. Thus we must have

$$\nu P_{j-1} = j P_j \qquad \text{for } 1 \le j \le r.$$

This gives us r equations among $r + 1$ unknowns, and the normalisation condition $\sum_{0 \le j \le r} P_j = 1$ completes the specification of the unique equilibrium distribution. These equations are easily solved to give

$$P_j = \frac{\frac{\nu^j}{j!}}{\sum_{0 \le i \le r} \frac{\nu^i}{i!}}.$$

This distribution is called the Erlang distribution [Er18]. It is a truncated Poisson distribution, which emerges upon taking $r = \infty$, whence the sum in the denominator becomes e^ν.

The rate at which requests are lost is equal to the probability P_r of being in the state S_r in which requests are lost, times the rate ν at which requests arrive when the system is in this state. Thus the fraction of the requests that are lost is simply

$$B(r, \nu) = \frac{\frac{\nu^r}{r!}}{\sum_{0 \le i \le r} \frac{\nu^i}{i!}}.$$

There is a remarkable identity expressing this fraction in terms of an integral rather than a sum,

$$B(r, \nu) = \frac{\nu^r e^{-\nu}}{\int_\nu^\infty x^r e^{-x}\, dx},$$

which we commend to the reader as an exercise. (The integral in the denominator is an "incomplete gamma function".) This representation has the merit of showing that $B(r, \nu)$ is an analytic function of r as well as ν, a property that will prove useful later.

A fundamental property of our class of problems is that the arrival rate for requests is independent of the state of the system. It often happens, however, that the arrival rate is diminished in proportion to the number of requests being served (as in a telephone exchange where subscribers, who cannot put calls on "hold", do not generate new requests

while they are engaged in a call). There is a simple device, however, whereby this effect can be modelled by problems in our class.

Consider n streams of requests of types $R_1, \ldots, R_n$, each arriving at a system with rate λ. Let the system contain n resources $C_1, \ldots, C_n$ called "customers", as well as r servers. A request of type R_m can be served only by a combination comprising the specific customer C_m and one (it is immaterial which) of the r servers. Again the mean service time is 1.

The states of the system again correspond to the numbers of busy servers, but now if we are in state S_j, we make a transition to state S_{j+1} at rate $(n-j)\lambda$ (the number of idle customers, times the arrival rate for requests at each idle customer). Proceeding as before we obtain the equilibrium distribution

$$P_j = \frac{\binom{n}{j}\lambda^j}{\sum_{0\le i\le r}\binom{n}{i}\lambda^i}.$$

This distribution is called the Engset distribution [En18]. If we set $p = \lambda/(1+\lambda)$, it can be rewritten as

$$P_j = \frac{\binom{n}{j}p^j(1-p)^{n-j}}{\sum_{0\le i\le r}\binom{n}{i}p^i(1-p)^{n-i}}.$$

This reveals it as a truncated Bernoulli distribution, which emerges upon taking $r = n$, whence the sum in the denominator becomes 1. We can also regard the Erlang distribution as a limiting case of the Engset distribution in which $n \to \infty$ and $\lambda \to 0$ in such a way that $n\lambda = \nu$.

Let us say that a request is "active" if when it arrives, its customer is idle. Since we are interested in the effect of the limited number of servers ($r < n$), we seek the fraction of active requests that are lost, rather than the fraction of all requests. To determine the fraction $A(r, n, \lambda)$ of active requests that are lost, we divide the rate at which active requests are lost, $(n-r)\lambda P_r$, by the rate at which active requests arrive, $\sum_{0\le j\le r}(n-j)P_j$. This gives

$$A(r,n,\lambda) = \frac{\binom{n-1}{r}p^r(1-p)^{n-1-r}}{\sum_{0\le j\le r}\binom{n-1}{j}p^j(1-p)^{n-1-j}}$$

or, in its integral representation,

$$A(r,n,\lambda) = \frac{p^r(1-p)^{n-1-r}}{(n-1-r)\int_p^1 x^r(1-x)^{n-2-r}\,dx}.$$

(The integral in the denominator is an "incomplete beta function".) The Erlang and Engset distributions play an important role not only in the exact analysis of simple systems, but also in the approximate analysis of more complex systems such as "link systems" (see Jacobæus [J50]).

Consider the lost requests from an Erlang system (with parameters r and ν) or the lost active requests from an Engset system (with parameters r, n and λ). Suppose that these requests are presented to a second, "overflow", system that has m servers. Here m may be large enough to serve all requests that arrive ($m = \infty$ for the Erlang system, $m = n - r$ for the Engset), or it may be smaller, so that requests are lost from the overflow system as well as from the "primary" system. What is the distribution of the number of requests being serviced by the overflow system?

To analyse this situation we must consider the two systems as a whole, and the states of this combined system (after grouping equivalent states) form a two-dimensional array:

$$\begin{array}{cccc} S_{0,0} & S_{0,1} & \cdots & S_{0,m} \\ S_{1,0} & S_{1,1} & \cdots & S_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ S_{r,0} & S_{r,1} & \cdots & S_{r,m} \end{array}$$

(there are j primary servers and k overflow servers busy in state $S_{j,k}$). Transitions are possible only from a state to one of its four neighbours at the cardinal compass points.

Finding the equilibrium distribution for this system is much more difficult than before, but Kosten [Kos37] managed to give expressions for $P_{j,k}$ in the case of an Erlang primary system followed by an infinite overflow system (see also Coffman, Kadota and Shepp [CoKS85]), and Brockmeyer [Br54] extended this to the case of a finite overflow system. The formulæ for the moments of the number of busy servers in the overflow system, together with the analytic continuation of $B(r, \nu)$ to non-integral r form the basis for a powerful method of approximate analysis (the "equivalent random method") due to Wilkinson [W56]. These Erlang overflow systems represent the outermost boundary of exactly solvable overflow systems at the present time, and the corresponding Engset overflow systems seem to lie just beyond this boundary.

3. Statistical Mechanics

The examples of the preceding section set the stage for what has been the most fruitful source of exactly solvable problems in telephone switching theory: the analogy

with statistical mechanics. This analogy was drawn in Beneš's paper 'A "Thermodynamic" Theory of Traffic in Connecting Networks' [Be63].

The analogy is based on the observation that in statistical mechanics one often encounters Markov processes with enormous numbers of states, but one can nevertheless obtain much useful information about what happens in the "thermodynamic limit". Roughly speaking, the thermodynamic limit is one where certain "extensive" parameters (number of particles, volume, internal energy and entropy) tend to infinity in such a way that ratios among them tend to definite limits (particles per unit volume, entropy per particle), while other "intensive" parameters (pressure, temperature) are held constant. One can easily imagine analogous limiting situations in telephone switching theory. For example, in an Erlang system we may let the arrival rate ν and the number of servers r tend to infinity in such a way that their ratio tends to a definite limit μ, the "mean number of requests per server". One can verify directly from the formulæ for $B(r, \nu)$ that in this limit we have $B(r, \nu) \to 0$ if $\mu \leq 1$, and $B(r, \nu) \to 1 - 1/\mu$ if $\mu \geq 1$. (There is a sort of "saturation" effect at $\mu = 1$.)

The methods of statistical mechanics rely heavily on the property of "reversibility", which all the systems it studies possess, but which may or may not be present in a telephone switching problem. Roughly speaking, a system is reversible if it satisfies the following criterion: if one makes a movie of the evolution of a sample of the random process in equilibrium, one cannot determine by statistical tests whether the movie is being run forwards or backwards. In terms of the Markov process describing the system, this amounts to the following: the rate at which transitions from a state S to a state S' occur is equal to the rate at which transitions from S' to S occur. (In the statistical mechanics literature, this is called the "principle of detailed balance".) It is easy to see that this criterion is fulfilled for the Erlang system; indeed, it was by exploiting it that the equations were solved by inspection. On the other hand, it is not fulfilled by the overflow systems considered by Kosten and Brockmeyer: requests can commence service in the overflow system only when all servers in the primary system are busy, but they can complete service at any time, without regard to the state of the primary system. Thus reversibility is a special property, possessed by some but not all telephone switching systems. Its presence, however, greatly simplifies the solution; see Kelly [Kell79] for a general account of the role of reversibility in the theory of stochastic service systems.

Kolmogorov [Kol36] gave the following necessary and sufficient condition for a Markov process to be reversible. For each pair of states (S, S'), define the ratio $\varrho(S, S') = \sigma(S, S')/\sigma(S', S)$, where $\sigma(S, S')$ is the rate at which the system makes a transition

into state S' when it is in state S. (We have $\varrho(S,S') = \varrho(S',S)^{-1}$ and $\varrho(S,S) = 1$.) Then the process is reversible if and only if, for every cycle $S_1, S_2, \ldots, S_n, S_1$, we have $\varrho(S_1,S_2)\,\varrho(S_2,S_3)\cdots\varrho(S_n,S_1) = 1$. If this criterion is fulfilled, then we may define a function τ on the states by choosing a reference state S_1, setting $\tau(S_1) = 1$, and then, for every other state S_n, setting $\tau(S_n) = \varrho(S_1,S_2)\,\varrho(S_2,S_3)\cdots\varrho(S_{n-1},S_n)$ for some path $S_1, S_2, \ldots, S_n$ from S_1 to S_n. (Kolmogorov's criterion ensures that this definition does not depend on the choice of path.) The equilibrium probability P_j of being in state S_j is then given by normalising the function τ: $P_j = \tau(S_j)/\sum_i \tau(S_i)$. (The normalisation ensures that this distribution does not depend on the choice of reference state.) The sum in the denominator is called the "partition function" for the system; it is usually a function of one or more parameters such as the temperature or arrival rate, though we have not indicated this dependence explicitly.

It statistical mechanics, it is common to write $\tau(S_j) = e^{-H(S_j)/T}$, where $H(S_j)$ is the "energy level" of state S_j and T is the temperature (in units for which "Boltzmann's constant" is 1). In telephone switching, it is more common to write $\tau(S_j) = \lambda^{H(S_j)}$, where λ is the arrival rate and $H(S_j)$ is the number of requests in service (the "occupancy") in state S_j. In either case P_j is called the "Gibbs distribution" associated with the potential function $H(S_j)$.

Although it is possible to define $T = -1/\log\lambda$ as the "temperature" of a telephone switching system, this leads to some unintuitive consequences: there are often negative temperatures, which turn out to be "hotter" than positive temperatures. It is better to adopt a "chemo-mechanical" rather than "thermo-mechanical" nomenclature, in which λ is taken as the fundamental independent variable. (Physicists refer to λ as the "activity", or to its reciprocal as the "fugacity".)

Beneš [Be63] acknowledged three drawbacks to the analogy with statistical mechanics. Two of these, the assumption of arrival rates that may be unrealistic, and the ability to deal with at most one of the possible policies for assigning resources to requests, stem directly from the requirement of reversibility. The third is present despite, rather than because of reversibility: "The problem of calculating the partition function [...] is, as in statistical mechanics, very difficult except in cases of unrealistic simplicity." In the remainder of this section we shall survey the cases in which this drawback has been surmounted and the calculation of the partition function has been carried out far enough to yield interesting asymptotic information.

If we seek situations in which we might determine a thermodynamic limit, two main classes emerge. In one, the number of types of requests and resources remains fixed,

while the rate at which each type of request arrives, and the number of resources of each type, tend to infinity in such a way that the ratios among them tend to definite limits. The thermodynamic limit of the Erlang system, described above, is of this type. A broad generalisation of this situation has been described by Kelly [Kell86]. Suppose that each type of request requires for service some fixed number of resources of each type, but that (beyond the equivalence of resources of the same type) no other alternatives exist. (Kelly calls this condition "fixed routing".) Then in the limiting situation described above, the fraction of the requests of each type that are lost tends to a definite limit, and these limiting values are the unique solution of a system of equations that can be written down by inspection from the data of the problem. This result permits the asymptotic analysis of many interesting systems, but the restriction to fixed routing is a strong one. It is plausible that an analogous result exists for a broader condition called "hierarchical routing", in which the types of resources are assigned to "levels", each type of request can be satisfied by any of a sequence of combinations of resources, all of the resources in a given combination are at the same level, and combinations of lower level receive preference over combinations of higher level. A still more general condition called "non-hierarchical routing" appears to lead to thermodynamic limits that violate the uniqueness present with fixed routing (see Nakagome and Mori [NM73] and Marbukh [Marb81]), but such systems have thus far resisted rigorous analysis.

In the second class of thermodynamic limits, the number of types of requests and the number of types of resources tend to infinity, while the arrival rate for each type of request and the number of resources of each type are held constant. This class corresponds more closely than the first to traditional statistical mechanics, and two large subclasses (which physicists call "monomer" systems and "dimer" systems) have drawn attention from the communication community. A monomer system is based on an undirected graph (finite or infinite), with types of requests corresponding to vertices, and types of resources corresponding to edges. In the simplest case, there is just one resource of each type, and the arrival rate λ is the same for each type of request. A request requires for service all the resources that are incident with it (no alternatives are possible). A dimer system is similar, but the resources correspond to vertices and the requests correspond to edges. (There is an intimate connexion between Gibbs distributions for systems on graphs such as these and what are called "Markov random fields"; see Preston [Pr74] for an introduction to this vast literature and Kelbert and Sukhov [KelbS83] for an application to communication networks.)

We may consider an infinite graph G to be the limit of a sequence of finite graphs G_n, which are usually subgraphs or quotient graphs of G. Let $\Psi_n(\lambda)$ denote the partition function for the system on G_n. If $\Psi_n(\lambda)^{1/n}$ tends to a definite limit $\psi(\lambda)$ as $n \to \infty$, we say that the thermodynamic limit exists. The limiting value $\psi(\lambda)$ is called the "partition function per unit"; equivalently $\log \psi(\lambda)$ is called the "free energy per unit". (Usually n denotes the number of vertices or edges in G_n, in which case we may speak of the partition function or free energy per vertex or per edge.)

In chemo-mechanical terminology, the free energy per unit may be identified with the "pressure" for many systems. The occupancy per unit (which physicists call the "density") is then given by $p = \lambda \partial(\log \psi(\lambda))/\partial\lambda$. The variance per unit is given by the next derivative, $\sigma^2 = \lambda\partial\lambda\partial(\log\psi(\lambda))/\partial\lambda^2$. (Physicists usually refer to the "compressibility" σ^2/p^2, which is equal to the derivative of density with respect to pressure, rather than to the variance.) Finally, the entropy per unit is $\eta = \log\psi(\lambda) - p\log\lambda$. (The Gibbs distribution is that distribution which maximises the entropy $-\sum_j P_j \log P_j$ among those with a given expected occupancy $\sum_j H(S_j)P_j$.)

As a simple example, consider the path Π_n in which n vertices alternate with $n-1$ edges. The monomer system on Π_n is isomorphic to the dimer system on Π_{n+1}; for definiteness, we shall consider the monomer system. A state of the system corresponds to a set of "occupied" vertices, no two of which are adjacent. It is easy to see that the number of states with k occupied vertices is $\binom{n-k}{k}$, so that the partition function is $\Psi_n(\lambda) = \sum_{k\geq 0}\binom{n-k}{k}\lambda^k$. (This partition function can be expressed in terms of the Chebyshev polynomials of the second kind.) Noting that $\Psi_0(\lambda) = \Psi_1(\lambda) = 1$ and $\Psi_{n+2}(\lambda) = \Psi_{n+1}(\lambda) + \lambda\Psi_n(\lambda)$, we see that $\Psi_n(\lambda)$ is the coefficient of z^n in the generating function $\Phi(\lambda, z) = (1+2z)/(1-z-\lambda z^2)$ (which physicists call the "grand partition function"). Other systems that can similarly be solved for their partition function include (1) the monomer or dimer system on the cycle C_n, in which n vertices alternate with n edges (this involves the Chebyshev poynomials of the first kind), (2) the dimer system for a balanced tree (this also involves the Chebyshev polynomials of the second kind), (3) the dimer system for the complete graph, with n vertices and $\binom{n}{2}$ edges (this involves the Hermite polynomials) and (4) the complete bipartite graph with $n+m$ vertices and nm edges (this involves the generalised Laguerre polynomials). See Heilmann and Lieb [HeL72], Godsil and Gutman [GoG81] and Pinsky and Yemini [PinY84].

For a problem that is solvable in the sense of the preceding paragraph (where a grand partition function can be written down), it is straightforward to prove the existence of the thermodynamic limit and to determine the limiting value of the partition function per

server. The procedure is to use Cauchy's contour integral to extract the partition function from the grand partition function, then to use the saddle-point method to determine the asymptotics of the integral. For the path, the partition function per vertex turns out to be $\psi(\lambda) = (1 + \sqrt{1 + 4\lambda})/2$, and the occupancy per vertex is given by $p = 1/2 - 1/2\sqrt{1 + 4\lambda}$. The path and the cycle are typical of a large class of "one-dimensional" systems that can be solved exactly in this way. Furthermore, both the path and the cycle may be regarded as tending to a common infinite limit graph as $n \to \infty$: the "infinite path" or "infinite cycle", in which vertices correspond to integers and edges correspond to pairs of integers that differ by 1. We may regard the limiting partition function per vertex as referring to this infinite system.

Our final example of an exactly solvable system comes directly from the statistical mechanics literature, with only a change of terminology. It may be described as "cellular radio-telephony with one frequency". Imagine a cellular radio-telephone system in which space, taken to be the Euclidean plane, is tessellated by regular hexagons in the usual "honeycomb" fashion. Suppose that requests to use one available frequency for communication arrive at cells, and that such a request can be served only if, at the time of its arrival, no request is being served in any of the six adjacent cells. We then have a monomer system on the infinite triangular lattice; the vertices correspond to cells, and the edges correspond to the frontiers between cells. (We do not consider the requests to be mobile, or their servers to be subject to rearrangement.) We may now consider a given arrival rate λ at each cell, and ask for the probability p that a cell is occupied. This is an infinite system, but we may consider a sequence of finite subsystems by considering either subgraphs (giving what physicists call "free boundary conditions") or quotient graphs (giving "periodic boundary conditions") of the infinite graph; for suitable sequences, we may prove that the thermodynamic limit exists.

The system just described is isomorphic to what physicists call the "lattice gas of hard hexagons", and it is one of a small number of models for which an exact expression can be written down for the partition function per cell. (To be precise, it should be said that, although there is a variety of evidence affording a considerable certainty that the solution is correct, this correctness has not been proved with mathematical rigour.) The solution was discovered by Baxter [Ba80, Ba82], and we shall briefly describe it here.

A particularly interesting feature of the system is that there is a critical value $\lambda_0 = (11 + 5\sqrt{5})/2 = 11.09017\ldots$ of the arrival rate per cell λ at which a "phase transition" occurs, with the solution having different analytic expressions above and below this value.

(The "saturation" of the Erlang system provides a trivial example of a phase transition, but the present example is much more interesting.)

Let

$$g(x) = \prod_{1 \le n < \infty} \frac{(1-x^{5n-4})(1-x^{5n-1})}{(1-x^{5n-3})(1-x^{5n-2})},$$

$$a(x) = \prod_{1 \le n < \infty} \frac{(1-x^{6n-4})(1-x^{6n-3})^2(1-x^{6n-2})(1-x^{5n-4})^2(1-x^{5n-1})^2(1-x^{5n})^2}{(1-x^{6n-5})(1-x^{6n-1})(1-x^{6n})^2(1-x^{5n-3})^3(1-x^{5n-2})^3}$$

and

$$b(x) = x^{-1/3} \prod_{1 \le n < \infty} \frac{(1-x^{3n-2})(1-x^{3n-1})(1-x^{5n-3})^2(1-x^{5n-2})^2(1-x^{5n})^2}{(1-x^{3n})^2(1-x^{5n-4})^3(1-x^{5n-1})^3}.$$

Then the partition function per cell $\psi(\lambda)$ is given by

$$\psi(\lambda) = \begin{cases} a(x), & \text{where } \lambda = -xg(x)^5 \text{ if } \lambda < \lambda_0; \\ \\ b(x), & \text{where } \lambda = 1/xg(x)^5 \text{ if } \lambda > \lambda_0. \end{cases}$$

The occupancy per cell is continuous at λ_0, where it assumes the value $(5-\sqrt{5})/10 = 0.276393\ldots$. The compresibility, and thus the variance per cell has a singularity of the form $|\lambda - \lambda_0|^{-1/3}$ at λ_0, so that there are large fluctuations in the occupancy near the critical arrival rate.

The nature of the phase transition may be described as follows. If one tries to "pack" requests as tightly as possible into the system, there are three ways to do this, corresponding to the three colours in the minimal colouring of the triangular lattice. When the arrival rate is less than λ_0, there will be small regions that resemble one of these packings, but all three packings will coexist in roughly equal extents. When the arrival rate exceeds λ_0, however, one of the packings will predominate over the other two, and a global order emerges. We may think of the disordered phase as a "gas", and the ordered phase as a "crystal", so the transition corresponds to "crystallisation" or "sublimation".

Heilmann and Lieb [HeL72] have shown that under rather general circumstances, a dimer system cannot have a phase transition. Even for monomer systems that are not exactly solvable, however, it may be possible to prove rigorously the existence of a phase transition. A fundamental technique for this, due to Peierls [Pe36], has been developed greatly in recent years. For examples of its application, see Dobrushin [D68], Heilmann [He72] and Heilmann and Præstgaard [HP74].

4. Conclusion

We have considered a rather restricted class of problems in this paper, and we should at least indicate what lies beyond its borders. Allowing refusals would not affect our discussion much; the theorem of Beneš [Be66] would no longer hold, but we would have no new exactly solvable systems to report. Allowing deferrals leads, of course, to the vast domain of queuing theory, which has a disjoint collection of exactly solvable systems of its own. Allowing rearrangements leads to yet another class of problems; for methods that can be brought to bear on dimer systems, see Karp and Sipser [KaS81], and for cellular radio-telephony, see Everitt and Macfadyen [EvM83] and Kelly [Kell86]. Finally, there is a class of problems that as are far from reversible as can be, in which requests arrive and are assigned resources, but never complete service. These are analogous to physical problems of "random sequential adsorption", and the methods used there can be applied to monomer and dimer systems; see Pippenger [Pip89].

We conclude by recapitulating what seem to us the most interesting open problems. Can the overflow from an Engset system be solved in closed form? Under what circumstances does hierarchical routing have a unique thermodynamic limit? What can be said rigorously about non-hierarchical routing? Can the validity of the exact solution for hard hexagons be proved rigorously?

5. References

[Ba80] R. J. Baxter, "Hard Hexagons: Exact Solution", *J. Phys. A: Math. and Gen.*, 13 (1980) L61–L70.

[Ba82] R. J. Baxter, *Exactly Solvable Models in Statistical Mechanics*, Academic Press, London, 1982.

[Be63] V. E. Beneš, 'A "Thermodynamic" Theory of Traffic in Connecting Networks', *Bell Sys. Tech. J.*, 42 (1963) 567–607.

[Be66] V. E. Beneš, "Programming and Control Problems Arising from Optimal Routing in Telephone Networks", *Bell Sys. Tech. J.*, 45 (1966) 1373–1438.

[Br54] E. Brockmeyer, "Det Simple Overflowproblem i Telefontrafikteorien", *Teletechnik*, 5 (1954) 361–374.

[CoKS85] E. G. Coffman, Jr., T. T. Kadota and L. A. Shepp, "A Stochastic Model of Fragmentation in Dynamic Storage Allocation", *SIAM J. Computing*, 14 (1985) 416–425.

[D68] R. L. Dobrushin, "The Problem of the Uniqueness of a Gibbsian Random Field and the Problem of Phase Transitions", *Funct. Anal. Appl.*, 2 (1968) 302–312.

[En18] T. Engset, "Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wahleranzahl in automatischen Fernsprechämtern", *Electrotech. Z.*, 31 (1918) 304–305.

[Er18] A. K. Erlang, "Solution of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchnges", *Post Office Elect. Eng. J.*, 10 (1918) 189–197.

[EvM83] D. E. Everitt and N. W. Macfadyen, Analysis of Multicellular Mobile Radiotelephone Systems with Loss", *British Telecomm. Technol. J.*, 1 (1983) 37–45.

[GoG81] C. D. Godsil and I. Gutman, "On the Theory of the Matching Polynomial", *J. Graph Theory*, 5 (1981) 137–144.

[He72] O. J. Heilmann, "Existence of Phase Transitions in Certain Lattice Gases with Repulsive Potential", *Let. al Nuovo Cimento*, 3, 3 (1972) 95–98.

[HeL72] O. J. Heilmann and E. H. Lieb, "Theory of Monomer-Dimer Systems", *Comm. Math. Phys.*, 25 (1972) 190–232.

[HeP74] O. J. Heilmann and E. Præstgaard, 'Phase Transition in a Lattice Gas with Third Nearest Neighbour Exclusion on a Square Lattice", *J. Phys. A: Math., Nucl. Gen.*, 15 (1974) 1913–1917.

[Hw79] F. K. Hwang, "Superior Channel Graphs", *Internat. Teletraffic Congr.*, 9 (1979).

[J50] C. Jacobæus, "A Study on Congestion in Link Systems", *Ericsson Technics*, 48 (1950) 1–68.

[KaS81] R. M. Karp and M. Sipser, "Maximum Matchings in Sparse Random Graphs", *IEEE Symp. on Foundations of Computer Science*, 22 (1981) 364–375.

[KelbS83] M. Ya. Kelbert and Yu. M. Sukhov, "Existence and Uniqueness Conditions for a Random Field Describing the State of a Switching Network", *Problems of Info. Transm.*, 19 (1983) 289–304.

[Kell79] F. P. Kelly, *Reversibility and Stochastic Networks*, Wiley, Chichester, 1979.

[Kell86] F. P. Kelly, "Blocking Probabilities in Large Circuit-Switched Networks", *Adv. Appl. Prob.*, 18 (1986) 473–505.

[Kol36] A. N. Kolmogorov, "Anfangsgründe der Theorie der Markoffschen ketten mit unendlich vielen möglichen Zuständen", *Mat. Sbornik*, 1 (1936) 607–610.

[Kos37] L. Kosten, "Über Sperrungswahrscheinlichkeiten bei Staffelschaltungen", *Elektr. Nachr.-Tech*, 14 (1937) 5–12.

[Marb81] V. V. Marbukh, "Asymptotic Investigation of a Complete Communication Network with a Large Number of Points and Bypass Routes", *Problems of Info. Transm.*, 7 (1981) 89–95.

[NM73] Y. Nakagome and H. Mori, "Flexible Routing in the Golbal Communication Network", *Internat. Teletraffic Congr.*, 7 (1973) 426.

[Pe36] R. Peierls, "Ising's Model of Ferromagnetism", *Proc. Cambridge Phil. Soc.*, 32 (1936) 477–481.

[PinY84] E. Pinsky and Y. Yemini, "A Statistical Mechanics of Some Interconnection Networks", *Performance '84*, 147–158.

[Pip89] N. Pippenger, "Random Sequential Adsorption on Graphs", *SIAM J. Discr. Math.*, 2 (1989) 393–401.

[Pr74] C. Preston, *Gibbs States on Countable Sets*, Cambridge University Press, London, 1974.

[S60] R. Syski, *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, Edinburgh, 1960; second edition: North-Holland, Amsterdam, 1986.

[W56] R. I. Wilkinson, "Theories for Toll Traffic Engineering in the USA", *Bell Sys. Tech. J.*, 35 (1956) 421–514.