SIAM J. COMPUT.
Vol. 18, No. 6, pp. 1149-1178, December 1989

006

APPROXIMATING THE PERMANENT*

MARK JERRUM† AND ALISTAIR SINCLAIR†

Abstract. A randomised approximation scheme for the permanent of a 0-1 matrix is presented. The task of estimating a permanent is reduced to that of almost uniformly generating perfect matchings in a graph; the latter is accomplished by simulating a Markov chain whose states are the matchings in the graph. For a wide class of 0-1 matrices the approximation scheme is fully-polynomial, i.e., runs in time polynomial in the size of the matrix and a parameter that controls the accuracy of the output. This class includes all dense matrices (those that contain sufficiently many 1's) and almost all sparse matrices in some reasonable probabilistic model for 0-1 matrices of given density.

For the approach sketched above to be computationally efficient, the Markov chain must be rapidly mixing: informally, it must converge in a short time to its stationary distribution. A major portion of the paper is devoted to demonstrating that the matchings chain is rapidly mixing, apparently the first such result for a Markov chain with genuinely complex structure. The techniques used seem to have general applicability, and are applied again in the paper to validate a fully-polynomial randomised approximation scheme for the partition function of an arbitrary monomer-dimer system.

Key words. permanent, perfect matchings, counting problems, random generation, Markov chains, rapid mixing, monomer-dimer systems, statistical physics, simulated annealing

AMS(MOS) subject classifications. 05C70, 05C80, 60J20, 68Q20

1. Summary.

The *permanent* of an $n \times n$ matrix A with 0-1 entries a_{ij} is defined by

$$\operatorname{per}(A) = \sum_{\sigma} \prod_{i=0}^{n-1} a_{i\sigma(i)},$$

where the sum is over all permutations σ of $[n] = \{0, \cdots, n-1\}$. Evaluating per (A) is equivalent to counting perfect matchings (1-factors) in the bipartite graph $G = (U, V, E)$, where $U = V = [n]$ and $(i, j) \in E$ if and only if $a_{ij} = 1$. The permanent function arises naturally in a number of fields, including algebra, combinatorial enumeration, and the physical sciences, and has been an object of study by mathematicians since first appearing in 1812 in the work of Cauchy and Binet (see [26] for background). Despite considerable effort, and in contrast with the syntactically very similar determinant, no efficient procedure for computing this function is known.

Convincing evidence for the inherent intractability of the permanent was provided in the late 1970s by Valiant [32], who demonstrated that it is complete for the class #P of enumeration problems, and thus as hard as counting *any* NP structures. Interest has therefore recently turned to finding computationally feasible approximation algorithms for this and other hard enumeration problems [18], [30]. To date, the best approximation algorithm known for the permanent is due to Karmarkar et al. [17] and has a runtime that grows exponentially with the input size.

The notion of approximation we will use in this paper is as follows. Let f be a function from input strings to natural numbers. A *fully-polynomial randomised approximation scheme* (fpras) for f is a probabilistic algorithm that, when presented with a string x and a real number $\varepsilon > 0$, runs in time polynomial in $|x|$ and $1/\varepsilon$ and outputs

* Received by the editors October 6, 1988; accepted for publication January 10, 1989. A preliminary version of this paper appeared in the Proceedings of the 20th ACM Symposium on Theory of Computing, Chicago, May 1988, under the title "Conductance and the Rapid Mixing Property for Markov Chains: The Approximation of the Permanent Resolved."

† Department of Computer Science, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JZ, United Kingdom.

a number that with high probability approximates $f(x)$ within ratio $1+\varepsilon$.[1] For definiteness we take the phrase "with high probability" to mean with probability at least $\frac{3}{4}$. This may be boosted to $1-\delta$ for any desired $\delta>0$ by running the algorithm $O(\lg \delta^{-1})$ times and taking the median of the results [16, Lemma 6.1]. An fpras therefore embodies a strong notion of effective approximation.

A promising approach to finding an fpras for the permanent was recently proposed by Broder [8], who reduces the problem of approximately counting perfect matchings in a graph to that of generating them randomly from an almost uniform distribution. The latter problem is then amenable to the following dynamic stochastic technique. Given a graph G, construct a Markov chain whose states correspond to perfect and "near-perfect" matchings in G and which converges to a stationary distribution that is uniform over the states. Transitions in the chain correspond to simple local perturbations of the structures. Then, provided convergence is fast enough, we can generate matchings almost uniformly by simulating the chain for a small number of steps and outputting the structure corresponding to the final state.

When applying this technique, we are faced with the thorny problem of proving that a given Markov chain is *rapidly mixing*, i.e., that after a short period of evolution the distribution of the final state is essentially independent of the initial state. "Short" here means bounded by a polynomial in the input size. Since the state space itself may be exponentially large, rapid mixing is a strong property: the chain must typically be close to stationarity after visiting only a very small fraction of the space.

Recent work on the rate of convergence of Markov chains has focused on stochastic concepts such as coupling [1] and stopping times [3]. While these methods are intuitively appealing and yield tight bounds for simple chains, the analysis involved becomes extremely complicated for more interesting processes that lack a high degree of symmetry. Using a complex coupling argument, Broder [8] claimed that the perfect matchings chain above is rapidly mixing provided the bipartite graph is *dense*, i.e., has minimum vertex degree at least $n/2$. This immediately implies the existence of an fpras for the dense permanent. However, the coupling proof is hard to penetrate; more seriously, as was first observed by Mihail [25], it contains a fundamental error that is apparently not correctable. As a result Broder has withdrawn his proof (see the erratum to [8]).

In this paper we propose a completely different approach to analysing the rate of convergence of Markov chains that is based on a structural property of the underlying weighted graph. Under fairly general conditions, a finite ergodic Markov chain is rapidly mixing if and only if the *conductance* of its underlying graph is not too small. This characterisation, discussed in detail in a companion paper [29], is related to recent work by Alon [4] and Alon and Milman [5] on eigenvalues and expander graphs.

While similar characterisations of rapid mixing have been noted by other authors (see, e.g., [2], [22]), they have been of little practical value because independent estimates of the conductance have proved elusive for nontrivial chains. Using a novel method of analysis, we are able to derive a lower bound on the conductance of Broder's perfect matchings chain under the same density assumption, thus verifying that it is indeed rapidly mixing. This is the first rapid mixing result we know of for a Markov chain with nontrivial structure. The existence of an fpras for the dense permanent is therefore established.

Reductions from approximate counting to almost uniform generation similar to that mentioned above for perfect matchings also hold for the large class of combinatorial

[1] For nonnegative real numbers a, $\tilde{a}$, ε, we say that *$\tilde{a}$ approximates a within ratio* $1+\varepsilon$ if $a(1+\varepsilon)^{-1} \leqq \tilde{a} \leqq a(1+\varepsilon)$.

structures that are *self-reducible* [16], [29]. Consequently, the Markov chain approach is potentially a powerful general method for obtaining approximation algorithms for hard combinatorial enumeration problems.

In fact, Markov chain simulation is a rather useful algorithmic tool with a variety of computational applications. Perhaps the most familiar of these are to be found in the field of statistical physics, where a physical system is typically modelled by a set of combinatorial structures, or *configurations*, each of which has an associated weight depending on its energy. Most interesting properties of the model can be computed from the *partition function*, which is just the sum of the weights of the configurations. An fpras for this function may usually be obtained with the aid of a generator that selects configurations with probabilities roughly proportional to their weights. This suggests looking for a Markov chain on configurations with the appropriate (nonuniform) stationary distribution. Such chains are in fact the basis of the ubiquitous Monte Carlo method of Metropolis et al. [6] that is extensively used among other things to estimate the expectation of certain operators on configurations under the weighted distribution.

In such applications efficiency again depends crucially on the rate of convergence of the Markov chain. Significantly, our proof technique for rapid mixing seems to generalise easily to other interesting chains. We substantiate this claim here by considering a Metropolis-style process for *monomer-dimer systems* [14], which are a model of physical systems involving diatomic molecules. In this case configurations correspond to matchings (independent sets of edges of any size) in a given weighted graph, and the weight of a configuration is the product of its edge weights. Using our earlier method of analysis, we are able to show that this Markov chain is rapidly mixing under very general conditions. As a result we deduce the existence of an fpras for the monomer-dimer partition function. This includes as a special case an fpras for the #P-complete problem of counting all matchings in an arbitrary graph.

The monomer-dimer chain also provides valuable new insight into our original problem of approximating the permanent. Most notably, by appending suitably chosen weights to the edges of the input graph we can use the chain to obtain a more elegant approximation scheme for counting perfect matchings. The scheme is immediately seen to be fully-polynomial if and only if the number of "near-perfect" matchings in the graph does not exceed the number of perfect matchings by more than a polynomial factor. This turns out to be rather a weak condition: it is satisfied not only by all dense graphs but also, in a sense that we will make precise later, by almost every bipartite graph that contains a perfect matching. Moreover, we present an efficient randomised algorithm for testing the condition for an arbitrary graph, allowing pathological examples to be recognised reliably.

A further byproduct of our work on the monomer-dimer process is the following. Consider the problem of finding a maximum cardinality matching in a given graph. The Markov chain above may be viewed as an application of the search heuristic known as *simulated annealing* [21] to this optimisation problem. Suppose the maximum cardinality of a matching is m and let $\varepsilon > 0$ be fixed. Then our results readily imply that, with a suitable choice of edge weights (or equivalently, "temperature"), the search finds a matching of size at least $(1-\varepsilon)m$ in polynomial time with high probability. This represents a considerable simplification of a recent result of Sasaki and Hajek [27].

The remainder of the paper is structured as follows. In § 2 we state our characterisation of rapid mixing in terms of conductance for a broad class of Markov chains, and illustrate by means of a simple example our technique for obtaining lower bounds on the conductance. Section 3 is devoted to a proof that Broder's method does indeed yield an fpras for the dense permanent. In § 4 we discuss the Markov chain for monomer-dimer systems and derive an fpras for the partition function. Other applications of this chain, including the improved algorithm for the permanent, appear in § 5. Section 6 deals with the approximation of the permanent of a randomly selected 0-1 matrix of given density. Finally, in § 7 we list a few open problems.

2. Markov chains and rapid mixing.

In this section we establish some general machinery for reasoning about the nonasymptotic behavior of Markov chains which will play a central role throughout the paper. We assume a nodding acquaintance with the elementary theory of finite Markov chains in discrete time. (For more information the reader is referred to [20].)

Let $(X_t)_{t=0}^{\infty}$ be a time-homogeneous Markov chain with finite state space $\mathcal{N}$ and transition matrix $P=(p_{ij})_{i,j\in\mathcal{N}}$. (All chains in this paper are assumed to be of this form.) If the chain is ergodic we let $\boldsymbol{\pi}=(\pi_i)_{i\in\mathcal{N}}$ denote its stationary distribution, the unique vector satisfying $\boldsymbol{\pi}P=\boldsymbol{\pi}$ and $\sum_i \pi_i=1$. In this case, if the chain is allowed to evolve for t steps from any initial state the distribution of its final state approaches $\boldsymbol{\pi}$ as $t\to\infty$. Necessary and sufficient conditions for ergodicity are that the chain is (a) *irreducible*, i.e., any state can be reached from any other in some number of steps; and (b) *aperiodic*, i.e., $\gcd\{s : i \text{ is reachable from } j \text{ in } s \text{ steps}\}=1$ for all $i,j\in\mathcal{N}$.

As explained in the previous section, our intention is to use simulation of an ergodic chain as a means of sampling elements of the state space $\mathcal{N}$ from a distribution close to $\boldsymbol{\pi}$. We shall always assume that individual transitions can be simulated at low cost. From the point of view of efficiency, our major concern is therefore the rate at which the chain approaches stationarity. As a time-dependent measure of deviation from the limit, we define the *relative pointwise distance* (r.p.d.) after t steps by

$$\Delta(t)=\max_{i,j\in\mathcal{N}}\frac{|p_{ij}^{(t)}-\pi_j|}{\pi_j},$$

where $p_{ij}^{(t)}$ is the t-step transition probability from state i to state j. Thus $\Delta(t)$ gives the largest relative difference between the distribution of the state at time t and $\boldsymbol{\pi}$, maximised over initial states i. Our aim is to establish conditions under which the chain is *rapidly mixing*, in the sense that $\Delta(t)$ tends to zero fast as a function of t.

An ergodic Markov chain is said to be *time-reversible* if it satisfies the detailed balance condition

$$p_{ij}\pi_i=p_{ji}\pi_j \quad \forall i,j\in\mathcal{N}. \tag{1}$$

Analysis of time-reversible Markov chains is simplified by the following observation.

LEMMA 2.1. *Suppose (X_t) is an ergodic Markov chain with finite state space $\mathcal{N}$ and transition matrix P. Let $\boldsymbol{\pi}=(\pi_i)_{i\in\mathcal{N}}$ be any vector satisfying the detailed balance condition* (1) *and the normalisation condition $\sum_i \pi_i=1$. Then the Markov chain (X_t) is time-reversible and $\boldsymbol{\pi}$ is its* (*unique*) *stationary distribution.* □

We may naturally identify a time-reversible chain with an underlying weighted graph as follows. The vertices of the graph are the states of the chain, and for each (not necessarily distinct) pair $i,j\in\mathcal{N}$ with $p_{ij}>0$ there is an edge (i,j) of weight $w_{ij}=p_{ij}\pi_i=p_{ji}\pi_j$. For convenience we set $w_{ij}=0$ for all pairs of states i,j between which no transition is possible. Note that this graph uniquely specifies the Markov chain.

As in [29], we define the *conductance* of a time-reversible chain with underlying graph H by

$$\Phi(H)=\min\frac{\sum_{i\in S,j\notin S} w_{ij}}{\sum_{i\in S}\pi_i} \tag{2}$$

where the minimisation is over all subsets S of states with $0<\sum_{i\in S}\pi_i\leqq\frac{1}{2}$. Note that the quotient in (2) is just the conditional probability that the stationary process escapes from S in a single step, given that it is initially in S. The conductance in some sense measures the rate at which the process can flow around the state space, so we might expect it to be connected with the rate of convergence of the chain. In fact, by relating $\Phi(H)$ to the second eigenvalue of the transition matrix that governs the transient behaviour of the chain, it is possible to obtain the following result. A proof can be found in [29].

THEOREM 2.2. *Let H be the underlying graph of a time-reversible ergodic Markov chain in which* $\min_i p_{ii}\geqq\frac{1}{2}$, *and let* $\pi_{\min}=\min_i \pi_i$ *be the minimum stationary state probability. Then the* r.p.d. *of the chain is bounded by*

$$\Delta(t)\leqq\frac{(1-\Phi(H)^2/2)^t}{\pi_{\min}}. \qquad \square$$

The following immediate corollary is useful. Define the function $\tau\colon \mathbb{R}^+\rightarrow\mathbb{N}$ by

$$\tau(\varepsilon)=\min\{t\in\mathbb{N}:\Delta(t')\leqq\varepsilon \text{ for all } t'\geqq t\}.$$

COROLLARY 2.3. *With the notation of Theorem* 2.2, *we have*

$$\tau(\varepsilon)\leqq\frac{2}{\Phi(H)^2}(\ln\pi_{\min}^{-1}+\ln\varepsilon^{-1}). \qquad \square$$

Theorem 2.2 allows us to investigate the rate of convergence of a time-reversible chain by examining the structure of its underlying graph: convergence will usually be rapid if the conductance is not too small. In our applications we will always be dealing with *families* of Markov chains $\mathcal{MC}(x)$ indexed by problem instances x. (Thus x might be a graph and $\mathcal{MC}(x)$ some Markov chain on the set of matchings in the graph.) Let $\tau^{(x)}$ denote the function τ above for the chain $\mathcal{MC}(x)$. Then the rapid mixing property referred to informally earlier requires that $\tau^{(x)}(\varepsilon)$ should be bounded above by a polynomial in the input size $|x|$ and $\lg\varepsilon^{-1}$. This means that the number of steps required to achieve some specified sampling accuracy increases only polynomially with the problem size. As is clear from Corollary 2.3, rapid mixing will usually follow from a lower bound of the form $1/\text{poly}(|x|)$ on the conductance Φ. In this and subsequent sections, we show how to derive such bounds for several interesting chains. Other examples are given in [29].

Remarks. (a) The condition $\min_i p_{ii}\geqq\frac{1}{2}$ is a technical device that simplifies the statement of the theorem by damping oscillatory or "near-periodic" behaviour [29]. Note that an arbitrary Markov chain can be modified to make the condition hold simply by replacing P by $\frac{1}{2}(I+P)$, where I is the $|\mathcal{N}|\times|\mathcal{N}|$ identity matrix. This operation leaves the stationary distribution unchanged and merely reduces the conductance by a factor of $\frac{1}{2}$. (In fact, we have just added a self-loop probability of $\frac{1}{2}$ to each state; for the purposes of practical implementation the waiting time at each state may be simulated more efficiently by a separate random process.)

(b) Theorem 2.2 has a converse stating that, under the same assumptions, $\Delta(t)\geqq(1-2\Phi(H))^t$ [28]. Hence we effectively have a *characterisation* of rapid mixing for a large class of time-reversible chains in terms of the graph-theoretic quantity Φ.

(c) Similar relationships between subdominant eigenvalues of graphs and their structural properties have appeared in the work of Alon [4] and Alon and Milman [5]. The significance of Alon's result as a sufficient condition for rapid mixing for certain Markov chains has been noted by several authors; in particular, Aldous [2] states a restricted version of Theorem 2.2 for random walks on regular graphs. Our *conductance* Φ is a weighted edge analogue of the *magnification* studied in [4], [2] and gives a cleaner and more natural formulation of this connection. Very recently, Lawler and Sokal [22] have independently discovered a characterisation similar to ours but in a rather more general context. □

As it stands, the rapid mixing criterion of Theorem 2.2 is essentially only of theoretical interest. To turn it into a useful practical tool, we need to develop some technology for estimating the conductance of the underlying graphs of natural Markov chains. This we now do with the aid of a simple example.

For a positive integer n, let $B(n)=\{0,1\}^n$ denote the set of bit vectors of length n. Consider the family of Markov chains $\mathcal{MC}(n)$ with state space $B(n)$ and transitions as follows. In any state $v=(v_0,\cdots,v_{n-1})$, select $i\in\{0,\cdots,n-1\}$ uniformly at random and flip the value of the bit v_i. To eliminate periodicity, remain at v with probability $\frac{1}{2}$.

$\mathcal{MC}(n)$ is obviously irreducible and aperiodic, and hence ergodic. Using Lemma 2.1, it is easily verified that $\mathcal{MC}(n)$ is time-reversible and that its stationary distribution is uniform over $B(n)$. The conditions of Theorem 2.2 are satisfied, so the rate of convergence of $\mathcal{MC}(n)$ depends on the conductance of its underlying graph $H(n)$. In $H(n)$, two states are adjacent if and only if they differ in at most one bit. Thus $H(n)$ is just the n-dimensional hypercube, each nonloop edge having weight $(2nN)^{-1}$, where $N=|B(n)|$ is the number of states.

PROPOSITION 2.4. *The conductance of* $H(n)$ *satisfies* $\Phi(H(n))\geqq 1/2n$.

Before presenting the proof of Proposition 2.4, which is the main point of this example, let us note that it implies rapid mixing for the family of chains $\mathcal{MC}(n)$. By Corollary 2.3 the number of simulation steps required to achieve an r.p.d. of ε is $O(n^2(n+\ln\varepsilon^{-1}))$, which is polynomially bounded in n and $\lg\varepsilon^{-1}$. Thus an algorithm that simulates $\mathcal{MC}(n)$ from some arbitrary initial state constitutes an efficient almost uniform sampling procedure for bit strings of length n. (Of course, there are more direct ways of doing this!)

Proof of Proposition 2.4. From the definition (2) we may write

$$\Phi(H(n))=\frac{1}{2n}\min_{0<|S|\leqq N/2}\frac{|\mathrm{cut}\,(S)|}{|S|}\tag{3}$$

where for each $S\subseteq B(n)$, $\mathrm{cut}\,(S)$ denotes the set of cut edges in $H(n)$ defined by S. Our argument hinges on the following observation. Suppose it is possible to specify a canonical simple path in $H(n)$ between each ordered pair of distinct states in such a way that no oriented edge of $H(n)$ is contained in more than bN of the paths. If S is any subset of states with $0<|S|\leqq N/2$, then the number of paths which cross the cut from S to its complement is clearly

$$|S|(N-|S|)\geqq|S|N/2.$$

Thus for any such S the number of cut edges must be at least $|S|N/2bN=|S|/2b$, and so from (3) we have

$$\Phi(H(n))\geqq\frac{1}{4nb}.\tag{4}$$

To get a lower bound on $\Phi(H(n))$ it therefore suffices to define a collection of canonical paths in $H(n)$ that are "sufficiently edge disjoint," as measured by the parameter b.

We now proceed to define a suitable set of paths. Let $u=(u_i)_{i=0}^{n-1}$ and $v=(v_i)_{i=0}^{n-1}$ be distinct elements of $B(n)$, and $i_1<\cdots<i_l$ be the positions in which u and v differ.

Then for $1 \leqq j \leqq l$, the jth edge of the canonical path from u to v corresponds to a transition in which the i_jth bit is flipped from u_{i_j} to v_{i_j}.

Consider now an arbitrary transition t of $\mathcal{MC}(n)$ (or, equivalently, an oriented edge of $H(n)$); our aim is to bound the number of paths that contain t. Suppose that t takes state $w=(w_i)$ to state $w'=(w_i')$ by flipping the value of w_k, and let $P(t)$ denote the set of paths containing t, viewed as ordered pairs of states. Rather than counting elements of $P(t)$ directly, we will set up an *injective* mapping from $P(t)$ into the state space $B(n)$; this will yield an upper bound on the ratio b appearing in (4).

The mapping σ_t: $P(t) \to B(n)$ is defined as follows. Given an ordered pair $\langle u, v\rangle \in P(t)$, set $\sigma_t(u, v) = (s_i)$, where

$$s_i = \begin{cases} u_i, & 0 \leqq i \leqq k, \\ v_i, & k < i < n. \end{cases}$$

Thus $\sigma_t(u, v)$ agrees with u on the first $k+1$ bits and with v on the remainder. Note that we can express this definition more succinctly as $\sigma_t(u, v) = u \oplus v \oplus w'$, where $\oplus$ denotes bitwise exclusive-or.

We claim that $\sigma_t(u, v)$ is an unambiguous *encoding* of the endpoints u and v, so that σ_t is indeed injective. To see this, simply note that

$$u_i = \begin{cases} s_i, & 0 \leqq i \leqq k, \\ w_i, & k < i < n, \end{cases} \qquad v_i = \begin{cases} w_i', & 0 \leqq i \leqq k, \\ s_i, & k < i < n. \end{cases}$$

Hence u and v may be recovered from knowledge of t and $\sigma_t(u, v)$, so σ_t is injective. It follows immediately that $|P(t)| \leqq N$; in fact, since all vectors (s_i) in the range of σ_t satisfy $s_k = w_k$, we have the slightly stronger result that $|P(t)| \leqq N/2$. Since t was chosen arbitrarily, the number of paths traversing *any* oriented edge cannot exceed $N/2$. Thus we may set $b = \frac{1}{2}$ in inequality (4) and deduce the desired bound on the conductance $\Phi(H(n))$. □

Remark. The bound of Proposition 2.4 is tight. To see this, let S be the subset of $B(n)$ consisting of all vectors with first bit 0 and note that $|\text{cut}\,(S)|/|S| = 1$. Hence $\Phi(H(n)) = 1/2n$. □

Some observations on the above proof are in order here. The idea of path counting is quite general and has been used before in the literature to investigate the connectivity properties of various graphs in other contexts (see e.g., [31], in which the hypercube is also studied). The novelty of our proof lies in the use of the injective mapping technique to bound the number of paths which traverse an edge. This is not actually necessary in this simple example as the paths could have been counted explicitly. The point is that in more complex cases the states of the chain will be less trivial structures, such as matchings in a graph, and we will have no useful information about their number—indeed, this is what we will ultimately be trying to compute. It is then crucial to be able to bound the maximum number of paths through any edge in terms of the number of states *without* explicit knowledge of these quantities. This is precisely what the injective mapping technique achieves. As we will see presently, it turns out to be rather generally applicable.

Other simple Markov chains may be analysed in a similar fashion [28]. Examples of rapidly mixing families include random walks on n-dimensional cubes of side d and a host of "card-shuffling" processes whose state space is the set of permutations of n objects and whose transitions correspond to some natural shuffling scheme. These and similar processes have been extensively studied using other methods such as coupling, stopping times, and group representation theory (see [1], [3], [10] for a variety of examples). The time bounds obtained by these methods are generally rather tighter than ours and can often be shown to be optimal. However, the full power of our approach will become apparent in the sequel where it will permit the analysis of highly irregular chains with only a little additional effort. Most significantly, such chains have seemingly not proved amenable to analysis by any of the other established methods.

3. Approximating the permanent.

In this section we consider Broder's method for approximating the permanent of a square 0-1 matrix, as sketched in § 1. Our main result is that the method yields an fpras for a large class of matrices, including all those that are sufficiently dense. Thus Broder's principal claim in [8] turns out to be true despite the fallacious coupling argument given there.

We will work with the perfect matching formulation of the permanent as described in § 1. Let $G=(U, V, E)$ be a bipartite graph with $U=V=\{0, \cdots, n-1\}$, and for $k \in \mathbb{N}$ let $M_k(G)$ denote the set of matchings of size k in G. Thus $M_n(G)$ is the set of perfect matchings in G and its cardinality is equal to the permanent of the $n \times n$ 0-1 matrix associated with G. We assume throughout this section that $M_n(G)$ is nonempty: it is well known that this property can be tested in polynomial time [11].

The method is based on the observation that an fpras for $|M_n(G)|$ can be constructed easily given an efficient procedure for sampling perfect and "near-perfect" matchings in G almost uniformly at random. First we must say more precisely what we mean by such a procedure. Let $\mathcal{N}$ be the set $M_n(G) \cup M_{n-1}(G)$. An *almost uniform generator* for $\mathcal{N}$ is a probabilistic algorithm that, when presented with G and a positive real *bias* ε, outputs an element of $\mathcal{N}$ such that the probability of each element appearing approximates $|\mathcal{N}|^{-1}$ within ratio $1+\varepsilon$. The generator is *fully polynomial* (f.p.) if it runs in time bounded by a polynomial in n and $\lg \varepsilon^{-1}$. (Generators for combinatorial structures are discussed in a more general framework in [16], [28], [29].) Actually, since all the generators we construct in this paper are based on rapidly mixing Markov chains, it should be clear that they embody effective procedures for sampling structures from a given distribution under *any* reasonable definition.

We call the bipartite graph G *dense* if its minimum vertex degree is at least $n/2$. It is shown in [8] that the problem of counting perfect matchings in dense graphs is no easier than counting them in general graphs, and hence is #P-complete. The following result is also proved in [8], and formalises the reduction from approximate counting to almost uniform generation in the case of dense graphs.

THEOREM 3.1 (Broder). *Suppose that, for all dense bipartite graphs G, there exists an f.p. almost uniform generator for $M_n(G) \cup M_{n-1}(G)$. Then there exists an fpras for $|M_n(G)|$ for all such graphs G.* □

Broder investigates the generation problem using a family of ergodic Markov chains $\mathcal{MC}_{\mathrm{pm}}(G)$ with state space $\mathcal{N}=M_n(G) \cup M_{n-1}(G)$ and uniform stationary distribution, in which transitions are made by adding and/or deleting edges locally. We now give a slightly modified definition of $\mathcal{MC}_{\mathrm{pm}}(G)$. View E as a subset of $U \times V$ and matchings in G as subsets of E. If $A, B \subseteq E$ and $e \in E$ then $A \oplus B$ denotes the symmetric difference of A and B, while $A+e$ and $A-e$ denote the sets $A \cup \{e\}$, $A \backslash \{e\}$, respectively. Transitions in $\mathcal{MC}_{\mathrm{pm}}(G)$ are specified as follows. In any state $M \in \mathcal{N}$, choose an edge $e=(u, v) \in E$ uniformly at random and then

(i) If $M \in M_n(G)$ and $e \in M$, move to state $M'=M-e$ (*Type* 1 *transition*);

(ii) If $M \in M_{n-1}(G)$ and u, v are unmatched in M, move to $M'=M+e$ (*Type* 2 *transition*);

(iii) If $M \in M_{n-1}(G)$, u is matched to w in M and v is unmatched in M, move to $M' = (M+e)-(u, w)$; symmetrically, if v is matched to w and u is unmatched, move to $M' = (M+e)-(w, v)$ (*Type* 0 *transition*);

(iv) In all other cases, do nothing.

We again eliminate periodicity by introducing an additional self-loop probability of $\frac{1}{2}$ for each state, i.e., with probability $\frac{1}{2}$ the process does not select a random edge as above but simply remains at M.

Using Lemma 2.1, it is a simple matter to check that $\mathcal{MC}_{\mathrm{pm}}(G)$ is ergodic and time-reversible with uniform stationary distribution. What is not at all obvious is that the family of chains is rapidly mixing. This surprising fact is a consequence of the following theorem.

THEOREM 3.2. *Let G be dense and H be the underlying graph of the Markov chain $\mathcal{MC}_{\mathrm{pm}}(G)$. Then $\Phi(H) \geqq 1/12n^6$.*

We shall prove the theorem in a moment after examining its implications.

COROLLARY 3.3. *There exists an f.p. almost uniform generator for $M_n(G) \cup M_{n-1}(G)$ in all dense bipartite graphs G.*

Proof. On input $\langle G, \varepsilon\rangle$, the generator deterministically constructs an initial state of $\mathcal{MC}_{\mathrm{pm}}(G)$ and then simulates the chain for some number $T \geqq \tau(\varepsilon/2)$ of steps, outputting the final state. (Assuming as we may that $\varepsilon \leqq 1$, an r.p.d. of $\varepsilon/2$ guarantees a bias of at most ε.) To see that the generator is f.p., note that for any G individual steps of $\mathcal{MC}_{\mathrm{pm}}(G)$ can be simulated, and a perfect matching to serve as initial state found, in polynomial time. Moreover, since $\pi_{\min}^{-1} = |\mathcal{N}|$ is certainly bounded above by 2^{n^2}, Theorem 3.2 and Corollary 2.3 imply that $\tau(\varepsilon/2) \leqq \mathrm{poly}\,(n, \lg \varepsilon^{-1})$, so T need only be this large. Hence the overall runtime is bounded as required. □

Combining this with Theorem 3.1 immediately yields the following Corollary.

COROLLARY 3.4. *There exists an fpras for $|M_n(G)|$ in all dense bipartite graphs G, and hence for the permanent of all dense square* 0–1 *matrices.* □

We return now to the proof of the key result above.

Proof of Theorem 3.2. As in (3) the conductance is given by

$$\Phi(H) = \frac{1}{2|E|} \min_{0<|S| \leqq |\mathcal{N}|/2} \frac{|\mathrm{cut}\,(S)|}{|S|} \tag{5}$$

where again cut (S) denotes the set of cut edges in H defined by S. We will proceed as in the proof of Proposition 2.4 by defining a set of canonical paths in H. If no transition occurs in more than $b|\mathcal{N}|$ of these, by analogy with (4) we will have the bound

$$\Phi(H) \geqq \frac{1}{4b|E|} \geqq \frac{1}{4bn^2}. \tag{6}$$

We begin by specifying, for each $M \in \mathcal{N}$, canonical paths to and from a unique "closest" perfect matching $\bar{M} \in M_n(G)$ as follows, where $u \in U$ and $v \in V$ denote the unmatched vertices (if any) of M:

(i) If $M \in M_n(G)$ then $\bar{M} = M$ and the path is empty;

(ii) If $M \in M_{n-1}(G)$ and $(u, v) \in E$, then $\bar{M} = M + e$ and the path consists of a single Type 2 transition;

(iii) If $M \in M_{n-1}(G)$ and $(u, v) \notin E$, fix some $(u', v') \in M$ such that $(u, v'), (u', v) \in E$: note that at least one such edge must exist by the density assumption on G. Then $\bar{M} = (M-(u', v')) + (u, v') + (u', v)$, and we specify one of the two possible paths of length two from M to $\bar{M}$, involving a Type 0 transition followed by a Type 2 transition.

The canonical path from $\bar{M}$ to M consists of the same edges of H traversed in the opposite direction.

For future reference, we observe that no perfect matching is involved in too many canonical paths of the above form: for $M \in M_n(G)$ define the set

$$\mathcal{K}(M) = \{M' \in \mathcal{N}: \bar{M}' = M\}.$$

Then, since each matching in $\mathcal{K}(M)$ has at least $n-2$ edges in common with M, it is easy to see that $|\mathcal{K}(M)| \leqq n^2$. Note that the sets $\mathcal{K}(M)$ partition $\mathcal{N}$, implying that $|\mathcal{N}| \leqq n^2 |M_n(G)|$. It is also worth noting that this is the only point in the proof at which the bipartite structure of G is used: we will have more to say about this later (see remark (d) below).

Next we define a canonical path in H between an ordered pair I, F of *perfect* matchings (refer to Fig. 1(a)). To do this, we first assume a fixed ordering of all even cycles of G, and distinguish in each cycle a *start vertex* in U. Now consider the symmetric difference $I \oplus F$; we may write this as a sequence $C_1, \cdots, C_r$ of disjoint even cycles, each of length at least four, where the indices respect the above ordering. The path from I to F involves *unwinding* each of the cycles $C_1, \cdots, C_r$ in turn in the following way. Suppose the cycle C_i has start vertex u_0 and consists of the sequence of distinct vertices $(u_0, v_0, u_1, v_1, \cdots, u_l, v_l)$, where $(u_j, v_j) \in I$ for $0 \leqq j \leqq l$ and the remaining edges are in F. Then the first step in the unwinding of C_i is a Type 1 transition that removes the edge (u_0, v_0). This is followed by a sequence of l Type 0 transitions, the jth of which replaces the edge (u_j, v_j) by (u_j, v_{j-1}). The unwinding is completed by a Type 2 transition that adds the edge (u_0, v_l).

Finally, the canonical path between any pair of matchings $I, F \in \mathcal{N}$ is defined as the concatenation of three segments as follows:

initial segment: follow the canonical path from I to $\bar{I}$,
main segment: follow the canonical path from $\bar{I}$ to $\bar{F}$,
final segment: follow the canonical path from $\bar{F}$ to F.

Now consider an arbitrary oriented edge of H, corresponding to a transition t in the Markov chain. We aim to establish an upper bound of the form $b|\mathcal{N}|$ on the number of canonical paths that contain this transition. Suppose first that t occurs in the *initial* segment of a path from I to F, where $I, F \in \mathcal{N}$. Then it is clear from the definition of initial segment that the perfect matching $\bar{I}$ is uniquely determined by t. But we have already seen that $|\mathcal{K}(\bar{I})| \leqq n^2$. Since $I \in \mathcal{K}(\bar{I})$, the number of paths that contain t in their initial segment is thus at most $n^2|\mathcal{N}|$. A symmetrical argument shows that the number of paths containing t in their final segment is similarly bounded.

To handle the main segments of the paths, we make use of the injective mapping technique seen in the proof of Proposition 2.4. This will obviate the need for any explicit counting of structures in $\mathcal{N}$, which is crucial here (cf. the discussion following the proof of Proposition 2.4). Let t be a transition from M to M', where $M, M' \in \mathcal{N}$ are distinct, and denote by $P(t)$ the set of ordered pairs $\langle I, F\rangle$ of *perfect* matchings such that t is contained in the canonical path from I to F. We proceed to define, for each pair $\langle I, F\rangle \in P(t)$, an encoding $\sigma_t(I, F) \in \mathcal{N}$ from which I and F can be uniquely reconstructed. The intention is that, if $C_1, \cdots, C_r$ is the ordered sequence of cycles in $I \oplus F$, and t is traversed during the unwinding of C_i, then the encoding should agree with I on $C_1, \cdots, C_{i-1}$ and on that portion of C_i that has already been unwound, and with F elsewhere.

With this in mind, consider the set $S = I \oplus F \oplus (M \cup M')$. Since $I \cap F \subseteq M \cup M' \subseteq I \cup F$ and $|I| = |F| = |M \cup M'| = n$, elementary set theory tells us that $|S| = n$. Furthermore, suppose that some vertex u is adjacent to two edges in S. Then both these edges necessarily lie in $I \oplus F$, which in turn implies that neither edge lies in $M \cup M'$. Hence

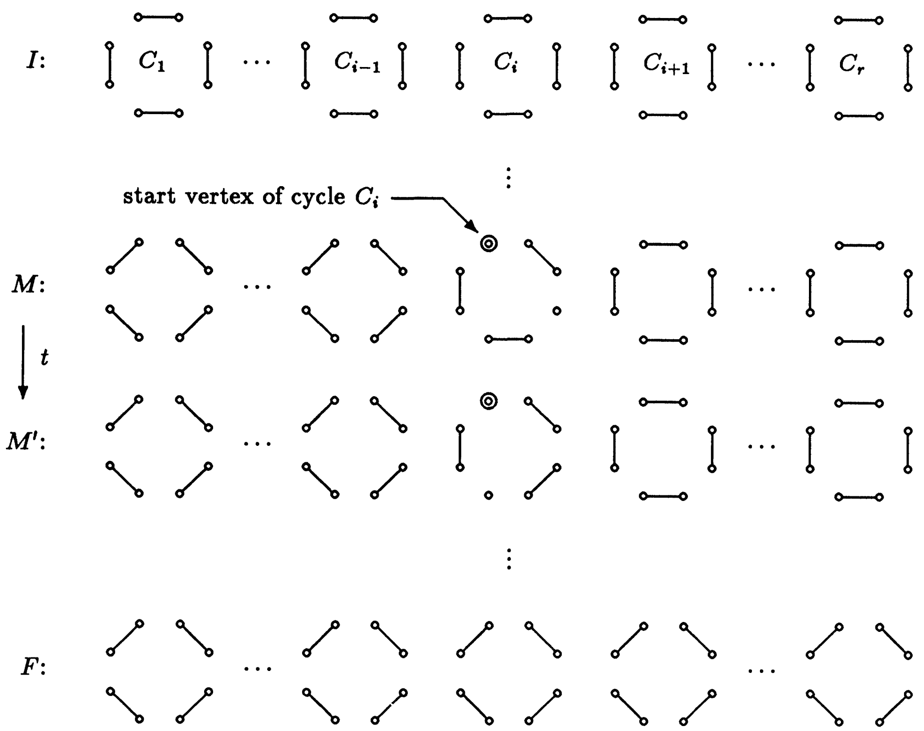

FIG. 1(a). *A transition t on the canonical path from I to F.*

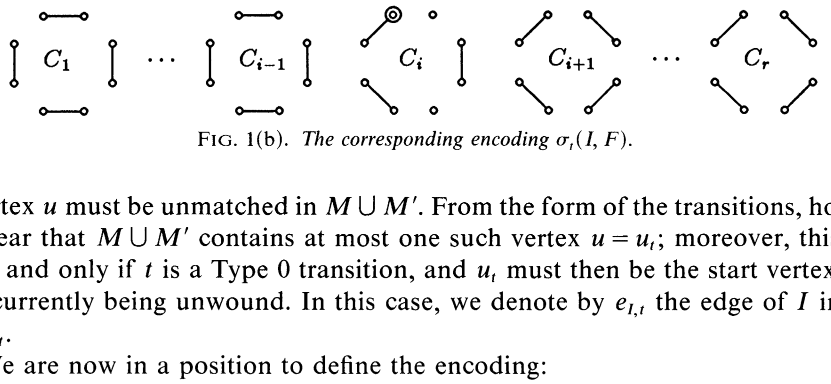

FIG. 1(b). *The corresponding encoding $\sigma_t(I, F)$.*

the vertex u must be unmatched in $M \cup M'$. From the form of the transitions, however, it is clear that $M \cup M'$ contains at most one such vertex $u = u_t$; moreover, this is the case if and only if t is a Type 0 transition, and u_t must then be the start vertex of the cycle currently being unwound. In this case, we denote by $e_{I,t}$ the edge of I incident with u_t.

We are now in a position to define the encoding:

$$\sigma_t(I, F) = \begin{cases} (I \oplus F \oplus (M \cup M')) - e_{I,t} & \text{if } t \text{ is Type 0}, \\ I \oplus F \oplus (M \cup M') & \text{otherwise}. \end{cases}$$

Figure 1(b) illustrates this definition for a Type 0 transition. In view of the above discussion, $\sigma_t(I, F)$ is always a matching of cardinality at least $n-1$, and hence an element of $\mathcal{N}$. It remains for us to show that I and F can be recovered from it.

First observe that $I \oplus F$ can be recovered immediately using the relation

$$I \oplus F = \begin{cases} (\sigma_t(I, F) \oplus (M \cup M')) + e_{I,t} & \text{if } t \text{ is Type 0}, \\ \sigma_t(I, F) \oplus (M \cup M') & \text{otherwise}. \end{cases}$$

(Note that $e_{I,t}$ is the unique edge that must be added to $\sigma_t(I, F) \oplus (M \cup M')$ to ensure that $I \oplus F$ is a union of disjoint cycles.) Thus we may infer the ordered sequence

$C_1, \cdots, C_r$ of cycles to be unwound on the path from I to F. The cycle C_i that is currently being unwound, together with its parity with respect to I and F, is then determined by the transition t. The parity of all remaining cycles may be deduced from M and the cycle ordering. Finally, the remaining portions of I and F may be recovered using the fact that $I \cap F = M \setminus (I \oplus F)$. Hence $\sigma_t(I, F)$ uniquely determines the pair $\langle I, F \rangle$, so σ_t is an injective mapping from $P(t)$ to $\mathcal{N}$.

The existence of σ_t ensures that $|P(t)| \leqq |\mathcal{N}|$ for any transition t. Since also $|\mathcal{K}(M)| \leqq n^2$ for any perfect matching M, we see that t is contained in the main segment of at most $n^4|\mathcal{N}|$ paths. Combining this with the results for initial and final segments derived earlier, we deduce that the maximum total number of paths that contain t is bounded by

$$(n^2 + n^2 + n^4)|\mathcal{N}| \leqq 3n^4|\mathcal{N}|.$$

To complete the proof we set $b = 3n^4$ in (6). □

Remarks. (a) In his original paper [8], Broder claimed that the above rapid mixing property holds under the same density assumption. However, as indicated in § 1, his proof based on coupling ideas is both complex and fundamentally flawed. The problem is that the "coupling" defined in [8] is not, in fact, a coupling because one of the two processes involved is not a faithful copy of $\mathcal{MC}_{\text{pm}}(G)$: this is explained in detail by Mihail [25]. As a result Broder has withdrawn his proof (see the erratum to [8]). We feel that this is compelling evidence of the unsuitability of coupling and related methods for the analysis of Markov chains that lack a high degree of symmetry.

(b) A chain with larger conductance is obtained by modifying $\mathcal{MC}_{\text{pm}}(G)$ slightly so that transitions are effected by selecting a random vertex in V rather than a random edge. This increases the transition probability in (5) from $1/2|E|$ to $1/2n$, saving a factor of n in the conductance bound.

(c) The f.p. almost uniform generator for $M_n(G) \cup M_{n-1}(G)$ may be adapted to one for *perfect* matchings in dense graphs G, by repeatedly generating elements of $M_n(G) \cup M_{n-1}(G)$ and outputting the first perfect matching which occurs. Since $|M_n(G)| \geqq n^{-2}|\mathcal{N}|$ we should not have to wait too long, but if so some arbitrary perfect matching may be output without affecting the bias too much. Among other things, with appropriate choice of G this provides a way of generating certain natural restricted classes of permutations that satisfy the density condition, such as displacements or ménage arrangements.

(d) So far we have concentrated exclusively on bipartite graphs because of their connection with the permanent. The Markov chain $\mathcal{MC}_{\text{pm}}(G)$ can be applied without essential modification to arbitrary graphs G. In fact, the only point at which we have relied on the bipartite structure of G is in the definition of the sets $\mathcal{K}(M)$ in the proof of Theorem 3.2 and the bound on their size. Let $G = (V, E)$ be an arbitrary graph with $|V| = 2n$. As before, we assume that G contains a perfect matching. Call G *dense* if its minimum vertex degree is at least n. This ensures that $\mathcal{K}(M)$ for $M \in M_n(G)$ is still well defined, and that $|\mathcal{K}(M)| \leqq 2n^2$. The rest of the proof carries through as before, yielding $b = 8n^4$ and consequently $\Phi(H) \geqq 1/64n^6$. (This can again be improved if transitions are implemented by random vertex selection.) Since a construction analogous to that of Theorem 3.1 holds for general dense graphs, we have:

COROLLARY 3.5. *There exists an fpras for* $|M_n(G)|$ *in arbitrary dense graphs.* □

We conclude this section by examining the role played in our results by the density assumption. In the proof of Theorem 3.2 it was used to show that each element of $M_{n-1}(G)$ is "close to" a perfect matching, so that near-perfect matchings could effectively be ignored in the conductance argument. In fact, it is enough to know that the total *number* of near-perfect matchings is not too large. More specifically, the above results hold under the considerably weaker assumption that

$$|M_{n-1}(G)|/|M_n(G)| \leqq q(n) \tag{7}$$

for some fixed polynomial q. (Recall that we are assuming $|M_n(G)|>0$.) This condition will arise more naturally in the context of the improved algorithm of § 5. Accordingly, we only sketch the proof modifications necessary to extend the method of this section. A fuller account can be found in [28].

First, it is not hard to see that the reduction of Theorem 3.1 still holds for graphs G satisfying the weaker condition (7). (This fact relies on Theorem 5.1 of § 5.) It is therefore enough to generate elements of $\mathcal{N} = M_n(G) \cup M_{n-1}(G)$ using the Markov chain $\mathcal{MC}_{\mathrm{pm}}(G)$: the only thing we have to check is that its conductance remains bounded below by an inverse polynomial function of n. This is a consequence of the following generalised version of Theorem 3.2.

THEOREM 3.6. *For any graph $G=(V, E)$ with $|V|=2n$ and $|M_n(G)|>0$, the conductance of the underlying graph of $\mathcal{MC}_{\mathrm{pm}}(G)$ is bounded below by*

$$\frac{1}{16|E|}\left(\frac{|M_n(G)|}{|M_{n-1}(G)|}\right)^2.$$

Sketch of proof. Let H be the underlying graph of $\mathcal{MC}_{\mathrm{pm}}(G)$. The proof differs from that of Theorem 3.2 in one or two details. First, we use a variant of the canonical path counting argument in which only paths from perfect to near-perfect matchings are considered. If these can be defined in such a way that no edge of H carries more than $b|\mathcal{N}|$ of them, then a little algebra yields (cf. (6))

$$\Phi(H) \geqq \frac{1}{16b|E|}\left(\frac{|M_n(G)|}{|M_{n-1}(G)|}\right). \tag{8}$$

The paths themselves are similar to those between perfect matchings in the proof of Theorem 3.2: if $I \in M_n(G)$ and $F \in M_{n-1}(G)$, the symmetric difference $I \oplus F$ consists of a sequence $C_1, \cdots, C_r$ of disjoint cycles as before, together with a single open path O that is unwound in the obvious way *after* all the C_i.

Let t be an arbitrary transition. The injective mapping technique can again be used to bound the cardinality of the set $P(t)$ of paths of the above kind that involve t. For $\langle I, F\rangle \in P(t)$ the encoding $\sigma_t(I, F)$ is defined exactly as before, and σ_t is again injective. Now, however, we find that $\sigma_t(I, F) \in M_{n-1}(G) \cup M_{n-2}(G)$, as we get an additional pair of unmatched vertices arising from the open path O. (Note that this time the encoding takes us outside the state space.) Since σ_t is injective we have

$$|P(t)| \leqq |M_{n-1}(G)| + |M_{n-2}(G)| \leqq \left(\frac{|M_{n-1}(G)|}{|M_n(G)|}\right)|\mathcal{N}|,$$

where the second inequality again appeals to Theorem 5.1. Thus we may take $b = |M_{n-1}(G)|/|M_n(G)|$ in (8), completing the proof of the theorem. □

COROLLARY 3.7. *Let q be any fixed polynomial. There exists an fpras for $|M_n(G)|$ in all $2n$-vertex graphs G that satisfy $|M_{n-1}(G)|/|M_n(G)| \leqq q(n)$.* □

Our earlier results for dense graphs can be derived as a special case of Corollary 3.7 with $q(n)=O(n^2)$. Note that the density bound quoted is tight in the sense that it is possible to construct, for any fixed $\delta>0$, a sequence of (bipartite) graphs (G_n) with $2n$ vertices and minimum vertex degree at least $n/(2+\delta)$ such that the ratio $|M_{n-1}(G_n)|/|M_n(G_n)|$ is exponentially large.

Remark. Dagum et al. [9] have shown that the reduction from counting to generation of Theorem 3.1 may be replaced by the following mechanism, with a small increase in efficiency. In analogous fashion to $\mathcal{MC}_{\mathrm{pm}}(G)$, we may define for $1 \leqq k \leqq n$ a Markov chain $\mathcal{MC}_k(G)$ whose states are k- and $(k-1)$-matchings in G. (Thus $\mathcal{MC}_n(G)$ is just $\mathcal{MC}_{\mathrm{pm}}(G)$.) By a simple extension of the proof of Theorem 3.6, whereby multiple rather than unique canonical paths between states are counted, it can be shown that each of the chains $\mathcal{MC}_k(G)$ is rapidly mixing under the same condition (7) on G. This allows the ratios $|M_k(G)|/|M_{k-1}(G)|$ to be estimated directly for each k in turn. We do not dwell on this point here as we will present a more natural algorithm in § 5. □

This concludes our discussion of perfect matchings for now. An alternative view of all these results will emerge as a by-product of our work on a different problem in the next section.

4. Monomer–dimer systems. This section is concerned with counting and generating at random all matchings (independent sets of edges) in a graph. Apart from their inherent interest, these problems arise in the theory of statistical physics, a rich source of combinatorial counting and generation problems.

A *monomer-dimer system* consists of a graph $G=(V, E)$, which is usually some form of regular lattice, together with a positive weight on each edge. The vertices of G represent physical sites, adjacent pairs of which may be occupied by diatomic molecules or *dimers*. Configurations of the system correspond to arrangements of dimers on the lattice in which no two dimers overlap, i.e., to matchings in G. In a configuration consisting of fewer than $|V|/2$ dimers, unoccupied sites are referred to as *monomers*. Monomer-dimer systems have been extensively studied as models of physical systems involving diatomic molecules. In the two-dimensional case they model the adsorption of dimers on the surface of a crystal. Three-dimensional systems occur in the theory of mixtures of molecules of different sizes and in the cell-cluster theory of the liquid state. For further information, see [14] and the references given there.

For each edge e of G, the weight $c(e)$ represents the relative probability of occupation by a dimer. This will depend on the contribution of such a dimer to the global energy of the system. Most thermodynamic properties of the system can be deduced from knowledge of the *partition function*

$$Z(G)=\sum_{M \in M_*(G)} W(G, M) \tag{9}$$

where $M_*(G)$ is the set of configurations (matchings in G) and $W(G, M)=\prod_{e \in M} c(e)$ is the *weight* of the configuration M. Counting matchings, i.e., computing (9) in the special case where all edge weights are unity, is a #P-complete problem even when restricted to planar graphs [15], [33]. The main result of this section is that the more general sum (9) can in fact be *approximated* efficiently for any weighted graph G.

We will proceed as in the previous section via a related random generation problem for configurations. Since the sum in (9) is weighted, however, configurations should be generated not uniformly but with probabilities proportional to their weights. In fact, this problem is of interest in its own right as a means of estimating the expectation of various physical operators on configurations by the so-called Monte Carlo method [6].

The notion of an almost uniform generator can be generalised to the weighted case in the obvious way. We will call a probabilistic algorithm an *almost W-generator* for matchings if, given a graph G with positive edge weights and a positive real bias $\varepsilon>0$, it outputs a matching M in G with probability that approximates

$W(G, M)/Z(G)$ within ratio $1+\varepsilon$. As usual, the generator is f.p. if its runtime is bounded by a polynomial in the size of the input G and in $\lg \varepsilon^{-1}$.

The problem of approximating the partition function (9) is reduced to the weighted generation problem as follows. Let $e=(u, v)$ be any edge of the weighted graph $G=(V, E)$, G^- the graph obtained by removing e from G, and G^+ the graph obtained by removing the vertices u and v together with all their incident edges. By partitioning matchings in G into two sets according to whether they do or do not contain e, it is readily seen that

(i) There is a (1-1)-correspondence between $M_*(G)$ and the disjoint union of $M_*(G^+)$ and $M_*(G^-)$;

(ii) $Z(G)=c(e)Z(G^+)+Z(G^-)$.

This suggests the following recursive procedure for estimating $Z(G)$:

(1) Using an almost W-generator, construct an independent sample of elements of $M_*(G)$ as detailed below.

(2) For some edge e of G, let z^+, z^- denote the proportions of elements in the sample that do and do not contain e, respectively. Note that these quantities estimate $c(e)Z(G^+)/Z(G)$ and $Z(G^-)/Z(G)$, respectively.

(3) If $z^+ \geqq z^-$, recursively estimate $Z(G^+)$ and multiply the result by $c(e)/z^+$; otherwise, recursively estimate $Z(G^-)$ and multiply by $1/z^-$.

The procedure terminates when the input graph contains no edges. Note that the choice of the larger ratio in step (3) maximises the accuracy of the method.

Some elementary but tedious statistics (see [16, Thm. 6.4]) confirms that, if the bias in the generator is set to $\varepsilon/\alpha|E|$ for some constant α, the sample size required in step (1) to ensure that the final answer approximates $Z(G)$ within ratio $1+\varepsilon$ with probability at least $\frac{3}{4}$ is only $O(|E|^3\varepsilon^{-2})$. We therefore have the following theorem.

THEOREM 4.1. *Suppose there exists an f.p. almost W-generator for matchings. Then there exists an fpras for the partition function of monomer-dimer systems.* □

Remark. The foregoing is an example of a more general reduction from approximate counting to random generation, justified in detail in [16], [28], that applies to all structures that are *self-reducible.* Informally, this means that the set of structures corresponding to any problem instance is in (1-1)-correspondence with the disjoint union of sets corresponding to a few smaller problem instances (subproblems). In the case of matchings this property is expressed in condition (i) above. Since we are dealing here with *weighted* combinatorial sums, we need to supplement the definition of self-reducibility by demanding that any sum can be computed easily given the sums for its subproblems: condition (ii) states this for the monomer-dimer partition function. This implies that the Markov chain approach to random generation studied in this paper is potentially a powerful *general* method for approximating hard combinatorial counting problems. (Note that in the previous section it was necessary to resort to the specialised reduction provided by Theorem 3.1. The reason is that, while perfect matchings are easily seen to be self-reducible in general, this property is apparently destroyed when restrictions are placed on the input graph as in § 3.) □

Theorem 4.1 says that we will get an fpras for the monomer-dimer partition function provided we can efficiently generate matchings with probabilities roughly proportional to their weights. This we achieve by simulating a Markov chain in the style of Metropolis et al. [6]. Given a graph $G=(V, E)$ with positive edge weights $\{c(e): e \in E\}$, we consider the chain $\mathcal{MC}_{\mathrm{md}}(G)$ with state space $\mathcal{N}=M_*(G)$ and transitions as follows. In any state $M \in \mathcal{N}$, choose an edge $e=(u, v) \in E$ uniformly at random and then

(i) If $e \in M$, move to $M-e$ with probability $1/(1+c(e))$ (*Type* 1 *transition*);

(ii) If u and v are both unmatched in M, move to $M+e$ with probability $c(e)/(1+c(e))$ (*Type* 2 *transition*);

(iii) If $e'=(u,w)\in M$ for some w, and v is unmatched in M, move to $(M+e)-e'$ with probability $c(e)/(c(e)+c(e'))$ (*Type* 0 *transition*);

(iv) In all other cases, do nothing.

As always, we simplify matters by adding a self-loop probability of $\frac{1}{2}$ to each state. It is then readily checked that $\mathcal{MC}_{\text{md}}(G)$ is irreducible and aperiodic, and hence ergodic. The stationary probability π_M of $M\in M_*(G)$ is easily seen, by Lemma 2.1, to be proportional to its weight $W(G,M)=\prod_{e\in M}c(e)$, so simulation of the chain will yield an f.p. almost W-generator for matchings provided the family $\mathcal{MC}_{\text{md}}(G)$ is rapidly mixing. Now $\mathcal{MC}_{\text{md}}(G)$ is clearly time-reversible by virtue of the detailed balance condition (1), so we may again apply Theorem 2.2. The crucial fact is the following.

THEOREM 4.2. *For a graph $G=(V,E)$ with positive edge weights $\{c(e):e\in E\}$, the conductance of the underlying graph of the Markov chain $\mathcal{MC}_{\text{md}}(G)$ is bounded below by $1/(8|E|c_{\max}^2)$, where $c_{\max}=\max\{1,\max_{e\in E}c(e)\}$.*

Proof. Let H be the underlying graph of $\mathcal{MC}_{\text{md}}(G)$. The first step is to establish a weighted version of the path counting argument that led to the bound (4). Suppose that between each ordered pair $\langle I,F\rangle$ of distinct states we have a canonical path in H, and let us associate with the path a weight $\pi_I\pi_F$. Also, for any subset S of states define

$$C_S=\sum_{M\in S}\pi_M \quad \text{the } \textit{capacity} \text{ of } S,$$

$$F_S=\sum_{M\in S,M'\notin S}\pi_M p_{MM'} \quad \text{the } \textit{ergodic flow} \text{ out of } S.$$

($p_{MM'}$ is the transition probability from M to M'.) Note that the conductance $\Phi(H)$ is just the minimum value of the ratio F_S/C_S over subsets S with $0<C_S\leqq\frac{1}{2}$. For any such S, the aggregated *weight* of all paths crossing the cut from S to its complement $\bar{S}$ in $\mathcal{N}$ is

$$\sum_{I\in S,F\in\bar{S}}\pi_I\pi_F=C_SC_{\bar{S}}\geqq\frac{C_S}{2}. \tag{10}$$

Now let t be a transition from a state M to a state $M'\neq M$, and denote by $P(t)$ the set of all ordered pairs $\langle I,F\rangle$ whose canonical path contains t. Suppose it is known that, for any such transition t, the aggregated weight of paths containing t satisfies

$$\sum_{\langle I,F\rangle\in P(t)}\pi_I\pi_F\leqq bw_t \tag{11}$$

where $w_t=\pi_M p_{MM'}=\pi_{M'}p_{M'M}$ is the weight of the edge in H corresponding to t. Taking (10) and (11) together, we have the following bound on the ergodic flow out of S, where cut(S) denotes the set of transitions crossing the cut from S to $\bar{S}$:

$$\begin{aligned}F_S=\sum_{t\in\text{cut}(S)}w_t&\geqq b^{-1}\sum_{t\in\text{cut}(S)}\sum_{\langle I,F\rangle\in P(t)}\pi_I\pi_F\\&\geqq b^{-1}\sum_{I\in S,F\in\bar{S}}\pi_I\pi_F\\&\geqq\frac{C_S}{2b}.\end{aligned}$$

By definition, the conductance of H therefore satisfies

$$\Phi(H)\geqq\frac{1}{2b}. \tag{12}$$

Our aim is thus to define a set of paths obeying a suitable bound b in (11).

To do this we generalise the proof of Theorem 3.2. Suppose there is an underlying order on all simple paths in G and designate in each of them a start vertex, which must be an endpoint if the path is not a cycle but is arbitrary otherwise. For distinct $I, F \in \mathcal{N}$, we can write the symmetric difference $I \oplus F$ as a sequence $Q_1, \cdots, Q_r$ of disjoint paths that respects the ordering. The canonical path from I to F involves unwinding each of the Q_i in turn as follows. There are two cases to consider:

Case i. *Q_i is not a cycle.* Let Q_i consist of the sequence $(v_0, v_1, \cdots, v_l)$ of vertices, with v_0 the start vertex. If $(v_0, v_1) \in F$, perform a sequence of Type 0 transitions replacing (v_{2j+1}, v_{2j+2}) by (v_{2j}, v_{2j+1}) for $j = 0, 1, \cdots$, and finish with a single Type 2 transition if l is odd. If on the other hand $(v_0, v_1) \in I$, begin with a Type 1 transition removing (v_0, v_1) and proceed as before for the reduced path $(v_1, \cdots, v_l)$.

Case ii. *Q_i is a cycle.* Let Q_i consist of the sequence $(v_0, v_1, \cdots, v_{2l+1})$ of vertices, where v_0 is the start vertex, $l \geqq 1$ and $(v_{2j}, v_{2j+1}) \in I$ for $0 \leqq j \leqq l$, the remaining edges belonging to F. Then the unwinding begins with a Type 1 transition to remove (v_0, v_1). We are left with an open path O with endpoints v_0, v_1, one of which must be the start vertex of O. Suppose v_k, $k \in \{0, 1\}$, is *not* the start vertex. Then we unwind O as in Case (i) above but treating v_k as the start vertex. This trick serves to distinguish cycles from open paths, as will prove convenient shortly.

Now let t be a transition from M to $M' \neq M$. The next step is to define our injective mapping $\sigma_t : P(t) \to \mathcal{N}$. As in the proof of Theorem 3.2, we set $\sigma_t(I, F)$ equal to $I \oplus F \oplus (M \cup M')$, and remove the edge $e_{I,t}$ of I adjacent to the start vertex of the path currently being unwound if necessary: this is so if and only if the path is a cycle and t is Type 0. It is now easily seen that $\sigma_t(I, F)$ consists of independent edges, and so is an element of $\mathcal{N}$. The difference $I \oplus F$ can be recovered from $\sigma_t(I, F)$ using the relation

$$I \oplus F = \begin{cases} (\sigma_t(I, F) \oplus (M \cup M')) + e_{I,t} & \text{if } t \text{ is Type 0 and the current path is a cycle;} \\ \sigma_t(I, F) \oplus (M \cup M') & \text{otherwise.} \end{cases}$$

Note that we can tell whether the current path is a cycle from the sense of unwinding. Recovery of I and F themselves now follows as before from the path ordering. Hence σ_t is injective.

Moreover, it should be clear that $\sigma_t(I, F)$ is very nearly the complement of M in the union of I and F viewed as a multiset, so that the product $\pi_I \pi_F$ is approximately equal to $\pi_M \pi_{\sigma_t(I,F)}$, giving us a handle on b in (11). We now make this precise.

CLAIM. *For any $\langle I, F \rangle \in P(t)$, we have*

$$\pi_I \pi_F \leqq 4|E| c_{\max}^2 w_t \pi_{\sigma_t(I,F)}. \tag{13}$$

The claim will be proved in a moment. First note that it immediately yields the desired bound b in (11), since for any transition t we have

$$\sum_{\langle I,F \rangle \in P(t)} \pi_I \pi_F \leqq 4|E| c_{\max}^2 w_t \sum_{\langle I,F \rangle \in P(t)} \pi_{\sigma_t(I,F)} \leqq 4|E| c_{\max}^2 w_t$$

where the second inequality follows from the fact that σ_t is injective. We may therefore take $b = 4|E| c_{\max}^2$, which in light of (12) gives the conductance bound stated in the theorem.

It remains only for us to prove the claim. We distinguish three cases:

Case i. *t is a Type* 1 *transition.* Suppose $M' = M - e$. Then $\sigma_t(I, F) = I \oplus F \oplus M$, so, viewed as multisets, $M \cup \sigma_t(I, F)$ and $I \cup F$ are equal. Hence we have

$$\begin{aligned} \pi_I \pi_F &= \pi_M \pi_{\sigma_t(I,F)} \\ &= (w_t / p_{MM'}) \pi_{\sigma_t(I,F)} \\ &= 2|E|(1 + c(e)) w_t \pi_{\sigma_t(I,F)}, \end{aligned}$$

from which (13) follows.

Case ii. *t is a Type* 2 *transition.* This is handled by a symmetrical argument to Case (i) above, with M replaced by M'.

Case iii. *t is a Type* 0 *transition.* Suppose $M' = (M+e) - e'$, and consider the multiset $\sigma_t(I, F) \cup M$. This is equal to the multiset $I \cup F$ except that the edge e, and possibly also the edge $e_{I,t}$ are absent from it. Assuming $e_{I,t}$ is absent, which happens precisely when the current path is a cycle, we have

$$\begin{aligned}\pi_I \pi_F &= c(e_{I,t}) c(e) \pi_M \pi_{\sigma_t(I,F)} \\ &= c(e_{I,t}) c(e) (w_t / p_{MM'}) \pi_{\sigma_t(I,F)} \\ &= 2|E| c(e_{I,t}) (c(e) + c(e')) w_t \pi_{\sigma_t(I,F)},\end{aligned}$$

again satisfying (13). If $e_{I,t}$ is not absent, the argument is identical with the factor $c(e_{I,t})$ omitted.

This concludes the proof of the claim and the theorem. □

COROLLARY 4.3. *There exists an f.p. almost W-generator for matchings in arbitrary weighted graphs provided the edge weights are positive and presented in unary.*

Proof. Define $c_{\min} = \min\{1, \min_{e \in E} c(e)\}$. Then the minimum stationary state probability in $\mathcal{MC}_{\mathrm{md}}(G)$ is at least $c^n_{\min} 2^{-|E|} c^{-n}_{\max}$, where $n = |V|$. The logarithm of this quantity is at least $-p(|G|)$, where $|G|$ is the size of the description of G and p is a polynomial. Hence by Theorems 2.2 and 4.2, the Markov chains are rapidly mixing. Simulation of $\mathcal{MC}_{\mathrm{md}}(G)$ is a simple matter, starting from the empty matching. □

In view of Theorem 4.1, we may now state the main result of this section.

COROLLARY 4.4. *There exists an fpras for the monomer-dimer partition function of arbitrary weighted graphs with edge weights presented in unary.* □

COROLLARY 4.5. *There exists an fpras for the number of matchings in arbitrary graphs.* □

5. Some applications: The permanent revisited. As we have already mentioned, our analysis of the monomer-dimer Markov chain $\mathcal{MC}_{\mathrm{md}}(G)$ sheds new light on the results of § 3. In this section we will demonstrate that it yields a more natural approximation algorithm for counting perfect matchings in $2n$-vertex graphs G for which the ratio $|M_{n-1}(G)|/|M_n(G)|$ is polynomially bounded, and in addition allows this condition to be probabilistically tested for an arbitrary graph in polynomial time. We will also discuss an application of the chain to finding a maximum matching in a graph by simulated annealing. The key to all these algorithms is the introduction of carefully chosen edge weights.

The results of this section (and indeed Corollary 3.7 of § 3) depend crucially on the following property of matchings. As usual, let $M_k(G)$ denote the set of k-matchings in a graph G.

THEOREM 5.1. *For any graph G, the sequence $\{|M_k(G)|: k \in \mathbb{N}\}$ is log-concave, i.e.,*

$$|M_{k+1}(G)|\,|M_{k-1}(G)| \leqq |M_k(G)|^2 \quad \forall k \in \mathbb{N}^+.$$

Proof. A proof that relies on machinery from complex analysis can be found in Theorem 7.1 of [14] (see also [23, Exercise 8.5.10]). We present an elementary combinatorial proof that uses ideas seen elsewhere in this paper. Since log-concavity results in combinatorics tend to be rather hard to come by, we believe the simpler proof to be of independent interest.

Let $k \in \mathbb{N}^+$. We may assume that $|M_{k+1}(G)| > 0$, since the inequality is trivially true otherwise. Define the sets $A = M_{k+1}(G) \times M_{k-1}(G)$ and $B = M_k(G) \times M_k(G)$. Our aim is to show that $|A| \leqq |B|$.

As in the proof of Theorem 4.2, the symmetric difference $M \oplus M'$ of any two matchings M, M' in G consists of a set of disjoint simple paths (possibly closed) in G. Let us call such a path an *M-path* if it contains one more edge of M than of M'; an M'-path is defined similarly. Clearly, all other paths in $M \oplus M'$ contain equal numbers of edges from M and M'. Now for any pair $\langle M, M' \rangle \in A$, the number of M-paths in $M \oplus M'$ must exceed the number of M'-paths by precisely two. We may therefore partition A into disjoint classes $\{A_r : 0 < r \leqq k\}$, where

$$A_r = \{\langle M, M' \rangle \in A : M \oplus M' \text{ contains } r+1 \ M\text{-paths and } r-1 \ M'\text{-paths}\}.$$

Similarly, the sets $\{B_r : 0 \leqq r \leqq k\}$ with

$$B_r = \{\langle M, M' \rangle \in B : M \oplus M' \text{ contains } r \ M\text{-paths and } r \ M'\text{-paths}\}$$

partition B. The lemma will follow from the fact that $|A_r| \leqq |B_r|$ for each $r > 0$.

Let us call a pair $\langle L, L' \rangle \in B_r$ *reachable* from $\langle M, M' \rangle \in A_r$ if and only if $L \oplus L' = M \oplus M'$ and L is obtained from M by taking some M-path of $M \oplus M'$ and flipping the parity of all its edges with respect to M and M'. (This is analogous to *unwinding* the path in the proof of Theorem 4.2.) Clearly, the number of elements of B_r reachable from a given $\langle M, M' \rangle \in A_r$ is just the number of M-paths in $M \oplus M'$, namely $r+1$. Conversely, any given element of B_r is reachable from precisely r elements of A_r. Hence if $|A_r| > 0$ we have

$$\frac{|B_r|}{|A_r|} = \frac{r+1}{r} > 1,$$

completing the proof of the lemma. □

Remark. In [14] the tight inequality

$$|M_k(G)|^2 \geqq \frac{(k+1)(m-k+1)}{k(m-k)} |M_{k+1}(G)|\,|M_{k-1}(G)|$$

is proved, where $m = \lceil n/2 \rceil$ and n is the number of vertices in G. The bound in our proof can also be improved, but we will not labour this point here as simple log-concavity is quite adequate for our purposes. □

Note that Theorem 5.1 immediately implies the following:

COROLLARY 5.2. *For a $2n$-vertex graph $G = (V, E)$ with $|M_n(G)| > 0$, the ratio $|M_{k-1}(G)|/|M_k(G)|$ increases monotonically with k in the range $0 < k \leqq n$; the maximum value of the ratio is $|M_{n-1}(G)|/|M_n(G)|$ and the minimum value is $|E|^{-1}$.* □

Armed with log-concavity, let us sketch how an algorithm that generates matchings from the weighted distribution of § 4 may be used to estimate the number of perfect matchings in a given *unweighted* $2n$-vertex graph $G = (V, E)$. Write m_k in place of $|M_k(G)|$, and assume that $m_n > 0$. The idea is to estimate the ratios m_{k+1}/m_k in turn in a sequence of n stages for $k = 0, \cdots, n-1$. Since $m_0 = 1$, an approximation to m_n is then obtained as the product of the estimated ratios.

In stage k we could in principle estimate m_{k+1}/m_k using an algorithm that generates matchings in G *uniformly*: just observe the relative numbers of $(k+1)$- and k-matchings in an independent sample produced by the generator. However, a very large sample may be necessary since these matchings might constitute only a tiny fraction of all matchings in G. This difficulty can be overcome by adding to every edge of G a weight c_k that is chosen so as to make the aggregated weight $m_k c_k^k$ of k-matchings maximal, i.e., at least as large as that of matchings of any other size. That such a weight exists is a direct consequence of the log-concavity of the m_i. To see this, take $c_k = m_{k-1}/m_k$ and let $G(c_k)$ denote the graph G augmented with weight c_k on every edge. (We will not have available the exact value of m_{k-1}/m_k, but it will suffice to substitute the *estimate* of this quantity obtained in the previous stage.) Then if p_i is the probability of being at an i-matching in the stationary distribution of the Markov chain $\mathcal{MC}_{\mathrm{md}}(G(c_k))$, we have for $i \geqq k$

$$(14) \qquad \frac{p_k}{p_i} = \frac{m_k c_k^k}{m_i c_k^i} = c_k^{k-i} \prod_{j=k}^{i-1} \frac{m_j}{m_{j+1}} \geqq \left(\frac{m_{k-1}}{m_k}\right)^{k-i} \left(\frac{m_{k-1}}{m_k}\right)^{i-k} = 1$$

where the inequality comes from Corollary 5.2. An identical bound holds for $i < k$. Since $\sum p_i = 1$, we conclude that $p_k \geqq (n+1)^{-1}$. Moreover, the probability of being at a $(k+1)$-matching satisfies

$$(15) \qquad p_{k+1} = \left(\frac{m_{k+1} c_k}{m_k}\right) p_k = \left(\frac{m_{k+1}}{m_k}\right)\left(\frac{m_{k-1}}{m_k}\right) p_k \geqq \frac{1}{|E|}\left(\frac{m_n}{m_{n-1}}\right) p_k.$$

Hence a lower bound of the form $1/\mathrm{poly}(n)$ holds for p_{k+1} also, provided the ratio m_{n-1}/m_n is polynomially bounded. These observations allow the ratio m_{k+1}/m_k to be estimated efficiently by sampling from the *weighted* distribution.

For the sampling itself we appeal to the Markov chain technique of § 4. (The algorithm is robust enough to cope with a small bias.) In view of Corollary 4.3, the generator will be efficient provided the various edge weights used in the algorithm are polynomially bounded. But each weight will be close to m_{k-1}/m_k for some k that by Corollary 5.2 lies in the range $[|E|^{-1}, m_{n-1}/m_n]$. This ensures rapid convergence of the Markov chain at all stages, provided again that m_{n-1}/m_n is polynomially bounded.

Notice how a polynomial bound on the ratio m_{n-1}/m_n plays a central role in the efficient operation of this algorithm. For an arbitrary function q of the natural number n, let us call a $2n$-vertex graph G *q-amenable* if either

(i) $|M_n(G)| = 0$, or

(ii) $|M_n(G)| > 0$ and $|M_{n-1}(G)|/|M_n(G)| \leqq q(n)$.

From the above discussion, we might expect to get an fpras for counting perfect matchings in q-amenable graphs for any fixed polynomial q.

The new approximation scheme for counting perfect matchings in q-amenable graphs G is spelled out in detail in Fig. 2. In line (4), $\mathcal{G}$ denotes the almost W-generator for matchings described in § 4, i.e., the call $\mathcal{G}(G(c), \cdot)$ invokes a simulation of the Markov chain $\mathcal{MC}_{\mathrm{md}}(G(c))$. The values of the sample size T and bias ξ will depend on n and the accuracy $0 < \varepsilon \leqq 1$ specified for the final estimate, as described below. The test in line (1) for the existence of a perfect matching may be implemented using any standard polynomial time algorithm.

THEOREM 5.3. *For an arbitrary polynomial q, the algorithm of Fig. 2 is an fpras for $|M_n(G)|$ in all q-amenable $2n$-vertex graphs G.*

Proof. In view of line (1), we need only consider graphs for which $m_n > 0$. Line (2) and the iterations of the **for**-loop correspond to the n stages of the computation mentioned above. Let c_{k+1} be the value of the weight parameter c at the end of stage k. We claim that, by making T a polynomial function of n and ε^{-1} and setting $\xi = \varepsilon/\alpha n$ for a suitable constant $\alpha > 1$, the following may be guaranteed:

$$(16) \qquad \forall k \quad \Pr\left(c_{k+1} \text{ approximates } m_k/m_{k+1} \text{ within ratio } 1 + \frac{\varepsilon}{2n}\right) \geqq \left(1 - \frac{1}{4n^2}\right)^k.$$

```
(1) if $|M_n(G)|=0$ then halt with output 0
    else begin
(2)   $c := |E|^{-1}$; $\Pi := |E|$;
      for $k := 1$ to $n-1$ do begin
(3)     if $c > 2q(n)$ or $c < (2|E|)^{-1}$ then halt with output 0
        else begin
(4)       make $T$ calls of the form $\mathcal{G}(G(c), \xi)$
            and let $Y$ be the set of outputs;
(5)       $\tilde{p}_k := T^{-1}|Y \cap M_k(G)|$; $\tilde{p}_{k+1} := T^{-1}|Y \cap M_{k+1}(G)|$;
(6)       if $\tilde{p}_k = 0$ or $\tilde{p}_{k+1} = 0$ then halt with output 0
(7)       else begin $c := c\tilde{p}_k/\tilde{p}_{k+1}$; $\Pi := \Pi/c$ end
        end
      end;
(8)   halt with output $\Pi$
    end
```

FIG. 2. *Approximation scheme for counting perfect matchings.*

This will imply that the product Π output in line (8) approximates m_n within ratio $(1+\varepsilon/2n)^n \leqq 1+\varepsilon$ with probability $(1-1/4n^2)^{n^2} \geqq 3/4$, as required. Moreover, the runtime of the procedure is bounded by a polynomial in n and ε^{-1}. (Note in particular that, by Corollary 4.3, the bounds on edge weights in line (3) ensure that each call to $\mathcal{G}$ is bounded in this way.) Hence the procedure is indeed an fpras.

The proof of (16) is a straightforward induction on k, the technical details of which are left to the reader. The important points to note are the following, assuming that c_k is a good estimate of m_{k-1}/m_k:

(i) In line (3), c will not violate the prescribed bounds because m_{k-1}/m_k lies in the range $[|E|^{-1}, q(n)]$.

(ii) From (14) and (15), the probabilities p_k, p_{k+1} of being at a k- and $(k+1)$-matching in the stationary distribution of the Markov chain $\mathcal{MC}_{\mathrm{md}}(G(c_k))$ used in stage $k+1$ are bounded below by a function of the form $1/\mathrm{poly}(n)$. Hence the modest sample size T in line (4) is enough to make the estimates $\tilde{p}_k$, $\tilde{p}_{k+1}$ of these quantities in line (5) good with high probability. (Note that the pathological cases of line (6) are therefore very unlikely to occur.)

The assignment to c in line (7) therefore makes c_{k+1} a good estimate of m_k/m_{k+1} with high probability. □

The algorithm of Theorem 5.3 is preferable to those described in § 3 in several respects. For a given input graph G, it makes use of a single Markov chain structure, the only manipulations being simple scaling of transition probabilities. It avoids any discussion of ad hoc processes with state space $M_k(G) \cup M_{k-1}(G)$, whose transition structure is not uniform over states. Moreover, the condition that the ratio $|M_{n-1}(G)|/|M_n(G)|$ should be polynomially bounded (if $|M_n(G)|>0$) is seen to arise directly from the log-concavity of the matching sequence.

Indeed, this condition seems to be a true characterisation of those graphs that can be handled by the algorithm, or equivalently of those matrices whose permanent we can efficiently approximate by this method. Since the condition is rather unfamiliar, it deserves further investigation. One worthwhile activity is to come up with simpler deterministic criteria that guarantee the condition holds. We have already seen one such criterion in § 3, namely that the graph is dense. Another criterion, due to Dagum et al. [9], is that the graph is bipartite and contains an αn-regular subgraph for some real $\alpha > 0$. However, as we will see in the next section, it turns out that the condition is a rather weak one and is, in fact, satisfied by almost all (bipartite) graphs. In other words, there exists a fixed polynomial q such that almost every graph is q-amenable. Thus of more practical interest is the problem of testing efficiently for any given graph whether the condition holds. Such a test would enable us not only to approximate the permanent in almost all cases, but also to reliably identify difficult instances.

We now present an efficient randomised algorithm that tests the condition in the following strong probabilistic sense. Let q be a polynomial. When given as input a $2n$-vertex graph G containing a perfect matching and a positive real $\delta > 0$, the algorithm

(i) Accepts with probability at least $1-\delta$ if $|M_{n-1}(G)|/|M_n(G)| \leqq q(n)$;

(ii) Rejects with probability at least $1-\delta$ if $|M_{n-1}(G)|/|M_n(G)| > 6q(n)$.

For intermediate values of the ratio, we do not care whether the algorithm accepts or rejects. (The value 6 here is used for illustrative purposes only and may be replaced by any fixed constant greater than 1.) Furthermore, the runtime of the algorithm will be bounded by a polynomial in n and $\lg \delta^{-1}$.

Before presenting the algorithm we make precise its implications for counting perfect matchings. Consider the following combined procedure, whose input is an *arbitrary* $2n$-vertex graph G:

(1) Using a standard polynomial time algorithm, test whether G contains a perfect matching. If not, output 0 and halt.

(2) Apply the above randomised test for the condition $|M_{n-1}(G)|/|M_n(G)| \leqq q(n)$, having set an error probability $\delta = 2^{-n}$. If the algorithm rejects, output "Graph is not q-amenable" and halt.

(3) Using the approximation scheme of Fig. 2 with $q(n)$ replaced by $6q(n)$ (and the test of line (1) omitted), estimate $|M_n(G)|$ and output the result.

This procedure will run in polynomial time for any desired polynomial q. There are two ways in which it may produce a misleading result. With probability at most δ it may falsely claim that the input graph G is not q-amenable. Or, again with probability at most δ, it may output an unreliable approximation to $|M_n(G)|$ obtained under the false assumption that $|M_{n-1}(G)|/|M_n(G)| \leqq 6q(n)$. Since δ decreases exponentially with n, the procedure will, with very high probability, produce a result that is *not* misleading. This will either be a statement that G is not q-amenable, or a reliable approximation of $|M_n(G)|$.

We now show how to construct the testing algorithm advertised above. It again makes use of the weighted Markov chain generator $\mathscr{G}$ for matchings and is extremely simple to describe:

(1) Make T calls of the form $\mathscr{G}(G(2q(n)), 1/16)$, and let $\tilde{p}$ be the proportion of perfect matchings among the outputs. (T will depend on the input δ as specified below.)

(2) If $\tilde{p} \geqq 3/8$ accept, otherwise reject.

To see that this algorithm works, consider the stationary distribution of the Markov chain $\mathcal{MC}_{\mathrm{md}}(G(2q(n)))$, and let p denote the probability of being at a perfect matching. Writing as usual m_k in place of $|M_k(G)|$, Corollary 5.2 implies that

$$\frac{m_k}{m_n} = \prod_{i=k}^{n-1} \frac{m_i}{m_{i+1}} \leqq \left(\frac{m_{n-1}}{m_n}\right)^{n-k}$$

for $0 \leqq k \leqq n$. Hence in the case that $m_{n-1}/m_n \leqq q(n)$ we have

$$p = \frac{m_n(2q(n))^n}{\sum_{k=0}^n m_k(2q(n))^k} \geqq \frac{2^n}{\sum_{k=0}^n 2^k} > \frac{1}{2}. \tag{17}$$

On the other hand, if $m_{n-1}/m_n > 6q(n)$ we have

$$p \leqq \frac{m_n(2q(n))^n}{m_n(2q(n))^n + m_{n-1}(2q(n))^{n-1}} < \frac{m_n}{m_n + 3m_n} = \frac{1}{4}.$$

An elementary statistical argument now shows that, by taking $T = \alpha \lg \delta^{-1}$ for a suitable constant α, we can arrange for

$$\Pr(\tilde{p} \geqq \tfrac{3}{8}) \begin{cases} \geqq 1 - \delta & \text{if } m_{n-1}/m_n \leqq q(n), \\ \leqq \delta & \text{if } m_{n-1}/m_n > 6q(n). \end{cases}$$

The runtime of the algorithm is therefore bounded as required.

Remark. It is often desirable to be able to count matchings of any specified cardinality in a given graph. In the context of the monomer-dimer systems of the previous section, these correspond to configurations with a given number of dimers. Obviously, the procedure of Fig. 2 may be modified so as to yield an fpras for $|M_k(G)|$ in graphs G for which the ratio $|M_{k-1}(G)|/|M_k(G)|$ is polynomially bounded. Such graphs may again be identified efficiently using a minor variant of the above randomised test. Note also that, under the same condition on G, we can easily adapt the algorithm of Fig. 2 to produce an f.p. almost uniform generator for $M_k(G)$ for any desired k. □

We close this section with a slight digression. In recent years, a stochastic search heuristic for combinatorial optimisation known as *simulated annealing* [21] has received much attention. The basic idea is that a Markov chain explores a space of configurations (feasible solutions), each of which has an associated cost or "energy." In the stationary distribution of the chain, low cost solutions have large weight so the chain tends to favour them asymptotically. By progressively reducing a "temperature" parameter, the weights are scaled so as to accentuate the depths of the energy wells. (Thus the process is not in general time-homogeneous.) While such a process is known to converge asymptotically under fairly general conditions (see, e.g., [13], [24]), virtually nothing useful is known about its *rate* of convergence when applied to nontrivial problems.

Consider the problem of finding a maximum cardinality matching in a graph G, which is nontrivial in the sense that all known polynomial time algorithms for solving it are far from simple. For any $c \geqq 1$ we may take the Markov chain $\mathcal{MC}_{\mathrm{md}}(G(c))$ as the basis for a simulated annealing algorithm for this problem: maximum cardinality matchings will certainly have maximum weight, and "temperature" may be reduced by increasing the edge weight c.

In [27], Sasaki and Hajek study the performance of algorithms of this kind. In particular, they prove a positive result of the following form.

THEOREM 5.4. *Let $\varepsilon > 0$ be any constant, $G = (V, E)$ be an input graph, and k_0 the maximum cardinality of a matching in G. Then a simulated annealing algorithm of the above type, operated at a fixed temperature (which depends on G and ε), finds a matching in G of cardinality at least $(1-\varepsilon)k_0$ with high probability in polynomial time.*

(In the same paper they also prove a strong negative result that says that no simulated annealing algorithm in this or a fairly large related class can be relied on to find a maximum cardinality matching in polynomial time with high probability.)

Sasaki and Hajek's proof is lengthy and complex. In contrast, we offer the following argument which rests directly on our earlier results.

Proof. Define $c_\varepsilon = 2|E|^{(1-\varepsilon)/\varepsilon}$. We claim that, in the stationary distribution of the Markov chain $\mathcal{MC}_{\mathrm{md}}(G(c_\varepsilon))$, the probability of being at a matching of size $k = \lceil (1-\varepsilon)k_0 \rceil$ or more is greater than $\frac{1}{2}$. Note that the theorem then follows at once: by

Corollary 4.3 a polynomially bounded simulation of $\mathcal{MC}_{\mathrm{md}}(G(c_\varepsilon))$ suffices to ensure that we visit a matching of size k or more with probability at least (say) $\frac{1}{4}$. This can be boosted to $1-\delta$ by repeating the entire experiment $O(\lg \delta^{-1})$ times. (However, in common with [27], our time bound increases exponentially with ε^{-1}.)

To justify the claim, note from Corollary 5.2 that

$$m_{k-1}=m_{k_0}\prod_{j=k}^{k_0}\frac{m_{j-1}}{m_j}\geqq\left(\frac{m_{k-1}}{m_k}\right)^{k_0-k+1}. \tag{18}$$

(Here we are using the fact that $m_{k_0}\geqq 1$.) But since j-matchings in G are subsets of E of size j, there is also the crude upper bound $m_{k-1}\leqq|E|^{k-1}$. Hence from (18) we conclude that

$$\frac{m_{k-1}}{m_k}\leqq|E|^{(1-\varepsilon)/\varepsilon}=\frac{c_\varepsilon}{2}. \tag{19}$$

A further application of Corollary 5.2 now shows that $m_i/m_k\leqq(c_\varepsilon/2)^{k-i}$ for $0\leqq i\leqq k$, so the aggregated weight of matchings of size less than k is

$$\sum_{i=0}^{k-1} m_i c_\varepsilon^i\leqq\left(\sum_{i=0}^{k-1}2^{i-k}\right)m_k c_\varepsilon^k< m_k c_\varepsilon^k.$$

It is now immediate that the probability of being at a matching of size k or more is at least $\frac{1}{2}$, completing the proof. □

Remark. Inequality (19) provides a polynomial upper bound on the ratio m_{k-1}/m_k, so from our earlier observations we are able to count and generate matchings of any cardinality up to $(1-\varepsilon)k_0$ in *arbitrary* graphs which contain a k_0-matching. □

6. Random permanents. For any polynomial q, the algorithm presented in Fig. 2 of the previous section is an fpras for the number of perfect matchings in a q-amenable graph G. For a bipartite graph G with $2n$ vertices, and $q(n)=n^2$, we have observed a sufficient condition for q-amenability, namely that the minimum vertex degree of G should be at least $n/2$. We have also observed that this result is the best possible, in the sense that, for any $\delta>0$, there exists a family of graphs of minimum vertex degree at least $n/(2+\delta)$ for which $|M_{n-1}(G)|/|M_n(G)|=\exp\{\Omega(n)\}$.

The aim of this section is to demonstrate that these counterexamples are pathological, and that a randomly selected bipartite graph with given edge density—even when that density is small—will almost surely be q-amenable for some suitably chosen (fixed) polynomial q.

Let n be a positive integer, and p a real number in the interval $(0,1)$. We will work with the probability space of bipartite graphs $G_{n,p}$ constructed according to the following random graph model. The vertex set of $G_{n,p}$ is $U+V$ where $U=V=\{0,\cdots,n-1\}$, and each potential edge (i.e., element of $U\times V$) is included in the edge set of $G_{n,p}$ independently and with probability p. (In the sequel, $G_{n,p}$ will always denote a graph randomly selected according to this model.) We say that an event A in this model occurs *with overwhelming probability* if $1-\Pr(A)=O(n^{-k})$ for all integer k. (The O-expression here is a function of n only, and is independent of p.)

The main result of the section (Theorem 6.4) is that for most values of p, and for a suitably chosen (fixed) polynomial q, the graph $G_{n,p}$ is q-amenable with overwhelming probability. Thus, in probabilistic terms, the approximation scheme of Fig. 2 has very wide applicability. Recall also that the rare examples that the scheme cannot handle reliably may be identified using the randomised test of the previous section. We approach Theorem 6.4 via a sequence of three technical lemmas, the proofs of which are deferred.

LEMMA 6.1. *Let $\varepsilon > 0$ be a fixed constant, and let $p \leqq (1-\varepsilon)n^{-1} \ln n$. Then, with overwhelming probability, $G_{n,p}$ has no perfect matching.*

Call a graph (k, m)*-expanding* [7, p. 327] if every k-subset of U is adjacent to at least m vertices of V, and vice versa.

LEMMA 6.2. *Let $\varepsilon > 0$ be a fixed constant, $p \geqq (1+\varepsilon)n^{-1} \ln n$, and $\alpha = pn/\ln \ln n$. Then, with overwhelming probability, $G_{n,p}$ is $(k, m+1)$-expanding for all integers k and m that satisfy the inequalities $k \geqq \ln n/\ln pn$, $m \leqq \alpha k$, and $m \leqq n/2$.*

LEMMA 6.3. *Let $p \geqq n^{-1} \ln n$. Then, with overwhelming probability, the maximum vertex degree of $G_{n,p}$ does not exceed $pn \ln n$.*

THEOREM 6.4. *Let $\varepsilon > 0$ be a fixed constant, and let p lie outside the interval*

$$((1-\varepsilon)n^{-1} \ln n, (1+\varepsilon)n^{-1} \ln n).$$

Then, with overwhelming probability, the graph $G_{n,p}$ is q-amenable for $q(n) = n^{10}$.

Proof. Let A_1 denote the event $|M_n(G_{n,p})| = 0$, and A_2 the event

$$|M_n(G_{n,p})| > 0 \quad \text{and} \quad |M_{n-1}(G_{n,p})|/|M_n(G_{n,p})| \leqq n^{10}.$$

The event that the graph $G_{n,p}$ is q-amenable is the disjunction of the events A_1 and A_2. If $p \leqq (1-\varepsilon)n^{-1} \ln n$ then, by Lemma 6.1, event A_1 occurs with overwhelming probability. So from now on we assume that $p \geqq (1+\varepsilon)n^{-1} \ln n$.

Let B be the event that $G_{n,p}$ is $(k, m+1)$-expanding for all k, m in the ranges allowed in the statement of Lemma 6.2; let C be the event that the maximum degree of $G_{n,p}$ does not exceed $pn \ln n$. Suppose, as we will prove, that the event A_2 is a logical consequence of the events B, C, and $\bar{A}_1$ (the complement of A_1), that is to say, $A_2 \supseteq B \cap C \cap \bar{A}_1$. Then, by elementary set theory, $A_1 \cup A_2 \supseteq A_1 \cup (B \cap C \cap \bar{A}_1) \supseteq (B \cap C)$. Thus, $\Pr(A_1 \cup A_2) \geqq \Pr(B \cap C) \geqq 1 - \Pr(\bar{B}) - \Pr(\bar{C})$. The theorem follows from the estimates for $\Pr(\bar{B})$ and $\Pr(\bar{C})$ provided by Lemmas 6.2 and 6.3.

To complete the proof, we need to show that any graph $G = G_{n,p}$ that satisfies B, C, and $\bar{A}_1$ must also satisfy A_2. Our strategy is to demonstrate that every $(n-1)$-matching M in G can be extended to a perfect matching of G by augmentation along a *short* alternating path. (An *alternating path* is a path in G whose edges lie alternately inside and outside the matching M.) Since every $(n-1)$-matching is "close to" some perfect matching, the ratio of $(n-1)$-matchings to perfect matchings cannot be very large. (A similar technique was used in the proof of Theorem 3.2.)

So let M be any $(n-1)$-matching in G. For $s \in U+V$, $T \subset U+V$ and i a positive integer, denote by $\Gamma^i_{\text{alt}}(s, T)$ the set of vertices in T that can be reached from vertex s by an alternating path of length at most i. Set $L = (1+\varepsilon_1(n)) \ln n/\ln pn$, where ε_1 is a positive real function that tends to zero (and that will be defined implicitly later in the proof). We will prove that M can be extended to a perfect matching via a path of length at most $8L$.

Let $u \in U$, $v \in V$ be the vertices of G that are left uncovered by M. Consider the set $\Gamma^{2L-1}_{\text{alt}}(u, V)$. If $v \in \Gamma^{2L-1}_{\text{alt}}(u, V)$ then we are done, so assume the contrary. For any i in the range $1 \leqq i < L$, we have the inequality $|\Gamma^{2i}_{\text{alt}}(u, U)| > |\Gamma^{2i-1}_{\text{alt}}(u, V)|$. To see this, note that $|\Gamma^{2i-1}_{\text{alt}}(u, V)|$ vertices in U can be reached from vertices in $\Gamma^{2i-1}_{\text{alt}}(u, V)$ via a single edge in M, and that these vertices do not include u, which can be reached by the null alternating path. Moreover, $|\Gamma^{2i+1}_{\text{alt}}(u, V)| \geqq |\Gamma^{2i}_{\text{alt}}(u, U)|$ since G is assumed to contain a perfect matching. (This is the trivial direction of Hall's Theorem.) Putting the inequalities together we obtain $|\Gamma_{\text{alt}}^{2i+1}(u, V)| > |\Gamma_{\text{alt}}^{2i-1}(u, V)|$, and, by iteration, $|\Gamma_{\text{alt}}^{2L-1}(u, V)| \geqq L$.

We continue the process of computing lower bounds on $|\Gamma_{\text{alt}}^{i}(u, V)|$ for increasing i, but now using the improved expansion factor provided by Lemma 6.2. (Note that $k = |\Gamma_{\text{alt}}^{2L-1}(u, V)| \geqq L > \ln n/\ln pn$, the threshold stipulated in the lemma.) For $i \geqq L$ we have $|\Gamma_{\text{alt}}^{2i+1}(u, V)| \geqq \min\{\alpha|\Gamma_{\text{alt}}^{2i-1}(u, V)|, \lceil n/2 \rceil\}$, where $\alpha = pn/\ln\ln n$. Thus, $|\Gamma_{\text{alt}}^{4L-1}(u, V)| \geqq \min\{|\Gamma_{\text{alt}}^{2L-1}(u, V)|\alpha^L, \lceil n/2 \rceil\}$. Since

$$\alpha^L = \exp\left\{(1+\varepsilon_1(n))\frac{\ln n}{\ln pn}(\ln pn - \ln\ln\ln n)\right\} \geqq n$$

for suitably chosen $\varepsilon_1(n) \to 0$, we deduce that $|\Gamma_{\text{alt}}^{4L-1}(u, V)| \geqq \lceil n/2 \rceil$. A symmetrical argument gives $|\Gamma_{\text{alt}}^{4L-1}(v, U)| \geqq \lceil n/2 \rceil$. Since some pair of vertices in $\Gamma_{\text{alt}}^{4L-1}(u, V)$ and $\Gamma_{\text{alt}}^{4L-1}(v, U)$ must be connected by an edge of M, there must exist an augmenting path for M of length not greater than $8L-1$.

Finally, associate with each $(n-1)$-matching M of G a perfect matching $\bar{M}$ that is reachable from M via an augmenting path of length at most $8L-1$. For each perfect matching $P \in M_n(G)$, let $\mathcal{K}(P) = \{M \in M_{n-1}(G): \bar{M} = P\}$ be the set of $(n-1)$-matchings associated with P; clearly, $\{\mathcal{K}(P): P \in M_n(G)\}$ is a partition of the set $M_{n-1}(G)$. To complete the proof, it is sufficient to show that the cardinality of $\mathcal{K}(P)$ is bounded above by n^{10}, for sufficiently large n.

Let M be an element of $\mathcal{K}(P)$. By definition, M can be reached from P by unwinding an alternating path of length less than $8L$. We can view the construction of such an alternating path as a sequence of choices. First select one of the n vertices of U as a starting point. Then, at each of at most $4L$ points during the tracing of the path, namely, each time the path visits a vertex v in V, select one of at most $pn \ln n$ possible next moves: either terminate the path at v, or extend it along one of the $pn \ln n - 1$ free edges incident at v. (Recall that G has maximum degree $pn \ln n$, and note that moves from U to V are forced.) Thus the number of possible augmenting paths, and hence the cardinality of $\mathcal{K}(P)$, is bounded above by

$$n(pn \ln n)^{4L} = n\exp\{4L(\ln pn + \ln\ln n)\} \leqq n\exp\{8L \ln pn\} = n^{1+8(1+\varepsilon_1(n))}.$$

Since $\varepsilon_1(n) \to 0$ as $n \to \infty$, the cardinality of $\mathcal{K}(P)$ is bounded by n^{10} for sufficiently large n. □

Remark. Neither event A_1 nor A_2 need, in isolation, occur with overwhelming probability, only their disjunction. This can be demonstrated by setting $p = \beta n^{-1} \ln n$, where β is any constant greater than 1. □

The condition on p in Theorem 6.4 is an unfortunate blemish. Alan Frieze [34] has indicated that the condition can be dropped at the expense of a slight weakening of the conclusion: q-amenability would now hold with probability tending to 1 as n tends to infinity, rather than with overwhelming probability.

We close the section by providing proofs of the three technical lemmas, using standard techniques from the theory of random graphs.

Proof of Lemma 6.1. The probability that $G_{n,p}$ has a perfect matching is certainly less than the probability that no vertex in U is isolated (has degree zero) so it is enough to bound the latter probability. Our calculations will make free use of the inequalities $1 - t \geqq \exp(-t - t^2)$ and $1 - t \leqq \exp(-t)$, the first of which is valid for $0 \leqq t < 0.69$, and the second valid unconditionally [7, p. 5]. First consider the probability that a *particular* vertex $u \in U$ is isolated:

$$\begin{aligned}\Pr(u \text{ is isolated}) &= (1-p)^n \\ &\geqq \exp\{n(-p-p^2)\} \\ &\geqq \exp\{-(1-\varepsilon)\ln n - n^{-1}\ln^2 n\} \\ &= n^{-(1-\varepsilon)} + o(n^{-1}).\end{aligned}$$

Thus, the probability that *no* vertex in U is isolated is $\{1-(1-p)^n\}^n \leqq \exp\{-n(1-p)^n\} \leqq \exp\{-n^{\varepsilon}+o(1)\}$. □

Proof of Lemma 6.2. Denote by A_{km} the event that $G_{n,p}$ is $(k, m+1)$-expanding. It is clearly enough to show that, for *arbitrary* k, m satisfying the given inequalities, the event A_{km} occurs with overwhelming probability. Furthermore, since $\Pr(A_{km})$ increases monotonically with k, it is sufficient to show that the event A_{km} occurs with overwhelming probability for k and m satisfying $\ln n/\ln pn \leqq k \leqq \lceil n/2\alpha \rceil$, and $m \leqq \alpha k$.

Let U' be an arbitrary k-subset of U. The set of vertices in V which are adjacent to U' in $G_{n,p}$ may be modelled as a sequence of n Bernoulli trials with success probability $q = 1-(1-p)^k$. Thus the probability that U' is adjacent to at most m vertices in V is

$$\sum_{t=0}^{m} \binom{n}{t} q^t (1-q)^{n-t},$$

and the probability $\Pr(\bar{A}_{km})$ that $G_{n,p}$ *fails* to be $(k, m+1)$-expanding is bounded above by

$$2\binom{n}{k} \sum_{t=0}^{m} \binom{n}{t} q^t (1-q)^{n-t}. \tag{20}$$

By Chernoff's bound [12, p. 18] and using the inequality

$$\binom{n}{k} \leqq \frac{1}{\sqrt{2\pi k}} \left(\frac{en}{k}\right)^k \leqq \frac{1}{2}\left(\frac{en}{k}\right)^k$$

the failure probability (20) may be bounded as follows:

$$\Pr(\bar{A}_{km}) \leqq \exp\left\{(n-m)\ln\frac{(1-q)n}{n-m} + m\ln\frac{qn}{m} + k\ln\frac{en}{k}\right\}. \tag{21}$$

Since $q = 1-(1-p)^k$, we have the relations $1-q \leqq \exp(-pk)$ and $q \leqq pk$; employing these in inequality (21), we obtain

$$\Pr(\bar{A}_{km}) \leqq \exp\left\{-pk(n-m) + (n-m)\ln\frac{n}{n-m} + m\ln\frac{pkn}{m} + k\ln\frac{en}{k}\right\}.$$

Further simplification, using the fact that $\ln(n/(n-m)) \leqq m/(n-m)$, yields

$$\Pr(\bar{A}_{km}) \leqq f(p, k, m) = \exp\left\{-pk(n-m) + m\left(1+\ln\frac{pkn}{m}\right) + k\ln\frac{en}{k}\right\}. \tag{22}$$

Our goal is to bound $\Pr(\bar{A}_{km})$ by maximising $f(p, k, m)$ (viewed as a function of three real variables) over the ranges $p \geqq (1+\varepsilon)n^{-1}\ln n$, $\ln n/\ln pn \leqq k \leqq \lceil n/2\alpha \rceil$, and $m \leqq \alpha k$.

By differentiating (22) with respect to m we discover that, with p, k fixed and n sufficiently large, $f(p, k, m)$ is an increasing function of m. (Indeed it is sufficient for n to be greater than 15, guaranteeing $m \leqq pkn/\ln\ln n < pkn$.) Thus, in attempting to bound $f(p, k, m)$, it is enough to consider those triples (p, k, m) for which the inequality that governs m, namely $m \leqq \alpha k$, is actually an equality. Substituting $k = m/\alpha = m \ln\ln n/pn$ in (22), our task is now to bound

$$f_1(p, m) = \exp\left\{-m \ln\ln n\left(1 - \frac{m}{n}\right) + m(\ln\ln\ln n + 1) + \frac{m \ln\ln n}{pn} \ln \frac{epn^2}{m \ln\ln n}\right\}$$

over feasible p, m. Now, for fixed m, the function f_1 decreases with p. (For this we need the inequality $m \leqq pkn/\ln\ln n$.) So making the substitution $p = (1+\varepsilon)n^{-1} \ln n$ we are further reduced to bounding the function

$$f_2(m) = \exp\left\{-m \ln\ln n\left(1 - \frac{m}{n}\right) + m(\ln\ln\ln n + 1) + \frac{m \ln\ln n}{(1+\varepsilon)\ln n} \ln \frac{e(1+\varepsilon)n \ln n}{m \ln\ln n}\right\}$$

over feasible m. The argument to the exponential is a convex function of m, so we can bound $f_2(m)$ by considering its values at the extremes of m's range. A lower bound for m is given by the chain of inequalities

$$m = \alpha k = \frac{pkn}{\ln\ln n} \geqq \frac{pn \ln n}{\ln\ln n \ln pn} \geqq \left(\frac{\ln n}{\ln\ln n}\right)^2 = m_{\min},$$

and an upper bound by

$$m = \alpha k \leqq \alpha\left\lceil \frac{n}{2\alpha} \right\rceil \leqq \frac{3n}{4} = m_{\max}$$

where we have used the known bounds on k and p, and assumed n sufficiently large. Substituting these extreme values in the expression for $f_2(m)$, we obtain

$$f_2(m_{\min}) = \exp\left\{-\left(\frac{\varepsilon}{1+\varepsilon} - o(1)\right)\frac{\ln^2 n}{\ln\ln n}\right\},$$

$$f_2(m_{\max}) = \exp\left\{-\left(\frac{3}{16} - o(1)\right) n \ln\ln n\right\}.$$

The former bound is the weaker and hence is the one that determines the overall bound on $f(p, k, m)$. □

Proof of Lemma 6.3. It is clearly enough to show that, with overwhelming probability, the degree of an arbitrary vertex $u \in U$ is bounded by $pn \ln n$. The set of vertices adjacent to u may be modelled as a sequence of n Bernoulli trials with success probability p. The probability that $\delta(u)$, the degree of u, exceeds m can be estimated from Chernoff's bound, using manipulations similar to those in the proof of Lemma 6.2:

$$\begin{aligned}\Pr(\delta(u) > m) &\leqq \exp\left\{(n-m)\ln\frac{(1-p)n}{n-m} + m \ln\frac{pn}{m}\right\}\\ &\leqq \exp\left\{-p(n-m) + m + m \ln\frac{pn}{m}\right\}\\ &\leqq \exp\left\{-m\left(\ln\frac{m}{pn} - 1\right)\right\}.\end{aligned}$$

(Clearly we may assume $m < n$.) Now, substituting $pn \ln n$ for m and noting $p \geqq n^{-1} \ln n$, we obtain

$$\Pr(\delta(u) > pn \ln n) \leqq \exp\{-(1-o(1)) \ln^2 n \ln\ln n\},$$

which decays faster than the reciprocal of any polynomial in n. □

7. Miscellaneous remarks and open problems. (i) The existence of an fpras for the unrestricted permanent remains an intriguing open question. Whereas the requirement that the ratio $|M_{n-1}(G)|/|M_n(G)|$ be polynomially bounded arises very naturally from our methods, there seems to be no a priori reason to suspect that graphs that violate this condition are particularly hard to handle. Perhaps an fpras of a different kind can be found for graphs in which the ratio is large. Alternatively, it is conceivable that counting perfect matchings approximately in general graphs is hard, in the sense that the existence of an fpras for this problem would imply that NP = RP. (Hardness results of this kind for other structures appear in [16], [28].)

(ii) It would be interesting to know whether the ratio $|M_{n-1}(G)|/|M_n(G)|$ is polynomially bounded for other natural classes of graphs, immediately yielding an fpras for the number of perfect matchings. For example, this question is pertinent for families of regular lattices encountered in statistical physics. Much effort has been expended on counting perfect matchings in such graphs, and an elegant exact solution obtained for planar lattices (or indeed arbitrary planar graphs [19]). The three-dimensional case, however, remains open even in approximate form.

(iii) From a practical point of view, it would be interesting to know whether the conductance bounds we have derived can be significantly improved. We make no claim of optimality here, preferring to concentrate on giving a clear exposition of the rapid mixing property. The practical utility of our algorithms, however, is likely to depend on rather tighter bounds being available.

Similar considerations apply to our methods for estimating the expectation of a 0-1 random variable under the stationary distribution of a Markov chain. We have chosen to view the chain as a generator of independent samples, partly to simplify the statistical concepts involved and partly because the random generation problems are of interest in their own right. In contrast, Aldous [2] considers estimates derived by observing a Markov chain continuously and formulates the definition of rapid mixing directly in terms of the variance of such an estimate. This approach may lead to increased efficiency.

(iv) A wider issue is the extent to which the techniques of this paper can be applied to the analysis of natural Markov chains whose states are combinatorial structures other than matchings. Since these chains are usually time-reversible, the conductance characterisation of rapid mixing presented in § 2 can in principle be applied. We conjecture that this is possible in practice for other interesting chains, and that the path counting technique developed in this paper is a promising general approach for obtaining positive results. It is to be hoped that this will yield efficient random generation and approximate counting procedures for further structures. Moreover, it may lead to rigorous performance guarantees for Monte Carlo experiments in statistical physics, and a demystification of currently fashionable stochastic optimisation techniques such as simulated annealing.

Acknowledgment. The authors thank Leslie Valiant for bringing reference [25] to their attention.

REFERENCES

[1] D. Aldous, *Random walks on finite groups and rapidly mixing Markov chains*, Séminaire de Probabilités XVII, 1981/82, Springer Lecture Notes in Mathematics 986, Springer-Verlag, Berlin, New York, 1983, pp. 243-297.

[2] ———, *On the Markov chain simulation method for uniform combinatorial distributions and simulated annealing*, Probab. Engrg. Inform. Sci., 1 (1987), pp. 33-46.

[3] D. ALDOUS AND P. DIACONIS, *Shuffling cards and stopping times*, Amer. Math. Monthly, 93 (1986), pp. 333-348.

[4] N. ALON, *Eigenvalues and expanders*, Combinatorica, 6 (1986), pp. 83-96.

[5] N. ALON AND V. D. MILMAN, λ_1, *isoperimetric inequalities for graphs and superconcentrators*, J. Combin. Theory Ser. B, 38 (1985), pp. 73-88.

[6] K. BINDER, *Monte Carlo investigations of phase transitions and critical phenomena*, in Phase Transitions and Critical Phenomena, Volume 5b, C. Domb and M. S. Green, eds., Academic Press, London, 1976, pp. 1-105.

[7] B. BOLLOBÁS, *Random Graphs*, Academic Press, London, 1985.

[8] A. Z. BRODER, *How hard is it to marry at random? (On the approximation of the permanent)*, in Proc. 18th Annual ACM Symposium on Theory of Computing, 1986, pp. 50-58; Erratum in Proc. 20th Annual ACM Symposium on Theory of Computing, 1988, p. 551, Association for Computing Machinery, New York.

[9] P. DAGUM, M. LUBY, M. MIHAIL, AND U. V. VAZIRANI, *Polytopes, permanents and graphs with large factors*, in Proc. 29th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computing Society, Washington, DC, 1988.

[10] P. DIACONIS AND M. SHAHSHAHANI, *Generating a random permutation with random transpositions*, Z. Wahrsch. Verw. Gebiete, 57 (1981), pp. 159-179.

[11] J. EDMONDS, *Paths, trees and flowers*, Canad. J. Math., 17 (1965), pp. 449-467.

[12] P. ERDOS AND J. SPENCER, *Probabilistic Methods in Combinatorics.*, Academic Press, New York, 1974.

[13] B. HAJEK, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988).

[14] O. J. HEILMANN AND E. H. LIEB, *Theory of monomer-dimer systems*, Comm. Math. Phys., 25 (1972), pp. 190-232.

[15] M. R. JERRUM, *Two-dimensional monomer-dimer systems are computationally intractable*, J. Statist. Phys., 48 (1987), pp. 121-134.

[16] M. R. JERRUM, L. G. VALIANT, AND V. V. VAZIRANI, *Random generation of combinatorial structures from a uniform distribution*, Theoret. Comput. Sci., 43 (1986), pp. 169-188.

[17] N. KARMARKAR, R. M. KARP, R. LIPTON, L. LOVÁSZ, AND M. LUBY, *A Monte Carlo algorithm to approximate the permanent*, preprint, 1988.

[18] R. M. KARP AND M. LUBY, *Monte-Carlo algorithms for enumeration and reliability problems*, in Proc. 24th Annual IEEE Symposium on Foundations of Computer Science, IEEE Computing Society, Washington, DC, 1983, pp. 56-64.

[19] P. W. KASTELYN, *Graph theory and crystal physics*, in Graph Theory and Theoretical Physics, F. Harary ed., Academic Press, London, 1967, pp. 43-110.

[20] J. KEILSON, *Markov Chain Models—Rarity and Exponentiality*, Springer-Verlag, Berlin, New York, 1979.

[21] S. KIRKPATRICK, C. D. GELLATT, AND M. P. VECCHI, *Optimisation by simulated annealing*, Science, 220 (1983), pp. 671-680.

[22] G. F. LAWLER AND A. D. SOKAL, *Bounds on the L^2 spectrum for Markov chains and Markov processes: A generalization of Cheeger's inequality*, Trans. Amer. Math. Soc., 309 (1988), pp. 557-580.

[23] L. LOVÁSZ AND M. D. PLUMMER, *Matching Theory*, North-Holland, Amsterdam, 1986.

[24] M. LUNDY AND A. I. MEES, *Convergence of an annealing algorithm*, Math. Programming, 34 (1986), pp. 111-124.

[25] M. MIHAIL, *On coupling and the approximation of the permanent*, Inform. Process. Lett., 30 (1989), pp. 91-96.

[26] H. MINC, *Permanents*, Addison-Wesley, Reading, MA, 1978.

[27] G. H. SASAKI AND B. HAJEK, *The time complexity of maximum matching by simulated annealing*, J. Assoc. Comput. Mach., 35 (1988), pp. 387-403.

[28] A. J. SINCLAIR, *Randomised algorithms for counting and generating combinatorial structures*, Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland, June 1988.

[29] A. J. SINCLAIR AND M. R. JERRUM, *Approximate counting, uniform generation and rapidly mixing Markov chains*, Inform. Comput., 82 (1989), pp. 93-133.

[30] L. STOCKMEYER, *The complexity of approximate counting*, in Proc. 15th Annual ACM Symposium on Theory of Computing, Association for Computing Machinery, New York, 1983, pp. 118-126.

[31] J. D. ULLMAN, *Computational Aspects of* VLSI, Computer Science Press, Rockville, MD, 1984.

[32] L. G. VALIANT, *The complexity of computing the permanent*, Theoret. Comput. Sci., 8 (1979), pp. 189-201.

[33] ———, *The complexity of enumeration and reliability problems*, SIAM J. Comput., 8 (1979), pp. 410-421.

[34] A. FRIEZE, *A note on computing random permanents*, unpublished manuscript.