Rapidly Mixing Markov Chains: A Comparison of Techniques

(A Survey)

Venkatesan Guruswami

MIT Laboratory for Computer Science
545 Technology Square
Cambridge, MA 02139.
Email: venkat@theory.lcs.mit.edu.

Abstract

For many fundamental sampling problems, the best, and often the only known, approach to solving them is to take a long enough random walk on a certain Markov chain and then return the current state of the chain. Techniques to prove how long "long enough" is, i.e., the number of steps in the chain one needs to take in order to be sufficiently close to the stationary distribution of the chain, are of great importance in obtaining estimates of running times of such sampling algorithms.

In this report, we survey existing techniques to bound the mixing time of Markov chains. The mixing time of a Markov chain is exactly captured by the "spectral gap" of its underlying transition matrix. The spectral gap is closely related to a geometric parameter called "conductance" which is a measure of the "edge-expansion" of the Markov chain. Conductance also captures the mixing time up to square factors. Lower bounds on conductance, which give upper bounds on the mixing time, are typically obtained by a technique called "canonical paths" where the idea is to find a set of paths, one between every unequal source-destination pair, such that no edge is very heavily congested.

Unlike conductance, the canonical paths approach cannot always show rapid mixing of a rapidly mixing chain. It is known that this "drawback" disappears if we allow the flow between a pair of states to be spread along multiple paths. We prove that for a large class of Markov chains, including all the ones that we use in the sampling applications we will be interested in, canonical paths does capture rapid mixing, i.e., we show that small mixing time implies the existence of some collection of paths with low edge congestion. Allowing multiple paths to route the flow still does help a great deal in the design of such flows, and this is best illustrated by a recent result of Morris and Sinclair [34] on the rapid mixing of a natural Markov chain for sampling 0-1 knapsack solutions; this result seems to rely critically on fractional flows.

An entirely different approach to prove rapid mixing, which in fact historically preceded the conductance/canonical paths based approach, is "Coupling". Coupling is a very elegant technique and has been used to prove rapid mixing of several chains where designing good canonical paths seems to be a hideous task. "Path Coupling" is a related technique discovered by Bubley and Dyer [5] that often tremendously reduces the complexity of designing good Couplings. We present several applications of Path Coupling in proofs of rapid mixing, and these invariably lead to much better bounds on mixing time than known using conductance, and moreover Coupling based proofs usually turn out to be much simpler. These applications motivate the question of whether Coupling indeed can be made to work whenever the chain is rapidly mixing. This question was answered in the negative in very recent work by Kumar and Ramesh [27], who showed that no Coupling strategy can prove the rapid mixing of the famous Jerrum-Sinclair chain for sampling perfect and near-perfect matchings (the chain is known to be rapidly mixing via a canonical paths argument).

1 Introduction

Suppose Ω is a large finite set of combinatorial structures (for example the set of feasible solutions to a combinatorial optimization problem), and let π be a probability distribution on Ω. The general "sampling" problem is then to pick an element of Ω at random according to the distribution π. The Markov chain Monte Carlo method, which is the subject of our study here, provides an elegant technique to efficiently solve this general computational task in a wide variety of contexts.

Sampling problems are inherently interesting, and in addition turn out to have many computational applications, the most notable ones being:

- *Approximate counting:* Here we want to estimate the size of Ω to a very good accuracy. It is well known [23] that, provided a certain technical condition known as *self-reducibility* is met, almost uniform sampling (that is sampling from a distribution that is *statistically close* to the uniform distribution) is possible in polynomial time if and only if approximate counting is. This has been one of the main motivations, at least from the computer science point of view, behind the rapid progress that has been made in this area. In particular, for a host of counting problems including several very hard $\#P$-complete problems, the Markov chain Monte Carlo method is the only known approach to approximate the number of feasible solutions.

- *Statistical physics:* Here the space Ω represents possible configurations of statistical mechanical system, and π is a "natural" probability distribution on Ω, in which the probability of a configuration is related to its energy. The task is to sample configurations according to π, in order to examine properties of a "typical" physical configuration.

In this report, we focus only on the sampling problem and omit the connections to counting since these involve by now standard reductions. The Markov chain Monte Carlo method has been a great success story in solving sampling problems. It solves the sampling problem by the following approach. An underlying "Markov chain" $\mathfrak{M}$ on the state space Ω is specified through a stochastic *transition probability matrix* of dimension $|\Omega| \times |\Omega|$ whose $(x, y)^{\text{th}}$ entry specifies the probability $P(x, y)$ that the chain moves from state x to state y in a single step (we assume states of $\mathfrak{M}$ are labeled by elements of Ω). Starting at any state x_0, there is a natural *random walk* $X_0 = x_0, X_1, X_2, \ldots$ defined on $\mathfrak{M}$ such that $\mathbf{Pr}[X_{t+1}|X_0, \ldots, X_t] = \mathbf{Pr}[X_{t+1}|X_t]$ where the latter conditional probability is specified by the matrix P, i.e., $\mathbf{Pr}[X_{t+1} = y|X_t = x] = P(x, y)$. In other words we start at state X_0 and at each time step t, we make a move to a next state X_{t+1} by moving to a random state from the current state X_t according to the transition probabilities of the chain. Note the crucial "forgetting property" of Markov chains: the state at time $t + 1$ depends probabilistically on the state at time t, but not on the state at any other time.

To sample according to a distribution π, the Markov chain $\mathfrak{M}$ is defined in such a way that it is *ergodic*, i.e., has a (unique) stationary distribution η on Ω such that $\mathbf{Pr}[X_t = y|X_0 = x] \to \eta(y)$ as $t \to \infty$, for all pairs of states $x, y \in \Omega$, and moreover the transition probabilities are set up so that $\eta = \pi$. Now we may sample from Ω according to π as follows: starting from an arbitrary state in Ω, take a random walk on the Markov chain (which we will loosely refer to as "simulating the Markov chain" in the sequel) for some number, T, of steps, and then output the final state. The ergodicity of $\mathfrak{M}$ implies that, by taking T large enough, we can ensure that the output state is arbitrarily close to the desired distribution π.

One of the most appealing things about this method is its simplicity – in fact in most applications it is not hard to construct a Markov chain having the above properties. The crux of the method, which is also its sticking point, is to obtain good upper bounds on the *mixing time* of the chain, i.e., the number of simulation steps T necessary before the Markov chain is close to its stationary distribution. This is critical as this forms the crucial factor in the running time of any sampling algorithm that uses the chain. Since our aim is to sample from a set Ω which is very large, we would like T to be much smaller than the size of Ω, say at most a polynomial in the logarithm of $|\Omega|$. We shall refer to such chains as *rapidly mixing*. Over the years several deep and novel analytic tools have been developed and refined to bound mixing times of Markov chains. It is the goal of this report to survey the known techniques for proving rapid mixing, to present representative examples of their use, and to compare and contrast their scope, their relative strengths and limitations, and their applicability to various contexts.

Organization. We begin in the next section by reviewing the relevant definitions and properties of Markov chains, and by giving a precise characterization of when a Markov chain mixes rapidly in terms of its *spectral* properties. In Section 3 we discuss the notion of conductance and its relation to the spectral gap of the chain. Section 4 discusses the canonical paths approach and some of its generalizations that yield bounds on the conductance and the spectral gap, and also proves that for a large class of chains a small mixing time implies the existence of some collection of good canonical paths. We then present an illustrative application of this technique to the problem of sampling 0-1 knapsack solutions in Section 5. Section 6 discusses Coupling which is an entirely different approach to bounding the mixing time, gives an illustrative example of Coupling in action, and also discusses Path Coupling, which is a useful design tool in constructing Couplings. Several elegant applications of Path Coupling are presented in Section 7. In Section 8 we discuss the recent result of [27] which proves that Coupling is in fact weaker than conductance, in that there are chains with large conductance which cannot be shown to be rapidly mixing by *any* Coupling strategy. Finally, we conclude with a few remarks and open questions in Section 9.

Acknowledgments. This survey was written as part of the author's Area Examination at MIT, the goal of which was to survey the papers by Bubley and Dyer [5], Anil Kumar and Ramesh [27], and Morris and Sinclair [34]. This survey (specifically Sections 5, 6.4, 7.1 and 8) uses liberal portions of the contents of these papers. This work was also influenced greatly by the reading of the survey by Jerrum [19], and the paper by Sinclair [37], among several other papers. I would like to thank Kumar and Ramesh for sending me a copy of the most recent version of their paper [27].

2 Preliminaries on Markov Chains

A Markov chain on state space Ω is completely specified by the *transition matrix* P whose entry $P(x, y)$ represents the probability that the chain moves from state x to state y is a single transition; i.e., $P(x, y) = \mathbf{Pr}[X_{t+1} = y | X_t = x]$ for all $t \geq 0$. Thus in order to study and analyze the properties of the Markov chain, it suffices to investigate the properties of this matrix P.

2.1 Basic definitions

Starting from an initial distribution $\mu^{(0)}$, the distribution of the chain after t steps $\mu^{(t)}$ is clearly given by $\mu^{(t)} = \mu^{(0)} P^n$ (here we view the distributions as row vectors in $\mathbb{R}^\Omega$). Thus, when using a Markov chain to randomly sample from its state space, we must study the evolution of $\mu^{(t)}$ as t increases, and we would like $\mu^{(t)}$ to (quickly) approach a limiting *stationary* distribution, say π; it is not surprising that π must be fixed under steps of the chain.

Definition 2.1 *A row vector $\pi \in \mathbb{R}^\Omega$ is a* stationary distribution *for a Markov chain $\mathfrak{M}$ with transition matrix P if (a) $\pi(x) \geq 0$ for all $x \in \Omega$, (b) $\sum_{x \in \Omega} \pi(x) = 1$, and (c) $\pi = \pi P$.*

Definition 2.2 *A Markov chain $\mathfrak{M}$ is said to be* ergodic *if it has a stationary distribution.*

Clearly, we would like (and need) all Markov chains we use for sampling to be ergodic, so next we turn to conditions on the chain which will ensure ergodicity.

Definition 2.3 *A Markov chain $\mathfrak{M}$ (with transition matrix P) is said to be* irreducible *if for all $x, y \in \Omega$, there is an m such that $P^m(x, y) > 0$, i.e y is eventually reachable from x with non-zero probability.*

Irreducibility guarantees that the underlying chain is connected, so that starting at any state it is possible to reach all the other states. It is clearly desirable (and necessary) to impose this requirement when using a Markov chain to sample from a set Ω. We next impose another condition on the chains we will study, namely *aperiodicity*; this is merely a technical condition imposed to simplify analysis, and does not cause any loss of generality as we can turn any (periodic) chain into an aperiodic one by simply adding loop probabilities of $1/2$ at each state, and this clearly does not affect the stationary distribution.

Definition 2.4 *A chain $\mathfrak{M}$ over state space Ω is* aperiodic *iff for all $x \in \Omega$,*

$$\gcd\{m : P^m(x, x) > 0\} = 1.$$

A central theorem in the classical theory of stochastic process is the following:

Theorem 2.1 *Any finite, irreducible, aperiodic Markov chain is ergodic.*

Definition 2.5 *Suppose $\mathfrak{M}$ (defined over state space Ω) has a stationary distribution π. $\mathfrak{M}$ is said to be* reversible *(with respect to π) iff*

$$\pi(x)P(x, y) = \pi(y)P(y, x) \textit{ for all } x, y \in \Omega \ . \tag{1}$$

The conditions of (1) are known as *detailed balance equations*. The condition of reversibility does cause some loss of generality, but the ease of analysis gained by making this requirement more than compensates the sacrifice made. Moreover, reversible chains will be general enough for our applications, and for the rest of the section we focus attention solely on finite, irreducible, aperiodic and reversible Markov chains.

The detailed balance conditions also permit an easy proof that a certain distribution is indeed the stationary distribution of an ergodic Markov chain, as is formalized below.

Lemma 2.2 *For a Markov chain $\mathfrak{M}$ defined on state space Ω, if there exists a probability distribution π on Ω that satisfies the conditions (1), then π is a stationary distribution of $\mathfrak{M}$ and $\mathfrak{M}$ is reversible with respect to π.*

Proof: We easily verify that $\pi P = \pi$. Indeed,

$$(\pi P)(x) = \sum_y \pi(y)P(y, x) = \sum_y \pi(x)P(x, y) = \pi(x) \sum_y P(x, y) = \pi(x) \ . \qquad \square$$

Note that in the definition of ergodicity we did not require the stationary distribution to be unique, but the conditions of Lemma 2.2 together with irreducibility, are sufficient to guarantee that π is in fact the unique stationary distribution.

2.2 Spectral theory of reversible Markov chains

Since a stationary distribution of a Markov chain is simply a left eigenvector of its transition matrix P, it is natural that in order to study the rate of convergence of the chain to its stationary distribution, we should try to investigate the spectral properties of P. The reversibility constraint implies that one can view P as a self-adjoint operator on a suitable inner product space and this permits us to use the well-understood spectral theory of self-adjoint operators. This approach was first undertaken in [8] (also see [39] for a nice exposition).

The relevant inner product space is $L^2(\pi^{-1})$ which is the space of real-valued functions on Ω, with the following inner product:[1]

$$\langle \phi, \psi \rangle = \sum_{x \in \Omega} \frac{\phi(x)\psi(x)}{\pi(x)}. \tag{2}$$

It is easy to check that the detailed-balance conditions (1) imply that $\langle \phi P, \psi \rangle = \langle \phi, \psi P \rangle$, so that P is a self-adjoint operator on $L^2(\pi^{-1})$. Now, by standard linear algebra, it is well known that such a P has $N = |\Omega|$ real eigenvalues $1 = \lambda_0 > \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{N-1} \geq -1$; the chain defined by P is ergodic iff $\lambda_{N-1} > -1$. Also, the space $L^2(\pi^{-1})$ has an orthonormal basis comprising of eigenvectors $\pi = v_0, v_1, v_2, \ldots, v_{N-1}$ of P corresponding to the eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{N-1}$.

Now, our initial distribution on Ω can be written as $\mu^{(0)} = c_0\pi + c_1 v_1 + \cdots + c_{N-1} v_{N-1}$ where $c_i = \langle \mu^{(0)}, v_i \rangle$ (so in particular $c_0 = \sum_x \frac{\mu^{(0)}(x)\pi(x)}{\pi(x)} = 1$). The distribution after t steps is then given by

$$\mu^{(t)} = \mu^{(0)} P^t = \pi + c_1 \lambda_1^t v_1 + \cdots + c_{N-1} \lambda_{N-1}^t v_{N-1}. \tag{3}$$

From the above, it is clear that the chain is ergodic whenever $\lambda_{N-1} > -1$, as then all eigenvalues λ_i, $1 \leq i \leq N-1$, have absolute value less than 1, and as $t \to \infty$, terms corresponding to them will become insignificant, and $\mu^{(t)} \to \pi$. For an ergodic chain, Equation (3) also clearly demonstrates that the rate of convergence to π is governed by the second-largest eigenvalue in absolute value, $\lambda_{\max} = \max\{\lambda_1, |\lambda_{N-1}|\}$. We now make this statement precise. For $x \in \Omega$, denote by $P^t(x, \cdot)$ the distribution of the state of the Markov chain at time t, when the chain starts at time $t = 0$ in state x.

Definition 2.6 *The* variation distance *at time t with initial state x is defined as the statistical difference between distributions $P^t(x, \cdot)$ and $\pi(\cdot)$, i.e*

$$\Delta_x(t) = \frac{1}{2} \sum_{y \in \Omega} |P^t(x, y) - \pi(y)|.$$

We will measure the rate of convergence using the function τ_x, which quantifies the *mixing time*, and which is defined for $\varepsilon > 0$ by

$$\tau_x(\varepsilon) = \min\{t : \Delta_x(t') \leq \varepsilon \text{ for all } t' \geq t\}. \tag{4}$$

(It is easy to see that if $\Delta_x(t) \leq \varepsilon$ then $\Delta_x(t') \leq \varepsilon$ for all $t' \geq t$ as well.) With this notation, we will say a Markov chain is *rapidly mixing* if $\tau_x(\varepsilon)$ is $O(\mathrm{poly}(\log(N/\varepsilon)))$ (in applications the number of states N will be exponential in the problem size n, so this amounts to saying that we need to simulate the chain only for $\mathrm{poly}(n)$ steps in order to get a "good" sample from Ω). The following makes precise our intuition that a large value of the *spectral gap* $(1 - \lambda_{\max})$ exactly *captures* the rapid convergence to stationarity. A proof can be found in [8, 2].

Propostion 2.3 *The quantity $\tau_x(\varepsilon)$ satisfies*

(i) $\tau_x(\varepsilon) \leq (1 - \lambda_{\max})^{-1}\Big(\ln \pi(x)^{-1} + \ln \varepsilon^{-1}\Big)$.

(ii) $\max_{x \in \Omega} \tau_x(\varepsilon) \geq \frac{1}{2}\lambda_{\max}(1 - \lambda_{\max})^{-1} \ln(2\varepsilon)^{-1}$.

[1] It is easy to see that the stationary distribution satisfies $\pi(x) > 0$ for all $x \in \Omega$ whenever the chain is irreducible, so the inner product is well-defined.

In light of the above Proposition, if we want rapid convergence to the stationary distribution irrespective of the starting state (which is desirable for our applications in sampling where we would like to start at some arbitrary state), a large gap $(1-\lambda_{\max})$ is both a necessary and sufficient condition. Moreover, in practice the smallest eigenvalue λ_{N-1} is unimportant: a crude approach is to add a holding probability of $1/2$ to every state, i.e., replace P by $\frac{1}{2}(I+P)$, where I is the $N \times N$ identity matrix. This ensures that all eigenvalues are positive while decreasing the spectral gap $(1-\lambda_1)$ only by a factor of 2. The upshot is that in order to study mixing times of Markov chains, one needs to focus attention on the second-largest eigenvalue λ_1, and bound it away from 1.

2.3 Characterizations of second-largest eigenvalue

We now present the known characterizations of the second largest eigenvalue λ_1 of self-adjoint matrices, which will be useful in obtaining good bounds on the spectral gap $(1-\lambda_1)$.

Lemma 2.4 (Rayleigh-Ritz) *Let P be a self-adjoint operator on a finite-dimensional inner product space with inner product $\langle\cdot,\cdot\rangle$. Suppose the eigenvalues of P are $\lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_m$ and v_0 is an eigenvector of eigenvalue λ_0. Then*

$$\lambda_1 = \sup_{x \perp v_0} \frac{\langle x, xP\rangle}{\langle x, x\rangle} . \tag{5}$$

Proof: Let $v_0, v_1, \ldots, v_m$ be an orthonormal basis of eigenvectors corresponding to the eigenvalues $\lambda_0, \ldots, \lambda_m$ respectively. Since $x \perp v_0$, we can write x as $x = c_1 v_1 + \cdots + c_m v_m$, so that

$$\langle x, xP\rangle = \sum_{i=1}^{m} \lambda_i c_i^2 \leq \lambda_1 \sum_{i=1}^{m} c_i^2 = \lambda_1 \langle x, x\rangle.$$

When $x = v_1$, equality is achieved, and hence the result follows. □

We next present another characterization which at first glance seems a bit unwieldy, but it turns out to be quite useful in that very natural geometrical arguments about a Markov chain can yield upper bounds on λ_1 via this characterization [8].

Lemma 2.5 (Variational characterization) *Let P be a self-adjoint operator on a finite-dimensional inner product space $L^2(\pi^{-1})$, and for $x, y \in \Omega$, let $Q(x,y) = \pi(x)P(x,y) = Q(y,x)$. Then, the second-largest eigenvalue of P satisfies:*

$$1 - \lambda_1 = \inf_{\psi} \frac{\sum_{x,y\in\Omega} (\psi(x) - \psi(y))^2 Q(x,y)}{\sum_{x,y\in\Omega} (\psi(x) - \psi(y))^2 \pi(x)\pi(y)} . \tag{6}$$

3 Two broad approaches to proving Rapid Mixing

We saw in the last section that establishing rapid mixing for a Markov chain amounts to bounding the second largest eigenvalue λ_1 of the transition matrix P away from 1 by a $\mathrm{poly}(\log N)^{-1}$ amount. The spectrum of the chain is very hard to analyze directly, so we either need tools to analyze the spectral gap (using the characterizations presented in the previous section), or somehow analyze the chain directly without resorting to spectrum.

3.1 Coupling

One simple and elegant approach to bound mixing times without explicitly bounding the spectral gap is *Coupling*. A "coupling" argument is in fact the classical approach to bound mixing times of Markov chains. Coupling was first used by Aldous [1] to show rapid mixing, and has since found several applications in proving rapid mixing of a variety of chains. We will define Coupling formally and discuss some of its applications in detail in later Sections, but at a very high level the idea behind Coupling is the following. One sets up two stochastic processes $\mathcal{X} = (X_t)$ and $\mathcal{Y} = (Y_t)$ on the state space Ω both of which individually are faithful copies of the Markov chain $\mathfrak{M}$ (whose mixing time we wish to bound). However, their joint evolution is set up in a way that encourages (X_t) and (Y_t) to *coalesce* rapidly, so that $X_t = Y_t$ for all sufficiently large t. The relevance to rapid mixing is obvious from the Coupling Lemma [1, 19] which states that the probability that the coupling time exceeds some value t for a certain distribution π' for X_0 is an upper bound on the variation distance between the stationary distribution π of $\mathfrak{M}$ and the distribution of the chain at time t starting from distribution π'. Note that we did not explicitly deal with the spectrum of the chain, and this is one advantage of this approach. We will come back to a detailed discussion of Coupling in Sections 6 through 8.

3.2 Conductance

Let us now look at approaches aimed at establishing rapid mixing via directly bounding the spectral gap. These use geometric properties of the chain and the characterizations of λ_1 given by Equations (5) and (6) to prove a lower bound on the spectral gap $(1 - \lambda_1)$. The relevant geometric parameter is the *conductance* of the chain which is defined below.

Definition 3.1 *The* conductance *of* $\mathfrak{M}$ *is defined by*

$$\Phi = \Phi(\mathfrak{M}) \stackrel{\text{def}}{=} \min_{\substack{S \subset \Omega \\ 0 < \pi(S) \leq 1/2}} \frac{Q(S, \bar{S})}{\pi(S)}, \tag{7}$$

where $Q(x, y) = \pi(x)P(x, y) = \pi(y)P(y, x)$*,* $\pi(S)$ *is the probability density of* S *under the stationary distribution* π *of* $\mathfrak{M}$*, and* $Q(S, \bar{S})$ *is the sum of* $Q(x, y)$ *over all* $(x, y) \in S \times (\Omega - S)$*.*

The conductance may be viewed as a weighted version of *edge expansion* of the graph underlying the chain $\mathfrak{M}$. For a fixed S, the quotient in Equation (7) is just the conditional probability that the chain in equilibrium escapes from the subset S of the state space in one step, given that it is initially in S. Thus Φ measures the ability of $\mathfrak{M}$ to escape from any small region of the state space, and hence to make rapid progress to the stationary distribution. It is not therefore very surprising that the conductance Φ would govern the rapid mixing properties of the chain, which in turn is related to the second-largest eigenvalue λ_1 (by Proposition 2.3). This is made precise in the following result from [36, 38]; related results appear in [3, 31, 33]. Note that the result proves that the conductance captures mixing rate up to square factors, and thus obtaining a good lower bound on Φ is *equivalent* to proving rapid mixing.

Theorem 3.1 *The second eigenvalue of a reversible chain satisfies*

$$1 - 2\Phi \leq \lambda_1 \leq 1 - \frac{\Phi^2}{2} \,. \tag{8}$$

Proof: We only prove the inequality $(1 - \lambda_1) \leq 2\Phi$ which shows, together with Proposition 2.3, implies that a large conductance (of the order of $1/\text{poly}(n)$ where n is the problem size) is necessary for rapid mixing. Our proof follows the elegant approach of Alon [3] who proved a similar result for expansion of unweighted graphs. A proof of the other direction: $(1-\lambda_1) \geq \frac{\Phi^2}{2}$, can be found in [38, 31].

In order to prove $\lambda_1 \geq (1-2\Phi)$, we use the characterization of Equation (5). The largest eigenvalue of P equals 1 and has π as its eigenvector. Define a vector $f \in \mathbb{R}^\Omega$ (specified as a real-valued function on Ω) as follows:

$$f(x) = \begin{cases} \pi(x)\pi(\bar{S}) & \text{if } x \in S \\ -\pi(x)\pi(S) & \text{if } x \notin S \end{cases}$$

Note that $\langle f, \pi \rangle = \sum_x \frac{f(x)\pi(x)}{\pi(x)} = \sum_x f(x) = 0$, hence by Equation (5), we have

$$\frac{\langle f, fP \rangle}{\langle f, f \rangle} \leq \lambda_1 \; . \tag{9}$$

Define $g(x) \stackrel{\text{def}}{=} \frac{f(x)}{\pi(x)}$. Now

$$\begin{aligned} \langle f, f \rangle &= \sum_x \frac{f^2(x)}{\pi(x)} = \sum_x g^2(x)\pi(x) \\ &= \sum_{x \in S} \pi(\bar{S})^2 \pi(x) + \sum_{x \in \bar{S}} (-\pi(S))^2 \pi(x) = \pi(S)\pi(\bar{S}) \end{aligned} \tag{10}$$

$$\begin{aligned} \langle f, fP \rangle &= \sum_x \frac{f(x) \sum_y f(y) P(y,x)}{\pi(x)} = \sum_{x,y} g(x) g(y) Q(x,y) \\ &= \sum_x g^2(x) \sum_y Q(x,y) + \sum_{x,y} g(x)(g(y)-g(x)) Q(x,y) \\ &= \sum_x g^2(x)\pi(x) + \Big(\sum_{x \in S} -\pi(\bar{S})(\pi(S)+\pi(\bar{S})) \sum_{y \in \bar{S}} Q(x,y) \Big) \\ &\qquad + \Big(\sum_{x \in \bar{S}} -\pi(S)(\pi(\bar{S})+\pi(S)) \sum_{y \in S} Q(x,y) \Big) \\ &= \langle f, f \rangle - Q(S, \bar{S}) = \pi(S)\pi(\bar{S}) - Q(S, \bar{S}). \end{aligned} \tag{11}$$

From (9), (10) and (11), we get that for any set S,

$$\frac{Q(S,\bar{S})}{\pi(S)\pi(\bar{S})} \geq 1 - \lambda_1.$$

Since $\pi(\bar{S}) \geq 1/2$, this implies $\Phi \geq \frac{1-\lambda_1}{2}$, as desired. □

Corollary 3.2 *Let $\mathfrak{M}$ be a finite, reversible, ergodic Markov chain with loop probabilities $P(x,x) \geq 1/2$ for all states x, and let Φ be the conductance of $\mathfrak{M}$. Then the mixing time of $\mathfrak{M}$ satisfies $\tau_x(\varepsilon) \leq 2\Phi^{-2}(\ln \pi(x)^{-1} + \ln \varepsilon^{-1})$.*

A direct analysis of the conductance is sometimes possible by exploiting an underlying geometric interpretation of $\mathfrak{M}$, in which states of $\mathfrak{M}$ are identified with certain polytopes, and transitions with their common facets. A lower bound on conductance then follows from an appropriate "isoperimetric inequality" of the graph under consideration. This has been fruitful in a few applications, for example the estimation of the volume of a convex body by Dyer, Frieze and Kannan [11], and a Markov chain over linear extensions of a partial order by Karzanov and Khachiyan [26]. A more recent example where the conductance is tackled "directly" is the work of Dyer, Frieze and Jerrum [10] who prove an upper bound on Φ to show that certain classes of Markov chains for sampling independent sets in sparse graphs do not mix rapidly. The conductance is still not very amenable to computation in general, and we need further tools that can be used to deduce good lower bounds on the conductance. It is this task to which we turn next.

4 Rapid mixing via canonical paths

We saw in the last section that in order to prove rapid mixing of a Markov chain, all we need is a good lower bound on the conductance (and hence the spectral gap) of the chain. In this section, we explore a useful piece of technology developed in [20, 36, 37] to prove such a lower bound. The basic idea behind the method is to try and associate *canonical paths* between every pair of states, in such a way that no transition of the chain is used by too many paths. Intuitively, if such a set of paths exists, this means that the chain has no severe bottlenecks which could impede mixing. We now turn to formalizing this intuition.

4.1 Bounding Conductance using Canonical paths

We first formalize some terminology and notation. Let $\mathfrak{M}$ be an ergodic Markov chain on a finite set Ω. We define the weighted directed graph $G(\mathfrak{M})$ with vertex set Ω and with an edge e between an ordered pair (x, y) of weight $Q(e) = Q(x, y) = \pi(x)P(x, y)$ whenever $P(x, y) > 0$. We call this the underlying graph of $\mathfrak{M}$.

A set of *canonical paths* for $\mathfrak{M}$ is a set Γ of simple paths $\{\gamma_{xy}\}$ in the graph $G(\mathfrak{M})$, one between each ordered pair (x, y) of distinct vertices. In order to bound the conductance, we would like to have a set of canonical paths that do not overload any transition of the Markov chain. To measure this "overloading", we define the *path congestion* parameter [20, 36] for a set of canonical paths Γ as:

$$\rho(\Gamma) = \max_{e \in G(\mathfrak{M})} \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y), \tag{12}$$

where the maximum is over all oriented edges e in $G(\mathfrak{M})$, and $Q(e) = Q(x, y)$ if $e = (x, y)$. Think of the Markov chain as a flow network in which $\pi(x)\pi(y)$ units of flow travel from x to y along γ_{xy}, and $Q(e)$, which equals the probability that the Markov chain in the stationary distribution will use the transition e in a single step, serves as the capacity of e. The quantity $\rho(\Gamma)$ measures the maximum overloading of any edge relative to its capacity. The path congestion $\rho = \rho(\mathfrak{M})$ of the chain $\mathfrak{M}$ is defined as the minimum congestion achievable by any set of canonical paths, i.e.,

$$\rho = \inf_{\Gamma} \rho(\Gamma). \tag{13}$$

The following simple result confirms our intuition that a set of paths with low congestion implies a large value of conductance.

Lemma 4.1 *For any reversible Markov chain and any set of canonical paths Γ, we have*

$$\Phi \geq \frac{1}{2\rho(\Gamma)} .$$

Proof: Pick $S \subset \Omega$ with $0 < \pi(S) \leq 1/2$ such that $\Phi = \frac{Q(S,\bar{S})}{\pi(S)}$. For any choice of paths Γ, the total flow from S to $\bar{S}$ is $\pi(S)\pi(\bar{S})$, and all this must flow across the cut $[S : \bar{S}]$, which has capacity $Q(S, \bar{S})$. Hence there must exist an edge e in the cut $[S : \bar{S}]$ such that

$$\frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y) \geq \frac{\pi(S)\pi(\bar{S})}{Q(S, \bar{S})} \geq \frac{\pi(S)}{2Q(S, \bar{S})} = \frac{1}{2\Phi} . \qquad \square$$

Corollary 4.2 *For any reversible Markov chain, and any choice of canonical paths Γ, the second-largest eigenvalue λ_1 satisfies*

$$\lambda_1 \leq 1 - \frac{1}{8\rho^2(\Gamma)} . \tag{14}$$

4.2 Relating Spectrum to congestion directly

Since the relation between ρ and $(1 - \lambda_1)$ above proceeded by appealing to the conductance, the bound of Corollary 4.2 is potentially rather weak because of the appearance of the square. So we now pursue a direct approach to bound λ_1 based on the existence of "good" canonical paths. This was first achieved by Diaconis and Strook [8], but we follow a treatment by Sinclair [37] as it gives the best bounds for the examples considered later.

In order to state the new bound on λ_1, we modify the measure $\rho(\Gamma)$ to take into account the lengths of the paths as well. For a set $\Gamma = \{\gamma_{xy}\}$ of canonical paths, the key quantity is now

$$\bar{\rho}(\Gamma) = \max_e \frac{1}{Q(e)} \sum_{\gamma_{xy} \ni e} \pi(x)\pi(y)|\gamma_{xy}|, \tag{15}$$

where $|\gamma_{xy}|$ stands for the length of the path γ_{xy}. The parameter $\bar{\rho}$ is defined analogously to Equation (13) by minimizing over the choice of Γ.

Note that it is reasonable to allow general length functions $l(e)$ on the edges e, compute $|\gamma_{xy}|$ in terms of this length function, and thus obtain a quantity similar to $\bar{\rho}(\Gamma)$ above. In fact, Diaconis and Strook use the length function $l(e) = 1/Q(e)$, and Kahale [24] considers good length functions that will lead to the best bounds for specific chains. We will be content with the unit length function for our purposes.

Intuitively, the existence of short paths which do not overload any edge should imply that the chain mixes very rapidly. Indeed, it turns out that the variational characterization (6) can now be used to bound λ_1 directly in terms of $\rho(\bar{\Gamma})$. This is stated in the theorem below; we will not prove this theorem, but will later prove a more general version of this result (namely Theorem 4.6, which also appears in [37].

Theorem 4.3 ([37]) *For any reversible Markov chain, and any choice of canonical paths Γ, the second-largest eigenvalue λ_1 satisfies*

$$\lambda_1 \leq 1 - \frac{1}{\bar{\rho}(\Gamma)} . \tag{16}$$

A useful way to use the above result is the following version which bounds the spectral gap in terms of the path congestion ρ and the length of a longest path used in Γ. This version of the result is the most used in bounding mixing times using this approach.

Corollary 4.4 *For any reversible Markov chain, and any choice of canonical paths Γ, the second-largest eigenvalue λ_1 satisfies*

$$\lambda_1 \leq 1 - \frac{1}{\rho(\Gamma)\ell} . \tag{17}$$

where $\ell = \ell(\Gamma)$ is the length of a longest path in Γ.

The above often leads to much sharper bounds on mixing times than (14) because the maximum path length ℓ will usually be significantly lesser than the estimate obtained for ρ.

4.3 Known applications of canonical paths

The "canonical paths" approach has been applied successfully to analyze a variety of Markov chains including those for sampling perfect matchings and approximating the permanent [20, 8], estimating the partition function of the Ising model [21], sampling bases of balanced matroids [17], sampling regular bipartite graphs [25], sampling 0-1 knapsack solutions [12], etc. All these papers with the exception of [17] use more or less the same technique to bound the path congestion that is due to [20] – they use the state space to somehow "encode" the paths that use any given transition, so that the number of paths through any edge will be comparable to the number of states of the chain. Feder and Mihail [17] give a random collection of canonical paths and use a variant of "Hall's condition" (for existence of perfect matchings in bipartite graph) to show a small expected congestion and maximum path length for this collection of paths. They also prove a version of Corollary 4.4 which applies with expected path lengths and congestion instead of worst case values.

4.4 Path congestion is weaker than Conductance

The canonical paths technique is very useful, but it is natural to ask whether, like conductance, it too *captures* rapid mixing up to some polynomial factor (recall that conductance captures mixing time up to square factors). In other words, does a large conductance or a large spectral gap $(1 - \lambda_1)$ always imply a small value of $\rho(\Gamma)$ for some choice of canonical paths Γ? Unfortunately we give a simple example below to show that the answer is no — the same example also appears in [37].

Example. Consider the complete bipartite graph $K_{2,n-2}$ on vertex set $\{1, 2, \ldots, n\}$ and edges $\{(1, i), (2, i) : 3 \leq i \leq n\}$ where n is even, and define transition probabilities corresponding to the random walk on this graph, namely at each step stay where you are with probability $1/2$, else move to a neighbor chosen uniformly at random. The stationary distribution π of this Markov chain is given by: $\pi(1) = \pi(2) = 1/4$ and $\pi(i) = 1/2(n-2)$ for $i = 3, 4, \ldots, n$, and hence $Q(e) = 1/4(n-2)$ for all edges e. Since n is even it is easy to verify that the conductance of this chain is $\Phi = 1/2$, and hence using Equation (8) we get $\lambda_1 \leq 7/8$. However, since $\pi(1)\pi(2) = 1/16$ and $Q(e) = 1/4(n-2)$ for all edges e, the path connecting states 1 and 2 alone implies that the best value for $\rho(\Gamma)$ or $\bar{\rho}(\Gamma)$ obtainable using canonical paths is $\Omega(n)$. Hence ρ and $\bar{\rho}$ could in fact be much larger than the quantity $(1 - \lambda_1)^{-1}$ which governs the mixing time.

4.5 Resistance: a generalization of path congestion

In order to alleviate the shortcoming of the canonical paths technique which was just discussed, we now present a natural generalization of this approach that will end up capturing mixing times exactly (and will thus be "as good as" conductance). The idea, again due to Sinclair [37], is to *spread the flow* on path γ_{xy} between a pair (x, y) of states among *several paths*. As before, we view $G(\mathfrak{M})$ as a flow network where one unit of flow has to be routed from x to y for every ordered pair (x, y) of distinct vertices, and each (oriented) edge e has "capacity" $Q(e)$. The difference from the canonical paths approach is that, we now allow the flow between x and y to be *split* among multiple paths, i.e., we are looking for a *fractional multicommodity flow* that minimizes the congestion. Considering the similarity with the earlier approach, it is natural to suppose that this new measure will yield similar bounds on the mixing rate. As we shall see, this will be the case, and in fact this seemingly innocuous generalization to multiple paths allows us to capture rapid mixing exactly!

Formally, a *flow* in $G(\mathfrak{M})$ is a function $f : \mathcal{P} \to \mathbb{R}^+$ which satisfies

$$\sum_{p \in \mathcal{P}_{xy}} f(p) = 1 \qquad \text{for all } x, y \in X, x \neq y,$$

where $\mathcal{P}_{xy}$ is the set of all simple directed paths from x to y in $G(\mathfrak{M})$ and $\mathcal{P} = \cup_{x \neq y} \mathcal{P}_{xy}$. The quality of a flow is measured by the *congestion parameter* $\mathfrak{R}(f)$, defined analogously to Equation (12) by

$$\mathfrak{R}(f) \stackrel{\text{def}}{=} \max_e \frac{1}{Q(e)} \sum_{x,y} \sum_{p \in \mathcal{P}_{xy}: p \ni e} \pi(x)\pi(y)f(p), \tag{18}$$

and one can define *elongated congestion* $\bar{\mathfrak{R}}(f)$, similar to Equation 15, by accounting for the lengths of the paths:

$$\bar{\mathfrak{R}}(f) \stackrel{\text{def}}{=} \max_e \frac{1}{Q(e)} \sum_{x,y} \sum_{p \in \mathcal{P}_{xy}: p \ni e} \pi(x)\pi(y)f(p)|p|. \tag{19}$$

We have the following results parallel to those of Lemma 4.1, Corollary 4.2, Theorem 4.3 and Corollary 4.4.

Lemma 4.5 *For any reversible Markov chain and any flow f, we have*

$$\Phi \geq \frac{1}{2\mathfrak{R}(f)} \quad \textit{and hence} \quad \lambda_1 \leq 1 - \frac{1}{8\mathfrak{R}(f)^2}\,.$$

Theorem 4.6 *For any reversible Markov chain, and any flow f, the second-largest eigenvalue λ_1 satisfies*

$$\lambda_1 \leq 1 - \frac{1}{\bar{\mathfrak{R}}(f)}\,. \tag{20}$$

Corollary 4.7 *For any reversible Markov chain, and any flow f, the second-largest eigenvalue λ_1 satisfies*

$$\lambda_1 \leq 1 - \frac{1}{\mathfrak{R}(f)\ell(f)}\,. \tag{21}$$

where $\ell(f)$ is the length of a longest path p with $f(p) > 0$.

We now provide a proof of Theorem 4.6 as we had promised before the statement of Theorem 4.3 (note that the statement of Theorem 4.6 clearly generalizes that of Theorem 4.3).

Proof of Theorem 4.6: We need to prove $(1-\lambda_1) \geq 1/\bar{\mathfrak{R}}(f)$ for any flow f. We use Equation (6) to bound $(1-\lambda_1)$, namely

$$1 - \lambda_1 = \inf_\psi \frac{\sum_{x,y\in\Omega}(\psi(x)-\psi(y))^2 Q(x,y)}{\sum_{x,y\in\Omega}(\psi(x)-\psi(y))^2\pi(x)\pi(y)}. \tag{22}$$

Now for any ψ, and any flow f, the denominator in the above expression can be written as:

$$\begin{aligned}
\sum_{x,y\in\Omega}(\psi(x)-\psi(y))^2\pi(x)\pi(y) &= \sum_{x,y}\pi(x)\pi(y)(\psi(x)-\psi(y))^2 \sum_{p\in\mathcal{P}_{xy}} f(p) \\
&= \sum_{x,y}\pi(x)\pi(y)\sum_{p\in\mathcal{P}_{xy}} f(p)\Big(\sum_{e\in p}(\psi(e^+)-\psi(e^-))\Big)^2 \\
&\leq \sum_{x,y}\pi(x)\pi(y)\sum_{p\in\mathcal{P}_{xy}} f(p)|p|\sum_{e\in p}(\psi(e^+)-\psi(e^-))^2 \\
&= \sum_e(\psi(e^+)-\psi(e^-))^2\sum_{x,y}\sum_{p\in\mathcal{P}_{xy}:p\ni e}\pi(x)\pi(y)f(p)|p| \\
&\leq \sum_e(\psi(e^+)-\psi(e^-))^2 Q(e)\bar{\mathfrak{R}}(f) \\
&= \bar{\mathfrak{R}}(f)\sum_{x,y} Q(x,y)(\psi(x)-\psi(y))^2.
\end{aligned}$$

(Here e^- and e^+ denote the start and end vertices of the oriented edge e, and we have used Cauchy-Schwartz inequality in the third step above.) The result now follows from (22). □

Definition 4.1 (Resistance) *The* resistance $\mathfrak{R} = \mathfrak{R}(\mathfrak{M})$ *of chain* $\mathfrak{M}$ *is defined as the minimum value of* $\mathfrak{R}(f)$ *over all flows* f*, and like the conductance is an invariant of the chain. Formally,*

$$\mathfrak{R} = \inf_f \mathfrak{R}(f) \ . \tag{23}$$

4.6 Resistance captures rapid mixing

By Lemma 4.5, note that $\lambda_1 \le 1 - \frac{1}{8\mathfrak{R}^2}$, so a small resistance leads to rapid mixing. We will now see that in fact the converse is true, in other words a small mixing time implies a small resistance, i.e., the existence of a flow f with small congestion $\mathfrak{R}(f)$. Thus resistance *overcomes* the shortcoming of path congestion (since low path congestion was not a necessary condition for rapid mixing, as was shown by the example in Section 4.4).

Theorem 4.8 ([37]) *Consider an irreducible, reversible, ergodic Markov chain* $\mathfrak{M}$ *over* Ω *and let* $\tau = \max_{x\in\Omega} \tau_x(1/4)$*. Then the resistance* $\mathfrak{R} = \mathfrak{R}(\mathfrak{M})$ *of* $\mathfrak{M}$ *satisfies* $\mathfrak{R} \le 16\tau$*.*

Proof: We will demonstrate a flow f with $\mathfrak{R}(f) \le 16\tau$. Let $t = 2\tau$. The flow between x and y will be routed as follows: Consider the set $\mathcal{P}_{xy}^{(t)}$ of all (not necessarily simple) paths of length t from x to y in $G(\mathfrak{M})$, and for each $p \in \mathcal{P}_{xy}^{(t)}$ route $f(p) \propto \mathrm{prob}(p)$ units of flow on p, where $\mathrm{prob}(p)$ is the probability that the Markov chain makes the sequence of transitions defined by p in the first t steps when starting in state x. Since $t = 2\tau$, it is easy to see that for any pair x, y, $\mathcal{P}_{xy}^{(t)} \neq \emptyset$, and in fact

$$\frac{P^t(x,y)}{\pi(y)} \ge \frac{1}{8} \ . \tag{24}$$

Thus for $p \in \mathcal{P}_{xy}^{(t)}$, we have $f(p) = \mathrm{prob}(p)/(\sum_{q\in\mathcal{P}_{xy}^{(t)}} \mathrm{prob}(q)) = \mathrm{prob}(p)/P^t(x,y)$. Now let us estimate the $\mathfrak{R}(f)$.

$$\begin{aligned}
\mathfrak{R}(f) &= \max_e \frac{1}{Q(e)} \sum_{x,y} \sum_{p\in\mathcal{P}_{xy}^{(t)}:p\ni e} \frac{\pi(x)\pi(y)\mathrm{prob}(p)}{P^t(x,y)} \\
&\le \max_e \frac{8}{Q(e)} \sum_{x,y} \sum_{\substack{p\in\mathcal{P}_{xy}^{(t)} \\ p\ni e}} \pi(x)\mathrm{prob}(p) \quad \text{(using (24))} \\
&\le \max_e \frac{8}{Q(e)} \cdot tQ(e) = 8t = 16\tau
\end{aligned}$$

where we used the fact that the final double summation is simply the probability that the Markov chain traverses the edge e within t steps when started in the stationary distribution π over Ω, and this probability, by the union bound, is at most t times the probability that this happens in one step, and is thus at most $tQ(e)$. □

Remark A. It is also possible to prove (see [37]), using techniques of the approximate max-flow min-cut theorem for uniform multicommodity flow [28], that $\lambda_1 \ge 1 - O(\frac{\log N}{\mathfrak{R}})$. This gives the weaker bound $\tau = \Omega(\mathfrak{R}/\log N)$, but is interesting in its own right.

Remark B. Note that since we used paths of length 2τ in the above proof, the flow f also satisfies $\bar{\mathfrak{R}}(f) = O(\tau^2)$. This, together with (20), implies that $\bar{\mathfrak{R}} = \inf_f \bar{\mathfrak{R}}(f)$ captures rapid mixing as well. The work of Kahale [24] actually shows that the bound on $\bar{\mathfrak{R}}(f)$, call it μ, obtained by minimizing over all length functions on the transitions and all flows, can be computed to arbitrary precision by reduction to a semidefinite program, and satisfies $\lambda_1 \geq 1 - O(\frac{\log^2 N}{\mu})$.

4.7 Path congestion almost always captures rapid mixing!

In the next section, we will see a resistance based proof (due to [34]) of rapid mixing of a natural Markov chain for sampling 0-1 Knapsack solutions. This problem was open for a long time, and had defied all attempts to prove rapid mixing based on canonical paths. In light of the example in Section 4.4, it is natural to ask if this chain (which we now know mixes rapidly) also cannot have low path congestion, and whether the generalization to resistance was really necessary.

In this section, we will show that, for a broad class of Markov chains, including all the ones we consider in applications here, the path congestion ρ (defined in Equations (12) and (13)) *characterizes rapid mixing* up to polynomial (in the problem size) factors. We show that if you can achieve low congestion with multiple paths, i.e., if the chain has low resistance, then you can also achieve low congestion by routing all the flow on just a single path. The proof is actually very simple, and is based on randomized rounding to relate the optimum congestion of "fractional" and "unsplittable" flows, but we were surprised that it does not seem to have been observed or made explicit in the literature.

Theorem 4.9 *Consider an ergodic, reversible Markov chain $\mathfrak{M}$ with stationary distribution π on a state space Ω of size N, and let the resistance of $\mathfrak{M}$ be $\mathfrak{R}$. Let $\Lambda = \max_{x \neq y} \pi(x)\pi(y)$, and let $Q_{\min} = \min_{e:Q(e)>0} Q(e)$. Then there exists a set of canonical paths Γ such that*

$$\rho(\Gamma) = O\Big(\mathfrak{R} + \log N \frac{\Lambda}{Q_{\min}}\Big).$$

Proof: By the definition of the resistance $\mathfrak{R}$, we know that there exists a flow f which routes $\pi(x)\pi(y)$ units of flow between every ordered pair (x, y) of distinct states $x \neq y$, such that every (oriented) edge e has at most $Q(e)\mathfrak{R}$ units of flow passing through it. Hence there is a feasible fractional flow f which routes $f_{xy} = \pi(x)\pi(y)/\Lambda \leq 1$ units of flow between x and y, and with "capacity" on edge e at most $C(e) = \max\{\frac{Q(e)\mathfrak{R}}{\Lambda}, 1\}$. We can now use a result of Raghavan and Thompson [35], who used randomized rounding to show the following: There are absolute constants b_0 and b_1 such that if all edge capacities equal 1, and all demands are at most 1, and there is a fractional flow satisfying all the demands with congestion on edge e at most $\mu^f(e) \geq 1$, then there is an *unsplittable* flow which satisfies all the demands by routing the demand for each source-destination pair along a *single* path, and which has congestion at most $b_0\mu^f(e) + b_1 \log N$ on edge e.

Applying this to our situation with $\mu^f(e) = C(e)$, we conclude that there exists a set Γ of canonical paths which can route f_{xy} units of flow from x to y such that at most $b_0 C(e) + b_1 \log N$ units flow through any edge e, or equivalently, it can route $\pi(x)\pi(y) = \Lambda f_{xy}$ units of flow between every pair (x, y) such that at most $b_0 \Lambda C(e) + b_1 \Lambda \log N$ units flow through any edge e. This implies that

$$\rho(\Gamma) \leq b_0 \max\{\mathfrak{R}, \frac{\Lambda}{Q_{\min}}\} + b_1 \frac{\Lambda}{Q_{\min}} \log N\ .$$

and the stated result follows. □

Theorem 4.9 actually implies that $\rho = O(\mathfrak{R})$ for a wide variety of Markov chains, and thus for these chains ρ also characterizes rapid mixing. Indeed, this will be the case whenever $\log N \frac{\Lambda}{Q_{\min}} = O(1)$, which will normally always be the case unless the stationary distribution varies widely in the mass it gives to points of the state space, or there are very small non-zero transition probabilities in the chain. As an example consider Markov chains with uniform stationary distribution. Then $\log N \frac{\Lambda}{Q_{\min}} = O(1)$ whenever $P(x,y) = \Omega(\frac{\log N}{N})$ for all x, y such that $P(x,y) > 0$. For most chains in applications to sampling, we will have $N = 2^{O(n)}$ where n is the problem size and each non-zero $P(x,y)$ will be at least $1/\text{poly}(n)$, hence this condition will indeed be met.

5 Sampling 0-1 Knapsack solutions

We describe an example of random walk on the truncated hypercube which was only very recently shown to be rapidly mixing using a fractional multicommodity flow with low congestion [34], but had resisted all efforts of proving such a result using canonical paths (with just one path between every source-destination pair). Our result from the previous section (Theorem 4.9) applies to this chain; this shows that even though spreading flow across multiple paths might in principle be not more powerful than sending all the flow along a single canonical path, it could be still be easier to deal with in actually designing the flow. (The example from this section is also covered by the framework of what Feder and Mihail [17] did, where they prove a version of the small path congestion implies small mixing time result using expected path lengths and congestion instead of worst case values.)

The Problem. We are interested in sampling from the set Ω of feasible solutions to the 0-1 knapsack problem defined by the vector $\mathbf{a}$ of item sizes and the knapsack capacity b; i.e., for a positive real vector $\mathbf{a} = (a_i)_{i=1}^n$ and a real number b,

$$\Omega = \Omega_{\mathbf{a},b} = \{\mathbf{x} \in \{0,1\}^n : \mathbf{a} \cdot \mathbf{x} = \sum_{i=1}^n a_i x_i \leq b\} .$$

There is a one-one correspondence between vectors $\mathbf{x} \in \Omega$ and subsets X of items whose aggregated weight does not exceed b, given by $X = \{i : x_i = 1\}$. We will write $a(X)$ for the weight of X, i.e., $a(X) = \sum_{i \in X} a_i$.

A particularly simple Markov chain $\mathfrak{M}_K$ on Ω has been proposed for the purposes of sampling uniformly at random from Ω. If the current state is $X \subseteq \{1, 2, \ldots, n\}$ then

1. With probability $1/2$ stay at X (this holding probability is to make the chain aperiodic), else
2. Pick an item $i \in \{1, 2, \ldots, n\}$ uniformly at random. If $i \in X$ move to $X - \{i\}$; if $i \notin X$ and $a(X \cup \{i\}) \leq b$, move to $X \cup \{i\}$, else stay at X.

The chain is aperiodic since $P(X, X) \geq 1/2$ for all states X, and it is irreducible since every pair of states can be connected via the empty set. Moreover, it is clear that each non-zero transition probability $P(X,Y)$, $X \neq Y$, equals $P(X,Y) = P(Y,X) = \frac{1}{2n}$. By Theorem 2.1 and Lemma 2.2 therefore, $\mathfrak{M}_K$ is ergodic with uniform stationary distribution. Despite all the recent activity in proving rapid mixing, this simple example was not known to be rapidly mixing until the work of [34]. The best prior known bound on the mixing time, obtained via the canonical paths technique, was $\exp(O(\sqrt{n}(\log n)^{5/2}))$ [12], which beats the trivial bound of $\exp(O(n))$ but is still exponential.

We will now sketch the proof of [34] that this chain has a mixing time of $O(n^8)$, and is thus indeed rapidly mixing. The proof will follow the *resistance* approach, i.e., we will find a flow f that routes one unit of flow between every pair of unequal states, using multiple paths for each pair to "spread" the flow, and then use Corollary 4.7 to bound the mixing time. Indeed, if $L(f)$ is the length of the longest flow carrying path, and $C(f)$ is the maximum flow across any (oriented) edge of the chain, then combining Corollary 4.7 and Proposition 2.3 shows that

$$\tau_X(\varepsilon) \leq 2n\frac{C(f)}{|\Omega|}L(f)(n + \ln \varepsilon^{-1}) \;. \tag{25}$$

Hence our goal now is to construct a flow f with $L(f) = \mathrm{poly}(n)$ and $C(f) = |\Omega|\mathrm{poly}(n)$. Note that a shortest path between states X and Y can be viewed as a permutation of the symmetric difference $X \oplus Y$, the set of items that must be added to or removed from the knapsack in passing from X to Y. A natural approach to defining a good flow seems to be to spread the unit flow from X to Y evenly among all permutations of $X \oplus Y$. The problem with this approach, however, is that many of these permutations will tend to violate the knapsack constraint, as too many items will have been added at some intermediate point; i.e., the permutation is unbalanced. The way to circumvent this problem is to define a family of permutations, which are all "balanced" and also "sufficiently random", and spread the flow evenly among them. Proving the existence of such permutations, called *balanced almost uniform permutations* in [34], forms the main technical component of this proof.

We will now define the notion of balanced almost uniform permutations formally, and state the Theorems from [34] guaranteeing their existence. (We will not prove these theorems as they are quite technical and doing so will take us too far away from our main theme of focusing on Markov chain techniques.) We will, however, show how to construct a good flow f for our purposes given the existence of the necessary balanced almost uniform permutations.

Definition 5.1 *Let $\{w_i\}_{i=1}^m$ be a set of real weights, and let $M = \max_{i \leq m} |w_i|$ and $W = \sum_i w_i$. Let ℓ be a non-negative integer. A permutation $\sigma \in S_m$ is ℓ-balanced, if for all k, $1 \leq k \leq m$,*

$$\min\{W, 0\} - \ell M \leq \sum_{i=1}^{k} w_{\sigma(i)} \leq \max\{W, 0\} + \ell M \;. \tag{26}$$

Definition 5.2 *Let σ be a random variable taking values in S_m, and let $\alpha \in \mathbb{R}$. We call σ a α-uniform permutation if*

$$\Pr_{\sigma}[\sigma\{1, 2, \ldots, k\} = U] \leq \alpha \times \binom{m}{k}^{-1}$$

for every k, $1 \leq k \leq m$, and every $U \subseteq \{1, 2, \ldots, m\}$ of size k.

The main theorem from [34] on the existence of balanced almost uniform permutations is the following:

Theorem 5.1 ([34]) *There is a universal constant C such that for any m and any set of weights $\{w_i\}_{i=1}^m$, there exists a 7-balanced Cm^2-uniform permutation on $\{w_i\}$. Moreover, if $|\sum_i w_i| > 15 \max_i |w_i|$, then there exists a 0-balanced Cm^2-uniform permutation on $\{w_i\}$.*

Constructing a good flow

Lemma 5.2 *For arbitrary weights $\{a_i\}$ and b, there exists a multicommodity flow f in $G(\mathfrak{M}_K)$ which routes one unit of flow between every pair of unequal vertices, with $C(f) = O(|\Omega| n^5)$ and $L(f) = O(n)$.*

Combining with Equation (25) we therefore conclude

Theorem 5.3 ([34]) *The mixing time of the Markov chain $\mathfrak{M}_K$ satisfies $\tau_X(\varepsilon) = O(n^8 \ln \varepsilon^{-1})$ for every starting state X. The chain is thus rapidly mixing.*

Proof of Lemma 5.2: Let X, Y be arbitrary states of Ω, $X \neq Y$. We wish to send one unit of flow from X to Y. As discussed earlier, our idea is to spread this flow evenly among a family of balanced almost uniform permutations of $X \oplus Y$, except that we isolate a constant number of "heavy" items H from $X \oplus Y$, and route the flow along balanced almost uniform permutations of $(X \oplus Y) \setminus H$ and add or remove some elements of H repeatedly along the path to maintain fine balance (we always want the knapsack to be filled to capacity between (roughly) $\min\{a(X), a(Y)\}$ and $\max\{a(X), a(Y)\}$: an upper bound on the weight packed in the knapsack is clearly necessary to define a feasible path, while the lower bound is used in the analysis to bound the total flow through any edge by "encoding" each flow path which uses that edge using an element of the state space).

We now proceed with the formal analysis. We wish to obtain an upper bound on the maximum flow that passes through any state Z (this will clearly also provide an upper bound on the flow through any transition (Z, Z_1) of the chain). Let X, Y be states such that the flow between them passes through Z. Let H be the 29 elements of $X \oplus Y$ with the largest weight (set $H = X \oplus Y$ if $|X \oplus Y| \leq 29$); breaking ties *according to index order*. Define $H_X = H \cap X$, $H_Y = H \cap Y$, $S = (X \oplus Y) \setminus H$ and $m = |S|$. Let $\{w_i\}_{i=1}^m$ be an arbitrary enumeration of the weights of items in S, where elements in Y receive positive signs and those in X receive in negative signs (since we want to add elements in $S \cap Y$ and remove those in $S \cap X$). The paths we use for our flow will correspond to permutations of indices in S that satisfy the specific "balance" requirement described below.

Claim. *There is an absolute constant C such that there exists a Cm^2-uniform family of permutations each one (call it σ) of which satisfies the following "balance" condition:*

$$\min\{a(Y) - a(X), 0\} - a(H_Y) \leq \sum_{i=1}^{k} w_{\sigma(i)} \leq \max\{a(Y) - a(X), 0\} + a(H_X), \tag{27}$$

for every k, $1 \leq k \leq m$.

Proof. We will assume $|X \oplus Y| \geq 29$, for otherwise $S = \emptyset$ and $m = 0$, and there is nothing to prove. Let $W = \sum_{i=1}^m w_i = a(Y) - a(X) + a(H_X) - a(H_Y)$, and $M = \max_i |w_i|$. Let us assume, w.l.o.g, that $W \geq a(H_X) - a(H_Y)$ (the other case is symmetric), so it is easy to see that the above condition (27) is equivalent to

$$-a(H_Y) \leq \sum_{i=1}^{k} w_{\sigma(i)} \leq W + a(H_Y) \tag{28}$$

Comparing with condition (26), and allowing for both cases $W \geq 0$ and $W < 0$, it is easy to see that an ℓ-balanced permutation satisfies (28) above whenever $\ell M \leq \min\{a(H_Y), W + a(H_Y)\}$. Thus, when $|W| > 15M$, we can use 0-balanced permutations guaranteed by Theorem 5.1 for our purposes. When $|W| \leq 15M$, we have $a(H_X) - a(H_Y) \leq W \leq 15M$. Also $a(H_X) + a(H_Y) = a(H) \geq 29M$. Combining these two inequalities we get $a(H_Y) \geq 7M$. Thus when $W \geq 0$, we have $7M \leq \min\{a(H_Y), W + a(H_Y)\}$, and thus we can use a Cm^2-uniform family of 7-balanced permutations to satisfy (28). When $W < 0$, we have $W \geq -15M$ and together with $a(H_X) - a(H_Y) \leq W$ this implies

$$a(H_Y) \geq \frac{29M - W}{2} \geq \frac{14M - 2W}{2} = 7M - W$$

and thus once again $7M \leq \min\{a(H_Y), W + a(H_Y)\}$, and we can use a 7-balanced Cm^2-uniform family of permutations. $\square$ *(Claim)*

We now specify the flow paths between X and Y (the flow will be evenly split among all these paths). The paths will follow the permutations σ of the family guaranteed by above Claim, except that along the way we will use elements of H to keep the knapsack as full as possible, and we will remove elements of H as necessary to make room for elements of $S \cap Y$ to be added. Hence each intermediate state will be of the form $H_0 \cup ((X \setminus H_X) \oplus \{\sigma(1), \ldots, \sigma(k)\})$ for some $k \leq m$ and $H_0 \subseteq H$. The path corresponding to a particular σ is defined by the following transitions:

- If $k < m$ and $w_{\sigma(k+1)} > 0$, then add $\sigma(k+1)$ if possible (i.e., current knapsack has room for the item); else delete an (arbitrary) element from H_0.
- If $k < m$ and $w_{\sigma(k+1)} < 0$, then add an element form $H - H_0$ if possible (so that knapsack is near full); otherwise remove $\sigma(k+1)$.
- If $k = m$ (i.e., all elements in S have been handled), add an element of H_Y if possible; otherwise delete an element from H_X.

By the upper bound of Condition (27), we have $a(X) - a(H_X) + \sum_{i=1}^{k+1} w_{\sigma(i)} \leq \max\{a(X), a(Y)\} \leq b$ so that we can always remove enough elements of H to make room for $w_{\sigma(k+1)}$ during its turn to be added. Moreover, the lower bound of Condition (27) implies that for any intermediate state Z on any flow path, $a(Z \cup H) \geq \min\{a(X), a(Y)\}$, and since we always keep the knapsack as full as possible, there exist elements $h_1, h_2 \in H$ such that $a(Z \cup \{h_1, h_2\}) \geq \min\{a(X), a(Y)\}$. In what follows h_1, h_2 are fixed elements of H that depend only on Z, X, Y.

To estimate the flow through Z, we will "encode" each pair X, Y of states whose flow paths use Z by a state $Z' \in \Omega$ (plus some auxiliary information), so that we can argue that $C(f)$ is not too large compared to $|\Omega|$. The encoding Z' is defined by

$$Z' = ((X \oplus Y) \setminus (Z \cup \{h_1, h_2\})) \cup (X \cap Y).$$

(Note that this is the complement of $Z \cup \{h_1, h_2\}$ in the multiset $X \cup Y$. Thus it is reasonable to expect that Z' will supply a lot of the "missing" information about X, Y that cannot be obtained from Z, h_1, h_2.) Now

$$\begin{aligned} a(Z') &= a(X) + a(Y) - a(Z \cup \{h_1, h_2\}) \\ &\leq a(X) + a(Y) - \min\{a(X), a(Y)\} \\ &= \max\{a(X), a(Y)\} \leq b \end{aligned}$$

so that $Z' \in \Omega$.

We now wish to upper bound the number of pairs (X, Y) that could be mapped to a given Z'. Note that $Z \cap Z' = X \cap Y$ and $Z' \oplus (Z \cup \{h_1, h_2\}) = X \oplus Y$, and knowing $X \oplus Y$, we also know H (since these form the 29 largest elements of $X \oplus Y$, ties broken according to index order). Thus Z, Z', h_1, h_2 together fix $X \cap Y$, $X \oplus Y$, H and $S = (X \oplus Y) \setminus H$. In order to completely specify X and Y, we add some more information to the encoding, namely the subset $U \subseteq S$ that have been "affected" (i.e., added/removed) by the time the path from X to Y reaches Z, and also $H' = H \cap X$.[2] Thus, the pair (X, Y) one of whose flow paths passes through Z is encoded by the 5-tuple:

$$f_Z(X, Y) = (Z', h_1, h_2, U, H')\ .$$

We now verify that Z and $f_Z(X, Y)$ do pinpoint X, Y. Indeed, we already argued Z, Z', h_1, h_2 alone fix $X \cap Y$, $X \oplus Y$, H and S. Now it is easy to verify that $X = (U \cap Z') \cup ((S \setminus U) \cap Z) \cup (X \cap Y) \cup H'$ and similarly $Y = (U \cap Z) \cup ((S \setminus U) \cap Z') \cup (X \cap Y) \cup (H \setminus H')$.

We are now ready to bound $C(f)$ by estimating the cumulative flow $f(Z)$ through Z. For each X, Y such that there is a flow path from X to Y passing through Z and whose encoding equals $f_Z(X, Y) =$

[2]The encoding U we use is slightly different from the one Morris and Sinclair [34] use in their proof.

(Z', h_1, h_2, U, H'), there will be non-zero flow only for paths corresponding to those permutations σ of $\{1, 2, \ldots, m\}$ (here $m = |S|$) that satisfy $\sigma\{1, 2, \ldots, |U|\} = U$. By the Cm^2-uniformity of the family of permutations we use to spread the flow, we can conclude that the total flow over all such paths is at most $Cm^2 \binom{m}{|U|}^{-1}$. Thus summing over all $U \subseteq S$, we still have only Cm^3 units of flow for each fixed (Z', h_1, h_2, H'). Now there are $|\Omega|$ choices for Z', and n^2 choices for the pair (h_1, h_2), and once (Z', h_1, h_2) are fixed, so is H, and thus there are at most 2^{29} possible choices of $H' \subseteq H$ for each choice of Z'. In all, we have

$$f(Z) \leq |\Omega| \cdot n^2 \cdot 2^{29} \cdot Cm^3 = O(|\Omega| n^5) .$$

Thus $C(f) = O(|\Omega| n^5)$ as well, and since all paths we use to route flows clearly have length $O(n)$, $L(f) = O(n)$, and the proof of Lemma 5.2 is complete. □

6 Coupling and Path Coupling

We have so far focused on conductance based techniques for proving rapid mixing, and saw a non-trivial application to sampling 0-1 knapsack solutions. The classical approach to bounding the mixing time is in fact via a different approach, viz. *Coupling*. The basic idea behind the coupling argument is very intuitive: suppose we wish to show that a Markov chain $\mathfrak{M}$ starting from distribution π' converges to its stationary distribution π within a small number of steps. Consider running the chain on a *joint* process $(\mathcal{X}, \mathcal{Y})$ where both $\mathcal{X}, \mathcal{Y}$ are individually *faithful* copies of $\mathfrak{M}$ and where $\mathcal{X}$ starts of at state X_0 distributed according to π' and $\mathcal{Y}$ starts of in state Y_0 distributed according to π. Thus at any time step t, the distribution of Y_t equals π. Now if the joint evolution of (X_t, Y_t) is designed to encourage them to *coalesce* rapidly, i.e., the "distance" between X_t and Y_t decreases rapidly, then for large enough t, say $t \geq t'$, we will have $X_t = Y_t$, with high probability, say $(1 - \varepsilon)$. Since the distribution of Y_t is π, it is easy to see that this implies that the mixing time to get within ε of the stationary distribution when the chain starts off in distribution π', is at most t' (by the "Coupling Lemma" which we will state and prove formally shortly).

6.1 The Coupling Lemma

Definition 6.1 (Coupling) *Let $\mathfrak{M}$ be a finite, ergodic Markov chain defined on state space Ω with transition probabilities $P(\cdot, \cdot)$. A (causal)* coupling *is a joint process $(\mathcal{X}, \mathcal{Y}) = (X_t, Y_t)$ on $\Omega \times \Omega$, such that each of the processes $\mathcal{X}$, $\mathcal{Y}$, considered marginally, is a faithful copy of $\mathfrak{M}$. In other words, we require that, for all $x, x', y, y' \in \Omega$,*

$$\mathbf{Pr}[X_{t+1} = x' | X_t = x \wedge Y_t = y] = P(x, x') \quad \textit{and,}$$
$$\mathbf{Pr}[Y_{t+1} = y' | X_t = x \wedge Y_t = y] = P(y, y') . \qquad \square$$

Note that the above conditions are consistent with (X_t) and (Y_t) being independent evolutions of $\mathfrak{M}$, but does not imply it. In fact the whole point of Coupling is to allow for the possibility that

$$\mathbf{Pr}[X_{t+1} = x' \wedge Y_{t+1} = y' | X_t = x \wedge Y_t = y] \neq P(x, x') P(y, y')$$

in order to encourage X_t and Y_t to *coalesce* rapidly.

Remark. In applications to bounding mixing time, (X_t) will typically be Markovian, while we allow $\mathcal{Y}$ to be Non-Markovian or history dependent, i.e., Y_t could depend upon $X_0, \ldots X_t$ and $Y_0, \ldots, Y_{t-1}$, as long as it remains faithful to the original chain $\mathfrak{M}$. One can also imagine allowing the process $\mathcal{Y}$ to make its moves dependent on future moves of $\mathcal{X}$, i.e., Y_t can depend upon X_{t+1}, X_{t+2}, etc. Such a coupling is called a *non-causal* coupling. We will only be concerned with causal couplings here, and the term "Coupling" will always refer only to a causal coupling.

If it can be arranged that coalescence occurs rapidly, independently of the initial states X_0, Y_0, we may then deduce that $\mathfrak{M}$ is rapidly mixing. The key result here is the *Coupling Lemma*, which seems to have first explicitly appeared in [2].

Lemma 6.1 (Coupling Lemma) *Let $\mathfrak{M}$ be a finite, ergodic Markov chain, and let (X_t, Y_t) be a coupling for $\mathfrak{M}$. Suppose that $\mathbf{Pr}[X_t \neq Y_t] \leq \varepsilon$, uniformly over the choice of initial state (X_0, Y_0). Then the mixing time $\tau(\varepsilon)$ of $\mathfrak{M}$ (starting from any state) is bounded above by t.*

Proof: Let $X_0 = x$ be arbitrary and let Y_0 be distributed according to the stationary distribution π of $\mathfrak{M}$. Let $A \subseteq \Omega$ be an arbitrary event. We have

$$\begin{aligned}
\mathbf{Pr}[X_t \in A] &\geq \mathbf{Pr}[Y_t \in A \wedge X_t = Y_t] \\
&\geq 1 - \mathbf{Pr}[Y_t \notin A] - \mathbf{Pr}[X_t \neq Y_t] \\
&\geq \mathbf{Pr}[Y_t \in A] - \varepsilon \\
&= \pi(A) - \varepsilon,
\end{aligned}$$

and this implies the variation distance between $P^t(x, \cdot)$ and π, $\Delta_x(t)$, is at most ε, as desired. $\square$

In light of the above Lemma, Coupling is a natural technique to prove rapid mixing of Markov chains. And as we will convince the reader in this section and the next, Coupling is a very crisp and elegant technique and when it works, it invariably establishes better bounds on mixing time than known through conductance, and avoids the slackness which is typical of conductance/canonical paths based proofs. We illustrate this by a simple example below.

6.2 An illustrative example of Coupling in action

We consider the "Bernoulli-Laplace diffusion model", whose state space Ω is the set of all k-element subsets of $[n] = \{1, 2, \ldots, n\}$, and we wish to sample an element u.a.r from Ω. We assume $k \leq n/2$ without loss of generality. A natural chain on Ω is the following (let the current state be the subset $X \subseteq [n]$ with $|X| = k$)

- Pick $r_X \in \{0, 1\}$ u.a.r; If $r_X = 0$, remain at X.
- If $r_X = 1$, pick $i \in X$ u.a.r and $j \in [n] \setminus X$ u.a.r and move to $Y = X \cup \{j\} \setminus \{i\}$.

It is easy to that this chain is ergodic with uniform stationary distribution $\pi(X) = N^{-1}$ for all $X \in \Omega$, where $N = \binom{n}{k}$. We will show using Coupling that this chain mixes in $O(k \log(k/\varepsilon))$ time (we will later mention the sort of weak bounds that more complicated conductance/resistance based proofs give even for this very simple example).

Theorem 6.2 *The mixing time of the above Markov chain satisfies $\tau_X(\varepsilon) = O(k \log(k/\varepsilon))$ irrespective of the starting state X.*

Proof: The proof is based on a Coupling that is actually quite simple to set up. The transition $(X_t, Y_t) \to (X_{t+1}, Y_{t+1})$ is defined as follows:

1. If $X_t = Y_t$, then pick X_{t+1} as $\mathfrak{M}$ would and set $Y_{t+1} = X_{t+1}$; else
2. If $r_{X_t} = 0$, set $X_{t+1} = X_t$, and $Y_{t+1} = Y_t$.

3. If $r_{X_t} = 1$, then: Let $S = X_t \setminus Y_t$ and $T = Y_t \setminus X_t$ (note that $|S| = |T|$); fix an arbitrary bijection $g : S \to T$. Pick $i \in X_t$ u.a.r and $j \in [n] \setminus X_t$ u.a.r and set $X_{t+1} = X_t \cup \{j\} \setminus \{i\}$. Define $i' \in Y_t$ and $j' \in [n] \setminus Y_t$ as follows:

   - If $i \in X_t \cap Y_t$, then $i' = i$, else $i' = g(i)$
   - If $j \notin Y_t$, $j' = j$, else (now $j \in T$) $j' = g^{-1}(j)$.

   Now set $Y_{t+1} = Y_t \cup \{j'\} \setminus \{i'\}$.

It is easy to see that (X_t) and (Y_t) are individually just copies of $\mathfrak{M}$, so the above is a legal (in fact Markovian) coupling. We assume $k \geq 2$ to avoid trivialities. Denote by D_t the random variable $X_t \oplus Y_t$. We wish to bound the expectation

$$\mathbf{E}[|D_{t+1}||D_t] \leq (1 - \frac{1}{k})|D_t| \; , \tag{29}$$

as this will imply $\mathbf{E}[|D_t||D_0] \leq (1 - \frac{1}{k})^t |D_0|$. Since $|D_t|$ is a non-negative integer random variable, and $|D_0| \leq 2k$, we obtain

$$\begin{aligned} \mathbf{Pr}[|D_t| > 0 | D_0] &\leq \mathbf{E}[|D_t||D_0] \\ &\leq 2k \cdot (1 - \frac{1}{k})^t \end{aligned}$$

which is at most ε provided $t \geq k \ln(2k\varepsilon^{-1})$. Invoking the Coupling Lemma 6.1, we obtain that the mixing time is $O(k \ln(k/\varepsilon))$, as promised. It remains therefore to establish (29) which basically quantifies the fact that X_t and Y_t tend to "coalesce".

Let $q = |X_t \oplus Y_t|$, and let $q' = |X_{t+1} \oplus Y_{t+1}|$. We want the expectation of q' for a given q. Consider now the choices in Step (3) of the Coupling. Four cases now arise:

(i) $j \in [n] \setminus (X_t \cup Y_t)$ and $i \in X_t \setminus Y_t$: Then $q' = q - 2$.

(ii) $j \in Y_t \setminus X_t$ and $i \in X_t \cap Y_t$: Then $q' = q - 2$.

(iii) $j \in Y_t \setminus X_t$ and $i \in X_t \setminus Y_t$, $j \neq g(i)$: Then $q' = q - 4$.

(iv) In all other cases $q' = q$.

Thus the expected value of the change $q' - q$ is

$$\mathbf{E}[q' - q] = \frac{n - k - q/2}{n - k} \cdot \frac{q/2}{k} \cdot (-2) + \frac{q/2}{n - k} \cdot \frac{k - q/2}{k} \cdot (-2) + \frac{q/2}{n - k} \cdot \frac{q/2 - 1}{k} \cdot (-4)$$

and this gives $\mathbf{E}[q'|q] \leq (1 - \frac{n-2}{k(n-k)})q \leq (1 - \frac{1}{k})q$ (as $k \geq 2$). □ *(Theorem 6.2)*

Comparison with performance of Canonical Paths. The best bound achievable for this problem via the canonical paths/conductance based approach seems to be (see [37]) to bound $\bar{\mathfrak{R}}$ by demonstrating a fractional flow that routes one unit between every pair of unequal states, and this gives $\bar{\mathfrak{R}} \leq \frac{k^2(n-k)^2}{n(n-1)}$. For $k = \Omega(n)$, say $k = n/2$, this gives a bound on mixing time equal to $O(n^2 \log(\binom{n}{k}\varepsilon^{-1})) = O(n^3 + n^2 \log \varepsilon^{-1})$, which is significantly worse than the $O(n(\log n + \log \varepsilon^{-1}))$ bound we proved using Coupling! In fact, in this case ($k = n/2$), the second-largest eigenvalue is known exactly: $\lambda_1 = 1 - 2/n$, so that even getting the best bound on the spectral gap, only yields a mixing time of $O(n^2 + n \log \varepsilon^{-1})$ (using Proposition 2.3)! These crisp and significantly improved bounds seem to be typical of Coupling whenever it works. We will later (in Section 7.2) also see an application of sampling from subsets of $[n]$ of size *at most* k (this is just the "uniform" version of the knapsack problem, where all items have the same size), where Coupling gives a much better bound mixing time than seems possible using techniques of Section 5.

6.3 Known applications of Coupling

Owing to its intuitive appeal, Coupling has been a very popular and successful technique in rapid mixing results. Some instructive examples of Coupling that have appeared in the literature are in sampling proper k-colorings of a graph [18, 5, 40], linear extensions of a partial order [30, 6], points in a convex body [7], independent sets in low-degree graphs [29, 5, 15], general contingency tables on 2 rows [13], etc. Even Broder's original paper [4] on sampling from the set of perfect matchings of a bipartite graph used a complicated Coupling argument, which was later found to have an error [32].

6.4 Path Coupling

Despite the conceptual simplicity and appeal of Coupling, it can often get very difficult to design couplings appropriate to specific situations that arise in sampling problems. The problem is one of "engineering": how do we encourage (X_t) and (Y_t) to coalesce while at the same time meeting the apparently contradicting requirement of keeping the individual processes faithful to $\mathfrak{M}$? This can lead to severe technical complexities (see [30] to get an impression of this). This led Bubley and Dyer [5] to invent an elegant solution to the task of designing Couplings: they called it "Path Coupling". The idea behind Path Coupling is to define the coupling only for "adjacent" states, i.e., only for pairs of states in a carefully chosen subset S of $\Omega \times \Omega$ (and hopefully the task is easier for such pairs of states), and then extend the coupling to arbitrary pairs of states by composition of adjacent couplings along a path. In fact, the discovery of Path Coupling has led to a spurt of Coupling based rapid mixing proofs, and indeed most of the applications cited in Section 6.3 use Path Coupling. We now state and prove the "Path Coupling" lemma (a version taken from [14]):

Lemma 6.3 (Path Coupling Lemma) *Let δ be an integer valued metric defined on $\Omega \times \Omega$ which takes values in $\{0, 1, \ldots, D\}$. Let S be a subset of $\Omega \times \Omega$ such that for all $(X_t, Y_t) \in \Omega \times \Omega$, there exists a path $X_t = Z_0, Z_1, \ldots, Z_r = Y_t$ between X_t and Y_t where $(Z_\ell, Z_{\ell+1}) \in S$ for $0 \leq \ell < r$, and $\sum_{\ell=0}^{r-1} \delta(Z_\ell, Z_{\ell+1}) = \delta(X_t, Y_t)$. (Equivalently, δ is defined by specifying a graph H with vertex set Ω and edge set S, and weights on edges in S, and $\delta(X, Y)$ is simply the shortest path between X and Y in this graph.) Suppose a Coupling $(X, Y) \mapsto (X', Y')$ of the Markov chain $\mathfrak{M}$ is defined on all pairs $(X, Y) \in S$ (note that (X', Y') need not lie in S) such that there exists a $\beta < 1$ such that $\mathbf{E}[\delta(X', Y')] \leq \beta\, \mathbf{E}[\delta(X, Y)]$ for all $(X, Y) \in S$. Then the mixing time $\tau(\varepsilon)$ of $\mathfrak{M}$ satisfies $\tau(\varepsilon) \leq \frac{\ln(D\varepsilon^{-1})}{(1-\beta)}$.*

Remark. One can also bound the mixing time in the case $\beta = 1$ [5, 14]. For the applications we will use to illustrate this technique, we will actually have $\beta < 1$, so to keep things simple we do not discuss the $\beta = 1$ case.

Proof: First, we observe that the Coupling on S can be extended in an obvious way to a Coupling on the entire space $\Omega \times \Omega$. Indeed let $(X_t, Y_t) \in \Omega \times \Omega$. Pick a "path" $X_t = Z_0, Z_1, \ldots, Z_r = Y_t$ such that $\delta(X, Y) = \sum_{\ell=0}^{r-1} \delta(Z_\ell, Z_{\ell+1})$ (use a deterministic choice rule for resolving ties). Define the coupling $(X_t, Y_t) \mapsto (X_{t+1}, Y_{t+1})$ as follows: First select $X_{t+1} = Z'_0 \in \Omega$ according to the probability distribution $P(X, \cdot)$. Now select Z'_1 according to the distribution induced by the pairwise coupling of the adjacent states Z_0 and Z_1, conditioned on the choice of Z'_0; then select Z'_2 using the pairwise coupling of (Z_1, Z_2), and so on, ending with $Z'_r = Y_{t+1}$. It is easy to verify, by induction of the path length r, that Y_{t+1} has been selected according to the distribution $P(Y_t, \cdot)$, so $(X_t, Y_t) \mapsto (X_{t+1}, Y_{t+1})$ does define a "legal" coupling that obeys conditions of Definition 6.1. Now

$$\mathbf{E}[\delta(X_{t+1}, Y_{t+1})] \quad \leq \quad \mathbf{E}[\sum_{\ell=0}^{r-1} \delta(Z'_\ell, Z'_{\ell+1})]$$

$$
\begin{aligned}
&= \sum_{\ell=0}^{r-1} \mathbf{E}[\delta(Z'_\ell, Z'_{\ell+1})] \\
&\leq \beta \sum_{\ell=0}^{r-1} \delta(Z_\ell, Z_{\ell+1}) \\
&= \beta\delta(X_t, Y_t),
\end{aligned}
$$

where we have used the fact that δ is a metric, and linearity of expectation. Now as in the proof of Theorem 6.2, this gives $\mathbf{E}[\delta(X_t, Y_t)] \leq \beta^t D$, and thus $\mathbf{Pr}[X_t \neq Y_t] \leq \mathbf{E}[\delta(X_t, Y_t)] \leq \varepsilon$ whenever $t \geq \frac{\ln(D\varepsilon^{-1})}{(1-\beta)}$. Invoking the Coupling Lemma 6.1, the claimed bound on the mixing time $\tau(\varepsilon)$ follows. □

Remark. The notion of "adjacency" in the graph H defined for Path Coupling need not have anything to do with the transitions in the Markov chain that is being studied. In fact, two states that are adjacent in the Path Coupling graph H need not even be reachable from one another in the Markov chain.

7 Some applications of Path coupling

In this section, we present a few applications of path coupling to Markov chains for interesting sampling problems.

7.1 Sampling k-colorings of a graph

Given a graph $G = (V, E)$ with maximum degree Δ, consider the task of sampling uniformly at random from the set $\Omega_k(G)$ of (proper) k-colorings of G. Let $C = \{1, 2, \ldots, k\}$ be the set of colors.

A natural Markov chain for the above problem, known in the literature as "Glauber dynamics", is the following. Suppose the current state is a coloring X:

- Choose $v \in V$ u.a.r, and $c \in C$ u.a.r. If $X_{v\to c}$ (i.e., X with the color of v changed to c) is a proper coloring of F, then move to $X_{v\to c}$, else remain at X.

Jerrum [18] (see also [19]) first proved, using Coupling, that the above chain rapidly mixes for $k > 2\Delta$. We will now present a simple proof due to Bubley and Dyer [5] of this fact using Path Coupling. We remark that Vigoda [40] recently established that this chain mixes rapidly for $k > \frac{11}{6}\Delta$, using Path Coupling on a different chain, and then using that to deduce the mixing time of the Glauber dynamics.

Following [5], we present the result in a more general set-up that captures "coloring-type" problems, and then deduce the result for coloring from that.

The general set-up is the following. Let V and C be finite sets, and let $n = |V|$ and $k = |C|$, and we consider a finite Markov chain $\mathfrak{M}$ with state space $\Omega \subseteq C^V$, the set of functions from V to C, and unique stationary distribution π. The transition structure of $\mathfrak{M}$ is similar to the graph coloring case we considered above: From a current state $X \in \Omega$, pick $v \in V$ according to a fixed distribution J on V, and and $c \in C$ according to a distribution $\kappa_{X,v}$ that depends only on X and v, and make the transition to $X_{v\to c}$ (where $X_{v\to c}(w)$ equals c if $w = v$, and equals $X(w)$ otherwise). We assume that $\kappa_{X,v}(c) = 0$ whenever $X_{v\to c} \notin \Omega$. Path Coupling yields the following result for this class of problems (for distributions A and B, $\|A - B\|$ denotes their statistical difference or variation distance):

Theorem 7.1 ([5]) *Let $\Omega = C^V$, and let*

$$\beta = \max_{X,Y\in\Omega, i\in V} \left\{ 1 - J(i) + \sum_{j\in V} J(j)\|\kappa_{X,j} - \kappa_{Y,j}\| \;\middle|\; Y = X_{i\to c} \textit{ for some } c \in C, \textit{ and } Y \neq X \right\}.$$

Then, if $\beta < 1$, the mixing time of $\mathfrak{M}$ satisfies $\tau(\varepsilon) \le \ln(n\varepsilon^{-1})/(1-\beta)$.

Proof: We set up a Path Coupling with "adjacency graph" being all non-equal pairs (X, Y) such that $Y = X_{i\to c}$ for some i, c, and the metric δ used is the Hamming metric (so $\delta(X, Y) = 1$ for adjacent pairs). For such a pair (X, Y) define the coupling to (X', Y') as follows: X' is distributed according to $P(X, \cdot)$, namely: pick $v \in V$ according to J and $c_0 \in C$ according to $\kappa_{X,v}$, and set $X' = X_{v\to c_0}$. Next pick $c_1 \in C$ as follows: with probability $\min\{1, \kappa_{Y,v}(c_0)/\kappa_{X,v}(c_0)\}$ let $c_1 = c_0$, otherwise pick c_1 according to the distribution $\gamma(c) = \frac{\max\{0, \kappa_{Y,v}(c) - \kappa_{X,v}(c)\}}{\|\kappa_{Y,v} - \kappa_{X,v}\|}$.

It is easy to see that marginally we choose c_1 according to $\kappa_{Y,v}$, so the above defines a "legal" coupling for the chain $\mathfrak{M}$. It is also easy to verify that $\mathbf{Pr}[c_1 \ne c_0] = \|\kappa_{Y,v} - \kappa_{X,v}\|$. Now since $\delta(X, Y)$ changes by at most 1 in one step of the chain, we have

$$\begin{aligned}
\mathbf{E}[\delta(X', Y')] &= 1 - \mathbf{Pr}[\delta(X', Y') = 0] + \mathbf{Pr}[\delta(X', Y') = 2] \\
&= 1 - J(i)\,\mathbf{Pr}[c_0 = c_1 | v = i] + \sum_{j\ne i} J(j)\,\mathbf{Pr}[c_0 \ne c_1 | v = j] \\
&= 1 - J(i)(1 - \|\kappa_{Y,i} - \kappa_{X,i}\|) + \sum_{j\ne i} J(j)\|\kappa_{Y,j} - \kappa_{X,j}\| \\
&\le \beta\delta(X, Y)
\end{aligned}$$

(since $\delta(X, Y) = 1$). The result now follows from the Path Coupling Lemma 6.3. □

Application to Coloring. Consider the Markov chain with state space all (not necessarily proper) k-colorings of G and transitions at state X defined as follows.

1. Choose v at random from V according to distribution J and c u.a.r from C.
2. If v is properly colored in $X_{v\to c}$, then move to $X' = X_{v\to c}$ else remain at X.

This is an extension of the Glauber dynamics we discussed earlier (except that we allow more general distributions to select v from), to all of C^V (we do so in order to be able to apply Theorem 7.1). This does not cause any problems since the non-proper colorings are transient states, and the stationary distribution is uniform over all proper k-colorings of G, and zero elsewhere. Moreover, if we start from a proper k-coloring, then we visit only states that correspond to proper k-colorings, so the mixing time of this chain is an upper bound on the mixing time of the Glauber dynamics. Note that this chain is not reversible, but Theorem 7.1 applies for such chains as well.

Let us now apply Theorem 7.1. Let $d(v)$ denote the degree of vertex v, and let m be the number of edges in G. We will use J to be proportional to the degree of the vertex, so that $J(v) = d(v)/2m$. If colorings X and Y differ only on vertex i, then $\kappa_{Y,j} = \kappa_{X,j}$ unless $j = i$ or $j \sim i$ (here $j \sim i$ stands for adjacency in the graph G). When $j = i$, $\kappa_{X,i}(X(i)) = \frac{d(i)+1}{k}$ and $\kappa_{Y,i}(X(i)) = \frac{1}{k}$, and similarly for the color $Y(i)$, while $\kappa_{X,i}(c) = \kappa_{Y,i}(c)$ for all colors $c \ne X(i), Y(i)$. Hence $\|\kappa_{Y,i} - \kappa_{X,i}\| = d(i)/k$. When $j \sim i$, every color that would be accepted in X (resp. Y), except possibly $Y(i)$ (resp. $X(i)$) would be accepted in Y (resp. X) as well, and hence $\|\kappa_{Y,j} - \kappa_{X,j}\| = \frac{1}{k}$. Thus the parameter β (from Theorem 7.1) satisfies

$$\beta \le 1 - \frac{d(i)}{2m}(1 - \frac{d(i)}{k}) + \sum_{j\sim i} \frac{d(j)}{2mk} .$$

Hence $\beta < 1$ whenever

$$k > \max_{v\in V}\{d(v) + \sum_{w\sim v} \frac{d(w)}{d(v)}\} .$$

This condition is certainly satisfied when $k > 2\Delta$, so using Theorem 7.1 we conclude

Theorem 7.2 ([5]) *The Glauber dynamics for sampling proper k-colorings of a graph G with maximum degree Δ is rapidly mixing (with mixing time $O(kn\log(n\varepsilon^{-1}))$) whenever $k > 2\Delta$.*

7.2 Sampling "Uniform Knapsack" solutions

We consider another elegant application of Path Coupling. We are interested in sampling from the space Ω of subsets of $[n] = \{1, 2, \ldots, n\}$ of size *at most* k. This resembles the problem of sampling k-element subsets of $[n]$ that we considered in Section 6.2, but turns out to be trickier. Note also that this problem is a special case of the 0-1 knapsack problem (which we considered in Section 5) when all items to be packed have the same size.

The Markov chain $\mathfrak{M}_K$ we will study will be the same as the one in Section 5, namely from a state $X \subseteq [n]$, $|X| \le k$, pick $r_X \in \{0,1\}$ u.a.r. If $r_X = 0$ remain at X. If $r_X = 1$, pick an $i \in [n]$ u.a.r and move to $X \setminus \{i\}$ if $i \in X$ and to $X \cup \{i\}$ if $i \notin X$ and $|X| < k$. We will use Path Coupling to prove

Theorem 7.3 *The mixing time of the Markov chain $\mathfrak{M}_K$ satisfies $\tau(\varepsilon) = O(n\log(k\varepsilon^{-1}))$.*

Proof: We will use Path Coupling with the (somewhat unusual) metric $\delta(X,Y) = |X \oplus Y| + ||X| - |Y||$. Note that $\delta(X,Y) \ge 2$ whenever $X \ne Y$. The set of "adjacent" pairs $S \subseteq \Omega \times \Omega$ for which we will define the Coupling is: $S = \{(X,Y) : X, Y \in \Omega \wedge \delta(X,Y) = 2\}$. It is easy to see that the metric δ and the set S satisfy the conditions required by the Path Coupling Lemma 6.3.

Now consider $(X,Y) \in S$ with $\delta(X,Y) = 2$; we wish to define a Coupling $(X,Y) \mapsto (X',Y')$. There are two possibilities for (X,Y):

(i) One of X,Y is a subset of the other, say $Y \subset X$ (the other case is symmetric), $|Y| = |X| - 1$.

(ii) $|X| = |Y|$ and $|X \oplus Y| = 2$.

We consider each of these cases in turn.

Case (i): Let $Y = X \setminus \{p\}$ for some $p \in [n]$. Now the Coupling (X',Y') is defined as follows:

(1) Pick $r_X \in \{0,1\}$ and $i \in [n]$ u.a.r. If $i = p$ then set $r_Y = 1 - r_X$; otherwise set $r_Y = r_X$.

(2) If $r_X = 0$ set $X' = X$. Else if $i \in X$ set $X' = X \setminus \{i\}$, else set $X' = X \cup \{i\}$ if $|X| < k$ and $X' = X$ otherwise.

(3) If $r_Y = 0$ set $Y' = Y$. Else if $i \in Y$ set $Y' = Y \setminus \{i\}$, else set $Y' = Y \cup \{i\}$ if $|Y| < k$ and $Y' = Y$ otherwise.

It is easy to see that $\delta(X',Y') = 2$ except when $i = p$, in which case, since we have cleverly designed the Coupling by setting $r_Y = 1 - r_X$ so that only one of X,Y "fires", $\delta(X',Y') = 0$. Thus we have $\mathbf{E}[\delta(X',Y')] = (1 - \frac{1}{n})\delta(X,Y)$.

Case (ii): $|X| = |Y|$ and $|X \oplus Y| = 2$. Let $X = S \cup \{p\}$ and $Y = S \cup \{q\}$ for some $p \ne q$. The Coupling (X',Y') is defined as follows:

(1) Pick $r_X \in \{0,1\}$ and $i \in [n]$ u.a.r. Set $r_Y = r_X$. If $i \notin \{p,q\}$, set $j = i$. If $i = p$ (resp. q) set $j = q$ (resp. p).

(2) If $r_X = 0$ set $X' = X$. Else if $i \in X$ set $X' = X \setminus \{i\}$, else set $X' = X \cup \{i\}$ if $|X| < k$ and $X' = X$ otherwise.

(3) If $r_Y = 0$ set $Y' = Y$. Else if $j \in Y$ set $Y' = Y \setminus \{j\}$, else set $Y' = Y \cup \{j\}$ if $|Y| < k$ and $Y' = Y$ otherwise.

Once again, the Coupling has been constructed so that $\delta(X', Y') = 2$ whenever $i \notin \{p, q\}$; $\delta(X', Y') = 0$ if $i = p$ and $r_X = 1$, and $\delta(X', Y') \le 2$ in all cases. Thus we have $\mathbf{E}[\delta(X', Y')] \le (1 - \frac{1}{2n})\delta(X, Y)$.

Combining both the above cases we get $\mathbf{E}[\delta(X', Y')] \le (1 - \frac{1}{2n})\delta(X, Y)$ always. Also, the maximum value D of $\delta(X_0, Y_0)$ over all pairs $(X_0, Y_0) \in \Omega \times \Omega$ is clearly $2k$. By Theorem 7.1 therefore, we have shown that $\mathfrak{M}_K$ has mixing time $\tau(\varepsilon) = O(n \log(k\varepsilon^{-1}))$, completing the proof. $\square$

Comparison with Canonical paths. Even for this special case of 0-1 knapsack, the best bound that we get using the multicommodity flow based analysis of Section 5 (without any change) is only $O(n^6)$, and it is almost inconceivable that such an approach can hope to yield a bound better than $O(n^3)$. Coupling gave us a much better $O(n \log(k\varepsilon^{-1}))$ bound, and the proof was in fact much easier than using canonical paths!

Remark. The uniformity of weights seems critical to our argument above. The "asymmetry" created when items have widely varying sizes seems to make it difficult for any natural Coupling strategy to work.

7.3 Linear extensions of a partial order

We are given a partially ordered set $(P, \preceq)$ where $|P| = n$, and we want to sample u.a.r from the space Ω of all linear orders that extend $\preceq$. (A linear order extending $\preceq$ is a permutation $a_1, a_2, \ldots, a_n$ of the elements of P such that $a_i \preceq a_j$ implies $i \le j$.)

A natural Markov chain with uniform stationary distribution over Ω was shown to be rapid mixing by Karzanov and Khachiyan via conductance arguments that exploited the geometry of the space [26]. Dyer and Frieze [9] improved the conductance estimate, and hence the bound on the mixing time, of this chain, and this gave a mixing time of $O(n^5 \log n + n^4 \log \varepsilon^{-1})$.

In this section, we will sketch a chain $\mathfrak{M}_{\mathrm{le}}^J$, which is a slight variant of the chain discussed above, and show (using Path Coupling) that it has a mixing time of $O(n^3 \log(n\varepsilon^{-1}))$, which significantly improves the best "conductance based" bound for this problem. The chain and its analysis are due to Bubley and Dyer [6] (see also [19] for an exposition).

Actually this algorithm can be used to sample u.a.r from any set Ω of permutations of elements of P that satisfies the following "closure" property: If $\sigma = (a_1, a_2, \ldots, a_n) \in \Omega$ and $\sigma \circ (i, j) = (a_1, \ldots, a_{i-1}, a_j, a_{i+1}, \ldots, a_{j-1}, a_i, a_{j+1}, \ldots, a_n) \in \Omega$ (i.e., the positions of a_i and a_j can be swapped and the resulting permutation still lies in Ω), then all permutations which are obtained from σ by placing a_i and a_j at arbitrary positions in the interval $[i, j]$, also lie in Ω. Clearly the linear extensions of a partial order have this closure property.

The transitions from one linear extension to another in the chain are obtained by (pre)-composing with a random transposition $(p, p+1)$ (if this yields a valid linear order); however, instead of selecting $p \in [n-1]$ uniformly, p is chosen according to a distribution J on $[n-1]$ that gives greater weight to values near the center of the range. Formally, the chain $\mathfrak{M}_{\mathrm{le}}^J$ is defined as follows. Let the current state be X_t. Then the next state X_{t+1} is defined by the following random experiment:

(1) Pick $p \in [n-1]$ according to the distribution J, and $r \in \{0, 1\}$ u.a.r

(2) If $r = 1$ and $X_t \circ (p, p+1) \in \Omega$, then $X_{t+1} = X_t \circ (p, p+1)$; otherwise $X_{t+1} = X_t$.

To use Path Coupling we need to specify an "adjacency" structure for the state space Ω. We say two states g and g' are adjacent if $g' = g \circ (i, j)$ for some transposition (i, j) with $1 \le i < j \le n$, and the "distance" $\delta(g, g')$ in this case is defined to be $j - i$. Since this distance is symmetric (i.e., $\delta(g, g') = \delta(g', g)$), this adjacency structure yields a weighted, undirected graph H on vertex set Ω. One can verify that the shortest path between adjacent states g, g' in H is the direct one that uses the edge (g, g'). We may thus extend δ to a metric on Ω by defining $\delta(g, h)$ for arbitrary states $g, h \in \Omega$ to be the length of a shortest path from g to h in H, and all conditions of the Path Coupling Lemma 6.3 are now met. It remains to define a coupling $(g, h) \mapsto (g', h')$ for adjacent states g, h and then bound $\mathbf{E}[\delta(g', h')]$.

The Coupling is defined as follows. Let (g, h) be a pair of adjacent states in H and let $h = g \circ (i, j)$. Then the transition to (g', h') is defined by the following experiment:

(i) Pick $p \in [n-1]$ according to distribution J, and $r_g \in \{0, 1\}$ u.a.r. If $j - i = 1$ and $p = i$, set $r_h = 1 - r_g$; otherwise set $r_h = r_g$.

(ii) If $r_g = 1$ and $g \circ (p, p+1) \in \Omega$ then set $g' = g \circ (p, p+1)$ else set $g' = g$.

(iii) If $r_h = 1$ and $h \circ (p, p+1) \in \Omega$ then set $h' = h \circ (p, p+1)$ else set $h' = h$.

Lemma 7.4 *For adjacent states g and h, for a suitable choice of the probability distribution J, we have*

$$\mathbf{E}\ [\delta(g', h') \mid g, h] \leq \Big(1 - \frac{6}{n^3 - n}\Big)\delta(g, h)\ .$$

In light of Lemma 6.3, this implies that the mixing time of $\mathfrak{M}_{\mathrm{le}}^J$ is $O(n^3 \log(n\varepsilon^{-1}))$ (since the "diameter" D of the graph H is easily seen to be at most $\binom{n}{2}$). It thus only remains to prove Lemma 7.4.

Proof of Lemma 7.4: We only provide the skeleton of the proof; details can be found in [6]. When $h = g \circ (i, j)$, it is easy to see that when $p \notin \{i-1, i, j-1, j\}$, we will have $h' = g' \circ (i, j)$ and thus $\delta(g', h') = \delta(g, h) = j - i$. When $p = i - 1$ or $p = j$, it is again easily checked that $\mathbf{E}\ [\delta(g', h') \mid g, h, p = i - 1 \vee p = j] \leq \delta(g, h) + 1/2$.

The "interesting case" is when $p = i$ or $p = j - 1$. These are symmetric, so let us focus on the case $p = i$. There are two sub-cases: $j - i = 1$ and $j - i \geq 2$. First, consider the case $j - i = 1$. In this case, we have made sure, by setting $r_h = 1 - r_g$, that only one of g or h "fires" in the Coupling, and thus $g' = h'$ and therefore $\delta(g', h') = 0$! In the case $j - i \geq 2$, by the "closure" property of Ω discussed earlier (this is the only place where we use this closure property), we know both $g \circ (i, i+1), h \circ (i, i+1) \in \Omega$, thus either $r_X = r_Y = 0$ and then $\delta(g', h') = \delta(g, h)$, or $r_X = r_Y = 1$ and $\delta(g', h') = j - i - 1 = \delta(g, h) - 1$. Hence $\delta(g', h')$ is less than $\delta(g, h)$ in expectation.

Summing up, it follows from the above discussion that

$$\mathbf{E}\ [\delta(g', h')\ | g, h] \leq \delta(g, h) - \frac{-J(i-1) + J(i) + J(j-1) - J(j)}{2}\ . \tag{30}$$

Specializing the probability distribution $J(\cdot)$ to be $J(p) \stackrel{\text{def}}{=} \zeta(p+1)(n-p-1)$ where $\zeta = 6/(n^3 - n)$ is a normalizing constant, and using $\delta(g, h) = j - i$, we get from (30) that $\mathbf{E}\ [\delta(g', h')] \leq (1 - \zeta)\delta(g, h)$. □ *(Lemma 7.4)*

8 Coupling is weaker than Conductance

We have seen several Coupling based proofs in the last Section which are not only extremely simple and elegant, but also end up giving much better bounds on mixing time than known via conductance based arguments. So, is Coupling the panacea as far as bounding mixing times goes? In particular, is Coupling as powerful as conductance, and does it *capture* rapid mixing exactly?

This fundamental question was unanswered for a long time until recently when Kumar and Ramesh [27] proved the following important result: For the famous Jerrum-Sinclair chain for sampling perfect and near-perfect matchings, no Coupling argument can show rapid mixing (the chain is known to be rapidly mixing using a canonical paths argument [20]). Hence Coupling is actually "weaker" than conductance! We discuss the salient features behind their proof in this section.

The Jerrum-Sinclair Chain. We are given a bipartite graph $G = (V_1, V_2, E)$ with $|V_1| = |V_2| = n$ and the goal is to sample u.a.r from the set $\mathfrak{P}$ of perfect and near-perfect matchings of G (a near-perfect matching is a matching that saturates all but two vertices of G). Jerrum and Sinclair [20] proposed the following natural Markov chain $\mathfrak{M}_{\mathrm{JS}}$ for sampling from $\mathfrak{P}$: At each state M, the chain moves to a state M' defined by the following random experiment:

(i) Pick $r \in \{0, 1\}$ u.a.r and an edge $e \in E$ u.a.r.

(ii) If $r = 0$ set $M' = M$; Else

(iii) If M is a perfect matching: Then set $M' = M \setminus \{e\}$ if $e \in M$, or else $M' = M$.

(iv) Suppose M is a near-perfect matching. Let $e = (u, v)$. There are two cases:

  (a) If u, v are both unmatched in M, set $M' = M \cup \{e\}$. [Add Move]

  (b) If exactly one of u, v is unmatched, then set $M' = M \setminus \{e'\} \cup \{e\}$ where e' is the edge in M incident on whichever of u, v is matched. [Swap Move]

(v) If none of the above conditions are met, set $M' = M$.

A special graph G. Anil Kumar and Ramesh [27] show that for a certain graph G, every Coupling strategy on the above chain will require time exponential in n. This graph has some special properties which are used in the proof; these are:

1. G has $\Omega(\frac{n!}{c^n})$ perfect matchings for some constant $c > 1$.

2. Each vertex of G has degree at least αn, for some $\alpha < 1/2$.

3. For every pair of vertices, the intersection of their neighborhoods has size at most $\alpha n/2$.

Such a graph G can be shown to exist using the probabilistic method (see for example the final version of [27]).

Modeling the Coupling Process. The coupling process $\mathcal{C} = (\mathcal{X}, \mathcal{Y})$ is specified by transition probabilities $p_{\mathcal{C}}(v, w)$ where $v = (a, b) \in \mathfrak{P} \times \mathfrak{P}$, and $w = (c, d) \in \mathfrak{P} \times \mathfrak{P}$ are pairs of states in $\mathfrak{P}$. Note that $p_{\mathcal{C}}(v, w)$ could even be a function of the history, i.e., the transition probabilities could vary with time (we do not show the time dependence for notational convenience, but it should be treated as implicit). Since we are aiming for a negative result and wish to rule out the existence of *any* Coupling based proof, the only thing we will (and can) assume about these probabilities is that the processes $\mathcal{X}$ and $\mathcal{Y}$ must individually be faithful copies of $\mathfrak{M}_{\mathrm{JS}}$, or in other words: If $v = (x, y)$, then for each $x' \in \mathfrak{P}$ and for each time instant t, $\sum_{w \in T(x')} p_{\mathcal{C}}(v, w) = P(x, x')$ where $T(x') = \{(x', z) \mid z \in \mathfrak{P}\}$ and $P(\cdot, \cdot)$ is the transition probabilities of the chain $\mathfrak{M}_{\mathrm{JS}}$, and a similar equation for $P(y, y')$ for each $y' \in \mathfrak{P}$.

Idea behind the Proof. The basic structure of the proof is the following: Define a "distance" between the two states X, Y in a Coupling, relative to which the states will have a tendency to *drift away* from each other in *any* Coupling, i.e., most transitions of any Coupling are *distance increasing*. Then analyze this drifting behavior and show that staring with two states (X_0, Y_0) at a distance $\Theta(n)$ apart, any Coupling will require exponential number of steps t before the states X_t, Y_t become equal, with say a probability of $1/2$. This gives an exponential lower bound on the Coupling time for any strategy, as desired.

8.1 Details of the Analysis

We partition the states of the Coupling chain $\mathcal{C}$ into layers $L(i)$, $i = 0, \ldots, 2n$ according to the "distance" i between its elements, where $L(i)$ contains of all pairs $(M, N) \in \mathfrak{P} \times \mathfrak{P}$ such that $|M \oplus N| = i$. We further partition each set $L(i)$ into two sets $Bot(i)$ and $Top(i)$, where $Bot(i) = \{(M, N)| \exists$ vertex v which is unmatched in exactly one of $M, N\}$, and $Top(i) = \{(M, N)|$ either both M and N are perfect matchings or both are near-perfect matchings with the same unmatched vertices$\}$.

A move in $\mathcal{C}$ from $L(i)$ to $L(j)$ is *leftwards* or *distance reducing* if $j < i$, and *rightwards* or *distance increasing* if $j > i$. Since G has $\Omega(\frac{n!}{c^n})$ perfect matchings, with overwhelming probability, the start state of the Coupling lies in $L(i)$ for some $i \geq n/4$. For simplicity therefore, we assume that the Coupling $\mathcal{C}$ begins at some state in $L(i_0)$, $i_0 \geq n/4$.

The idea now is to upper bound the probabilities of the leftward transitions and lower bound the probabilities of the rightward transitions, and then use these bounds to show that the Coupling has a tendency to drift towards the right. Finally, this will imply that the (expected) number of steps to reach a state in $L(0)$ will be exponentially large, giving us our desired result.

The Key Lemmas. We now state the main Lemmas which bound transition probabilities between different layers. We will later use the statements of these Lemmas give us the desired "rightward drift". We give a representative proof of one of the Lemmas (the proofs of the other Lemmas can be found in [27], and we do not reproduce them here).

Lemma 8.1 *No transition in $\mathcal{C}$ can change the distance by more than* 4.

Lemma 8.2 *For any coupling strategy, the sum of transition probabilities from $(M, N) \in Bot(i)$ to vertices in $L(j)$, $j < i$, is at most $\frac{2i+1}{m}$.*

Lemma 8.3 *For any coupling strategy, the sum of the transition probabilities from $(M, N) \in Bot(i)$ into $\bigcup_{j=i+1}^{i+4} Bot(j)$ is at least $\frac{\alpha n/2-i-2}{2m}$.*

Lemma 8.4 *For any coupling strategy, the sum of the transition probabilities from $(M, N) \in Bot(i)$ into $Top(i) \cup Top(i+1)$ is at most $\frac{i+3}{2m}$.*

Lemma 8.5 *For any coupling strategy, all transitions from $(M, N) \in Top(i)$ are to vertices in either $Top(i)$ or in $Bot(j)$ for some $j \geq i - 2$.*

We only prove Lemma 8.3 as it is the key Lemma that establishes a tendency of any Coupling to drift to the right. This should give a flavor of the sort of arguments necessary to prove the other Lemmas as well.

Proof of Lemma 8.3: Since $(M, N) \in Bot(i)$, three cases arise: (a) M is a near-perfect matching and N is a perfect matching; (b) M is a perfect matching and N is a near-perfect matching; and (c) Both M and N are near-perfect matchings with at most one common unmatched vertex. Case (b) is symmetric to Case (a), so we consider Cases (a) and (c) in turn.

Case (a): M is near-perfect and N is perfect. Let $a \in V_1$ and $b \in V_2$ be the unmatched vertices in M. We consider only one situation that will increase $|M \oplus N|$ and then lower bound the probability that this situation occurs. The situation is: M moves to $M' = M + e - (u, u')$ where $e = (a, u)$ and $(u, u') \in M \cap N$. Now $|M' \oplus N| = |M \oplus N| + 2$. N can move to N' where either $N' = N$ or $N' = N - f$ for some edge $f \in N$. In either case $|M' \oplus N'| \geq |M \oplus N| + 1$. Furthermore, u' and b are unmatched in M', and since $(u', b) \notin N$, at least one of them is matched in N'. We thus conclude $(M', N') \in Bot(j)$, for some $j > i$. Now the probability that this situation occurs is clearly at least $\frac{\alpha n - |M \setminus N|}{2m}$ which is at least $\frac{\alpha n - i}{2m}$, for any coupling strategy.

Case (b): M and N are both near-perfect. Suppose M have vertices $a \in V_1$ and $b \in V_2$ unmatched and N has vertices $c \in V_1$ and $d \in V_2$ unmatched. Let us assume that $b \neq d$ (while c could equal a).

We once again focus on a particular class of moves which M makes. Suppose M chooses an edge $e = (b, u)$, where u is *not adjacent* to d and $(u, u') \in M \cap N$ for some $u' \in V_2$ (by our assumption about G there exist at least $\alpha n/2 - |M \setminus N| \geq \alpha n/2 - i$ such edges e. If e is picked (i.e., $M' = M + e - (u, u')$) then $|M' \oplus N| = |M \oplus N| + 2$. It is easy to verify now that the only moves for N that can reduce the distance back by 2 are when it choose the unique edge $(c, c') \in M$, if any, or the unique edge $(d, d') \in M$, to swap in. The probability of either of these happening is at most $\frac{2}{2m}$ for any coupling strategy. Furthermore, in this case $u' \in V_2$ is unmatched in M' and must be matched in N' because $(u, u') \in N$ (it lies in $M \cap N$) and $(u, d) \notin E$ by the choice of u. Hence (M', N') lies in $Bot()$. Summing up, $(M', N') \in Bot(j)$ for $j > i$ with probability at least $\frac{\alpha n/2 - i - 2}{2m}$. □ *(Lemma 8.3)*

8.2 Bounding the Coupling Time

With the above Lemmas in place, we are ready to finish off the analysis bounding the coupling time. The rightward drifting behavior of any Coupling $\mathcal{C}$ can be predicted (qualitatively) given the above Lemmas. We now see how to quantify this intuition. We define a sequence of random variables $Z_0, Z_1, \ldots$ which represent the layer number of some intermediate states of the Coupling. We will show that $\mathbf{Pr}[Z_t = 0] \sim te^{-\Theta(n)}$, and this will imply an exponential lower bound on the Coupling time.

Define Z_0 to the layer number of the starting state of the Coupling $\mathcal{C}$. As discussed earlier, we assume $Z_0 \geq n/4$. Also assume, by virtue of Lemma 8.5, that the starting state is in a $Bot()$ set rather than a $Top()$ set.

For $i > 0$, the random variable Z_i is defined as follows. If $Z_{i-1} = 0$ then $Z_i = 0$. Otherwise, Z_i is the layer number of the first state A reached in the Coupling $\mathcal{C}$ that has the following properties:

1. $A \notin L(Z_{i-1})$.
2. A is in some $Bot()$ set or in $L(0)$.

Lemma 8.6 *For every $i \geq 1$, $|Z_i - Z_{i-1}| \leq 8$*

Proof: Follows easily from Lemmas 8.1 and 8.5. □

The Lemma below quantifies the "rightward drifting" behavior of the sequence $Z_0, Z_1, \ldots$.

Lemma 8.7 *Define $p_i = \frac{\alpha n/2 - Z_{i-1} - 2}{2m}$ and $q_i = \frac{5(Z_{i-1}+1)}{2m}$. Then $\mathbf{Pr}[Z_i > Z_{i-1} | Z_{i-1}] \geq \frac{p_i}{p_i + q_i}$.*

Proof: By Lemma 8.3, $Z_i > Z_{i-1}$ happens with probability at least p_i. By Lemma 8.5, $Z_i < Z_{i-1}$ only if the first vertex visited after leaving $Bot(Z_{i-1})$ for the last time is either in $L(j)$, $j < i$, or is in $Top(Z_{i-1}) \cup Top(Z_{i-1} + 1)$. By Lemmas 8.2 and 8.4, this probability is at most q_i. The claimed result now follows. □

Let $\beta > 0$ be a constant such that $\frac{5(\beta n+1)}{\alpha n/2 - \beta n - 2} \leq \frac{1}{16}$. Then it is easy to see using the above Lemma that

$$\mathop{\mathbf{E}}_{Z_i}[Z_i - Z_{i-1} \mid Z_{i-1}; 0 < Z_{i-1} \leq \beta n] \geq \frac{1}{4}. \tag{31}$$

Combining Lemma 8.6 with the above Equation, we will be able to bound the Coupling time by appealing to the following submartingale inequality [27] (see also [16]).

Propostion 8.8 *Let $Z_0, Z_1, Z_2, \cdots$ be a sequence of random variables with the following properties (for some $R, \Delta, M > 0$):*

1. $Z_i \geq 0$, *for all* $i \geq 0$. *Further* $Z_i = 0 \Rightarrow Z_{i+1} = 0$, *for all* $i \geq 0$.

2. $|Z_i - Z_{i-1}| \leq \Delta$ *for all* i, $i \geq 1$.

3. $\mathbf{E}\ [Z_i - Z_{i-1} \mid Z_{i-1}; 0 < Z_{i-1} \leq R] \geq M$, *for all* i, $i \geq 1$.

Let T *be the random variable defined as* $\min\{i \geq 0 | Z_i = 0\}$. *Then*

$$\mathbf{Pr}[T \leq t | Z_0] \leq e^{-\frac{MZ_0}{\Delta^2}} + te^{-\frac{M(R-\Delta)}{\Delta^2}} .$$

Note that the above is very similar in spirit to Azuma's inequality applied to submartingales, except that the assumption (3) above is made only when conditioned on $0 < Z_{i-1} \leq R$, and not for any value of Z_{i-1} (as is done in Azuma's inequality).

Let us now apply the above Proposition to our setting. Let t_ε be the earliest instant at which the probability that coupling time exceeds t_ε falls below ε. Define $T = \min\{i \geq 0 | Z_i = 0\}$. Then, applying Proposition 8.8 with $\Delta = 8$, $M = 1/4$, $R = \beta n$ and $Z_0 \geq n/4$, we get

$$1 - \varepsilon \leq \mathbf{Pr}[T \leq t_\varepsilon | Z_0] \leq t_\varepsilon e^{-\Theta(n)} .$$

It follows that $t_\varepsilon \geq (1-\varepsilon)\exp(\Theta(n))$. We have thus proved the following:

Theorem 8.9 ([27]) *Consider any Coupling process for the Markov chain* $\mathfrak{M}_{\mathrm{JS}}$ *for sampling from perfect and near-perfect matchings. The probability that this process has "coupled" exceeds* $(1-\varepsilon)$ *only after time* $\Omega((1-\varepsilon)e^{\Theta(n)})$. *Thus, no proof of rapid mixing of* $\mathfrak{M}_{\mathrm{JS}}$ *exists based on the Coupling Lemma.*

9 Concluding Remarks and Open Questions

We have seen that the mixing rate of a Markov chain is captured by the spectral gap and also by a geometric parameter called Conductance. We discussed ways to bound the conductance, and also ways to bound the spectral gap directly, based on construction of canonical paths or flows between every pair of states that do not overload any transition of the Markov chain. The "flow" based approach led to the notion of resistance which also *captures* the spectral gap (up to square factors). We showed that for a large class of chains, the existence of "good" canonical paths with low edge-congestion also captures mixing time, and thus is no weaker than the resistance based approach. We nevertheless demonstrated that spreading the flow along multiple paths might still be a very useful design tool by discussing the recent result of [34] on the rapid mixing of a natural chain for sampling 0-1 knapsack solutions.

We then turned to an entirely different approach to proving rapid mixing: Coupling. We discussed "Path Coupling" which is a useful tool in designing good Couplings. We saw several simple and elegant applications of Coupling which invariably gave much better bounds on mixing time than known through conductance. One of these examples was the 0-1 knapsack problem with uniform item sizes for which we proved a much better mixing time bound than seems possible using the (more difficult) approach of [34].

Despite the appeal of Coupling in several applications, it turns out that Coupling is weaker than conductance in the sense that there are Markov chains with an exponential gap between their actual mixing time and that which can be deduced using any Coupling strategy. We discussed the result of [27] which showed such a result for the famous Jerrum-Sinclair chain for sampling uniformly from the set of perfect and near-perfect matchings of a bipartite graph.

There are several natural questions on the relative power of the various techniques that are worthy of more detailed study. We list some of them below.

- The result of Kumar and Ramesh [27] is quite natural and says that Coupling cannot work when there is a measure of distance relative to which the states have a tendency to drift away from each other in any Coupling strategy. It will be nice to find other chains for which Coupling cannot prove rapid mixing. This might shed some light on how to tackle the question we raise next.
- Is there a subclass of Markov chains for which Coupling characterizes rapid mixing (up to polynomial factors)? What kinds of structure in the underlying problem enables easy design of good couplings, i.e., what makes a problem "Coupling friendly"?
- It almost seems that whenever Path Coupling works there is a "natural" notion of adjacency and a distance metric fixing which gives a rather easy proof of rapid mixing. For several problems for which the natural choice for these notions does not work, no known Coupling based proof seems to be in sight as well. It will be interesting to shed some light on this, and investigate how one may make Coupling work when most natural choices for doing Path Coupling do not work out.
- Finally there are several questions still open about designing and analyzing rapidly mixing Markov chains for specific sampling problems. Some of our favorite ones are:
  - Bipartite graphs with a given degree sequence (for sampling regular bipartite graphs, a rapidly mixing Markov Chain was given in [25]). More generally, contingency tables with given row and column sums (the $2 \times n$ case was solved in [13] using Path Coupling).
  - Independent sets in graphs with maximum degree 5. (The case $\Delta \leq 4$ has been considered in [29, 15], and a "negative" result for $\Delta \geq 6$ appears in [10].)
  - Proper k-colorings of a graph when $k < \frac{11}{6}\Delta$.
  - Perfect matchings in a general bipartite graph.

References

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

[2] D. Aldous. Some inequalities for reversible Markov chains. *Journal of the London Mathematical Society*, 25 (1982), pp. 564-576.

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

[4] A. Broder. How hard is it to marry at random? (On the approximation of the permanent). *Proc. of 18th STOC*, pp. 50-58, 1986.

[5] R. Bubley and M. Dyer. Path coupling: a technique for proving rapid mixing in Markov chains. *Proc. of 38th FOCS*, pp. 223-231, 1997.

[6] R. Bubley and M. Dyer. Faster random generation of linear extensions. *Proc. of the 9th ACM Symposium on Discrete Algorithms*, pp. 35-354, 1998.

[7] R. Bubley, M. Dyer and M. Jerrum. An elementary analysis of a procedure for sampling points in a convex body. *Random Structures and Algorithms*, 12 (1998), pp. 213-235.

[8] P. Diaconis and D. Strook. Geometric bounds for eigenvalues of Markov chains. *Annals of Applied Probability*, 1 (1991), pp. 36-61.

[9] M. Dyer and A. Frieze. Computing the volume of convex bodies: a case where randomness provably helps. In Probabilistic Combinatorics and its Applications, *Proc. of AMS Symposia in Applied Mathematics*, 44 (1991), pp. 123-170.

[10] M. Dyer, A. Frieze and M. Jerrum. On counting independent sets in sparse graphs. *Proc. of 40th FOCS*, pp. 210-217, 1999.

[11] M. Dyer, A. Frieze and R. Kannan. A random polynomial time algorithm for approximating the volume of convex bodies. *Journal of the ACM*, 38 (1991), pp. 1-17.

[12] M. Dyer, A. Frieze, R. Kannan, A. Kapoor, L. Perkovic and U. Vazirani. A sub-exponential time algorithm for approximating the number of solutions to a multidimensional knapsack problem. *Combinatorics, Probability and Computing*, 2 (1993), pp. 271-284.

[13] M. Dyer and C. Greenhill. A genuinely polynomial-time algorithm for sampling two-rowed contingency tables. *Proc. of the 25th International Colloquium on Automata, Languages and Programming (ICALP)*, pp. 339-350, Aalborg, Denmark (1998),

[14] M. Dyer and C. Greenhill, A more rapidly mixing Markov chain for graph colourings. *Random Structures and Algorithms*, 13 (1998), pp. 285-317.

[15] M. Dyer and C. Greenhill. On Markov chains for independent sets. *Journal of Algorithms*, 35 (2000), pp. 17 - 49.

[16] B. Hajek. Hitting time and occupation time bounds implied by drift analysis with applications. *Advances in Applied Probability*, 14 (1982), pp. 502-525.

[17] T. Feder and M. Mihail. Balanced Matroids. *Proc. of 24th STOC*, pp. 26-38, 1992.

[18] M. Jerrum. A very simple algorithm for estimating the number of k-colourings of a low-degree graph. *Random Structures and Algorithms*, 7 (1995), pp. 157-165.

[19] M. Jerrum. Mathematical foundations of the Markov chain Monte Carlo method. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, Algorithms and Combinatorics 16, Springer-Verlag, 1998, pp. 116-165.

[20] M. Jerrum and A. Sinclair. Approximating the permanent. *SIAM Journal on Computing*, 18 (1989), pp. 1149-1178.

[21] M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM Journal on Computing*, 22 (1993), pp. 1087-1116.

[22] M. Jerrum and A. Sinclair. The Markov chain Monte Carlo method: an approach to approximate counting and integration. In *Approximation Algorithms for NP-hard problems*, D.S Hochbaum ed., PWS Publishing, Boston, 1997, pp. 482-520.

[23] M. Jerrum, L. G. Valiant and V. V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43 (1986), pp. 169-188.

[24] N. Kahale. *A semidefinite bound for mixing rates of Markov chains*. DIMACS Technical Report 95-41, September 1995.

[25] R. Kannan, P. Tetali and S. Vempala. Simple Markov chain algorithms for generating bipartite graphs and tournaments. *Proc. of 8th SODA*, 1997.

[26] A. Karzanov and L. Khachiyan. *On the conductance of order Markov chains*, Technical Report DCS 268, Rutgers University, June 1990.

[27] V.S. Anil Kumar and H. Ramesh. Coupling vs. conductance for the Jerrum-Sinclair chain. *Proc. of 40th FOCS*, pp. 241-251, 1999.

[28] T. Leighton and S. Rao. An approximate max-flow min-cut theorem for uniform multicommodity flow problems with applications to approximation algorithms. *Proc. of 29th STOC*, pp. 422-431, 1988.

[29] M. Luby and E. Vigoda. Approximately counting up to four. *Proc. of 29th STOC*, pp. 682-687, 1997.

[30] P. Matthews. Generating random linear extensions of a partial order. *The Annals of Probability*, 19 (1991), pp. 1367-1392.

[31] M. Mihail. Conductance and convergence of Markov chains: a combinatorial treatment of expanders. *Proc. of the 30th FOCS*, pp. 526-531, 1989.

[32] M. Mihail. On coupling and the approximation of the permanent. *Information Processing Letters*, 30 (1989), pp. 91-95.

[33] B. Mohar. Isoperimetric numbers of graphs. *Journal of Combinatorial Theory, Series B*, 47 (1989), pp. 274-291.

[34] B. Morris and A. Sinclair. Random walks on truncated cubes and sampling 0-1 knapsack solutions. *Proc. of 40th FOCS*, pp. 230-240, 1999.

[35] P. Raghavan and C. D. Thompson. Randomized rounding: a technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7 (1987), pp. 365-374.

[36] A. Sinclair. *Algorithms for random generation and counting: a Markov chain approach*. Ph.D thesis, University of Edinburgh, June 1988.

[37] A. Sinclair. Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combinatorics, Probability and Computing*, 1 (1992), pp. 351-370.

[38] A. Sinclair and M. Jerrum. Approximate counting, uniform generation, and rapidly mixing Markov chains. *Information and Computation*, 82 (1989), pp. 93-133.

[39] S. Vadhan. Rapidly mixing Markov chains and their applications. Essay, *Churchill College, Cambridge University*, May 1996.

[40] E. Vigoda. Improved bounds for sampling colorings. *Proc. of 40th FOCS*, pp. 51-59, 1999.