

Mixing Times of Markov Chains for Self-Organizing Lists and Biased Permutations

Prateek Bhakta ^{*} Sarah Miracle [†] Dana Randall [‡] Amanda Pascoe Streib [§]

Abstract

We study the mixing time of a Markov chain \mathcal{M}_{nn} on permutations that performs nearest neighbor transpositions in the non-uniform setting, a problem arising in the context of self-organizing lists. We are given “positively biased” probabilities $\{p_{i,j} \geq 1/2\}$ for all $i < j$ and let $p_{j,i} = 1 - p_{i,j}$. In each step, the chain \mathcal{M}_{nn} chooses two adjacent elements k , and ℓ and exchanges their positions with probability $p_{\ell,k}$. Here we define two general classes and give the first proofs that the chain is rapidly mixing for both. In the first case we are given constants r_1, \dots, r_{n-1} with $1/2 \leq r_i \leq 1$ for all i and we set $p_{i,j} = r_i$ for all $i < j$. In the second we are given a binary tree with n leaves labeled $1, \dots, n$ and constants q_1, \dots, q_{n-1} associated with all of the internal vertices, and we let $p_{i,j} = q_{i \wedge j}$ for all $i < j$. Our bounds on the mixing time of \mathcal{M}_{nn} rely on bijections between permutations, *inversion tables* and *asymmetric simple exclusion processes* (ASEPs) that allow us to express moves of the chain in the context of these other combinatorial families. We also demonstrate that the chain is not always rapidly mixing by constructing an example requiring exponential time to converge to equilibrium. This proof relies on a reduction to biased lattice paths in \mathbb{Z}^2 .

1 Introduction

Sampling from the permutation group S_n is one of the most fundamental problems in probability theory. A natural Markov chain that has been studied extensively is a symmetric chain, \mathcal{M}_{nn} , that iteratively makes

nearest neighbor transpositions on adjacent elements. We are given a set of input probabilities $\mathbf{P} = \{p_{i,j}\}$ for all $1 \leq i, j \leq n$ with $p_{i,j} = 1 - p_{j,i}$. At each step, the Markov chain \mathcal{M}_{nn} uniformly chooses a pair of adjacent elements, i and j , and puts i ahead of j with probability $p_{i,j}$, and j ahead of i with probability $p_{j,i} = 1 - p_{i,j}$. The problem of biased permutations arises naturally from the Move-Ahead-One list update algorithm and was considered by Fill [8, 9]. In the MA1 protocol, elements are chosen according to some underlying distribution and they move up by one in a linked list after each request is serviced, if possible. Thus, the most frequently requested elements will move toward the front of the list and will require less access time. The transposition rates vary depending on i and j and we are always more likely to put things in order (of their request frequencies) than out of order. Fill asked for which $\mathbf{P} = \{p_{i,j}\}$ the chain is rapidly mixing.

Despite the simplicity of the model, only a few special cases are known. Wilson [18] showed that in the unbiased case when $p_{i,j} = 1/2$ for all i, j the mixing time is $\Theta(n^3 \log n)$, with upper and lower bounds within a factor of two. Subsequently Benjamini et al. [1] considered a constant bias version of this chain, where we are given a fixed parameter $0 \leq p \leq 1$ such that $p \neq 1/2$ and $p_{i,j} = p$ for all $i < j$ and $p_{i,j} = 1 - p$ for $i > j$. They relate this biased shuffling Markov chain to a chain on an asymmetric simple exclusion process (ASEP) and showed that they both converge in $\Theta(n^2)$ time. These bounds were matched by Greenberg et al. [10] who also generalized the result on ASEPs to sampling biased surfaces in two and higher dimensions in optimal $\Theta(n^d)$ time. Note that when the bias is a constant for all $i < j$ there are other methods for sampling from the stationary distribution, but studying the Markov chain \mathcal{M}_{nn} is of independent interest, partly because of the connection to ASEPs and other combinatorial structures. Finally, we also have polynomial bounds on the mixing time when each of the $p_{i,j}$ for $i < j$ is equal to $1/2$ or 1 ; in this case we are sampling linear extensions of a partial order over the set $\{1 \dots n\}$, and the chain \mathcal{M}_{nn} was shown by Bubley and Dyer [3] to mix in $O(n^3 \log n)$ time.

^{*}College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0765; pbhakta@gatech.edu. Supported in part by NSF CCF-0830367 and a Georgia Institute of Technology ARC Fellowship.

[†]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0765; sarah.miracle@gatech.edu. Supported in part by a DOE Office of Science Graduate Fellowship, NSF CCF-0830367 and an ARCS Scholar Award.

[‡]College of Computing, Georgia Institute of Technology, Atlanta, GA 30332-0765; randall@cc.gatech.edu. Supported in part by NSF CCF-0830367 and CCF-0910584.

[§]National Institute of Standards and Technology, Gaithersburg, MD 20899-8910; amanda.streib@nist.gov. Supported in part by the National Physical Sciences Consortium Fellowship and NSF CCF-0910584.

It is easy to see that \mathcal{M}_{nn} is not always rapidly mixing. Consider, for example, n elements $1 \dots n$ such that $p_{i,j} = 1$ for all $1 \leq i < j \leq n-1$, $p_{n,i} = .9$ for $j \leq n/2$ and $p_{i,n} = .9$ for $j > n/2$. Then the first $n-1$ elements will stay in order once they become ordered. All n places where the last element can be placed have nonzero stationary probability, but the configurations that have this last element at the beginning or end of the permutation will have exponentially larger stationary probability than the configuration that has this last element near the middle of the permutation. This defines an exponentially small cut in the state space and we can conclude that the nearest neighbor transposition chain must be slowly mixing for this choice of \mathbf{P} .

To avoid such situations, we restrict our attention to the *positively biased* setting where for all $i < j$, we have $1/2 \leq p_{i,j} \leq 1$. Note that any transposition that puts elements in the proper order has probability at least $1/2$, so starting at any permutation, we can always perform a series of transpositions to move to the ordered permutation $1, 2, \dots, n$ without ever decreasing the stationary probability. It is also worth noting that the classes for which the chain is known to mix rapidly are all positively biased. Fill [8, 9] conjectured that when \mathbf{P} is positively biased and also satisfies a monotonicity condition where $p_{i,j} \leq p_{i,j+1}$ and $p_{i,j} \geq p_{i+1,j}$ for all $1 \leq i < j \leq n$, then the chain is always rapidly mixing. In fact, he conjectured that the spectral gap is minimized when $p_{i,j} = 1/2$ for all i, j , a problem he refers to as the “gap problem.” Fill verified the conjecture for $n = 4$ and gave experimental evidence for slightly larger n .

In this paper, we make progress on the question of determining for which values of \mathbf{P} the chain \mathcal{M}_{nn} is rapidly mixing. First, we show that restricting \mathbf{P} to be positively biased is not sufficient to guarantee fast convergence to equilibrium. Our example uses a reduction to ASEPs and biased lattice paths. The construction is motivated by models in statistical physics that exhibit a phase transition arising from a “disordered phase” of high entropy and low energy, an “ordered phase” of high energy and low entropy, and a bad cut separating them that is both low energy and entropy. We note that this example does not satisfy the monotonicity condition of Fill, thus leaving his conjecture open, but does give insight into why bounding the mixing rate of the chain in more general settings has proven quite challenging.

In addition, we identify two new classes of input probabilities \mathbf{P} for which we can prove that the chain is rapidly mixing. It is important to note that these classes are not necessarily monotone. The first, which we refer to as “Choose Your Weapon,” we are given a set of input parameters $1/2 \leq r_1, \dots, r_{n-1} \leq 1$

representing each player’s ability to win a duel with his or her weapon of choice. When a pair of neighboring players are chosen to compete, the dominant player gets to choose the weapon, thus determining his or her probability of winning the match. In other words, we set $p_{i,j} = r_i$ when $i < j$. We show that the nearest neighbor transposition chain \mathcal{M}_{nn} is rapidly mixing for any choice of $\{r_i\}$. The second class, which we refer to as “League Hierarchies,” is defined by a binary tree with n leaves labeled $1, \dots, n$. We are given q_1, \dots, q_{n-1} with $1/2 \leq q_i \leq 1$ for all i , each associated with a distinct internal node in the tree. We then set $p_{i,j} = q_{i \wedge j}$ for all $i < j$. We imagine that the two subtrees under the root represent two different leagues, where each player from one league have a fixed advantage over each player from the other. Moreover, each league is subdivided into two sub-leagues, and each player from one has a fixed advantage over a player from the other, and so on recursively. We prove that there is a Markov chain based on transpositions (not necessarily nearest neighbors) that is always rapidly mixing for positively biased \mathbf{P} defined as League Hierarchies. Moreover, if the $\{q_i\}$ additionally satisfy “weak monotonicity” (i.e., $p_{i,j} \leq p_{i,j+1}$ if $j > i$) then the nearest neighbor chain \mathcal{M}_{nn} is also rapidly mixing. Note that both the choose-your-weapon and the tree-hierarchy classes are generalizations of the constant bias setting, which can be seen by taking all parameters r_i or q_i to be constant.

Our proofs rely on various combinatorial representations of permutations, including Inversion Tables and families of ASEPs. In each case there is a natural Markov chain based on (non necessarily adjacent) transpositions for which we can more easily bound the mixing time in the new context. We then interpret these new moves in terms of the original permutations in order to derive bounds on the mixing rate of the nearest neighbor transposition via comparison methods. These new chains that allow additional, but not necessarily all, transpositions are also interesting in the context of permutations and these related combinatorial families. Finally, we note that the choose-your-weapon class is actually a special case of the league-hierarchy class, but the proofs bounding the mixing rate of \mathcal{M}_{nn} are simpler and yield faster mixing times, so we present these proofs separately in Sections 4 and 5.

2 The Markov Chains \mathcal{M}_{nn} and \mathcal{M}_{tr}

We begin by formalizing the nearest neighbor and transposition Markov chains. Let $\Omega = S_n$ be the set of all permutations $\sigma = (\sigma(1), \dots, \sigma(n))$ of n integers. We consider Markov chains on Ω whose transitions transpose two elements of the permutation. Recall we are given a set \mathbf{P} , consisting of $p_{i,j} \in [0, 1]$ for each

$1 \leq i \neq j \leq n$, where for any $i < j$, $p_{i,j} \geq 1/2$ and $p_{j,i} = 1 - p_{i,j}$. The Markov chain \mathcal{M}_{nn} will sample elements from Ω as follows.

The Nearest Neighbor Markov chain \mathcal{M}_{nn}

Starting at any permutation σ_0 , repeat:

- At time t , select index $i \in [n-1]$ uniformly at random (u.a.r).
 - Exchange the elements $\sigma_t(i)$ and $\sigma_t(i+1)$ with probability $p_{\sigma_t(i+1), \sigma_t(i)}$ to obtain σ_{t+1} .
 - With probability $p_{\sigma_t(i), \sigma_t(i+1)}$ do nothing so that $\sigma_{t+1} = \sigma_t$.

The chain \mathcal{M}_{nn} connects the state space, since every permutation σ can move to the ordered permutation $(1, 2, \dots, n)$ (and back) using the bubble sort algorithm. Since \mathcal{M}_{nn} is also aperiodic, this implies that \mathcal{M}_{nn} is ergodic. For an ergodic Markov chain with transition probabilities \mathcal{P} , if some assignment of probabilities π satisfies the *detailed balance condition* $\pi(\sigma)\mathcal{P}(\sigma, \tau) = \pi(\tau)\mathcal{P}(\tau, \sigma)$ for every $\sigma, \tau \in \Omega$, then π is the stationary distribution of the Markov chain [13]. It is easy to see that for \mathcal{M}_{nn} , the distribution $\pi(\sigma) = \prod_{(i < j)} p_{\sigma(i), \sigma(j)} / Z$, where Z is the normalizing constant $\sum_{\sigma \in \Omega} \prod_{(i < j)} p_{\sigma(i), \sigma(j)}$, satisfies detailed balance, and is thus the stationary distribution.

The Markov chain \mathcal{M}_{tr} can make any transposition at each step, while maintaining the stationary distribution π . The transition probabilities of \mathcal{M}_{tr} can be quite complicated, since swapping two distant elements in the permutation consists of many transitions of \mathcal{M}_{nn} , each with different probabilities. In the following sections, we will introduce two other Markov chains whose transitions are a subset of those of \mathcal{M}_{tr} , but for which we can describe the transition probabilities succinctly.

The relevant measure of the number of times we need to repeat steps of a Markov chain \mathcal{M} so that we are close (within total variation distance ϵ of stationarity) is the *mixing time* $\tau(\epsilon)$. The *total variation distance* between the stationary distribution π and the distribution of the Markov Chain at time t is

$$\|\mathcal{P}^t, \pi\|_{tv} = \max_{x \in \Omega} \frac{1}{2} \sum_{y \in \Omega} |\mathcal{P}^t(x, y) - \pi(y)|,$$

where $\mathcal{P}^t(x, y)$ is the t -step transition probability. For all $\epsilon > 0$, we define

$$\tau(\epsilon) = \min\{t : \|\mathcal{P}^t, \pi\|_{tv} \leq \epsilon, \forall t' \geq t\}.$$

We say that a Markov chain is *rapidly mixing* if there exists a polynomial p such that $\tau_\epsilon = O(p(n, \log(\epsilon^{-1})))$ where n is the size of each configuration in Ω .

3 A Positively Biased \mathbf{P} that is Slowly Mixing

We begin by presenting an example that is positively biased yet takes exponential time to mix. In particular, we show that there are positively biased \mathbf{P} for which the chains \mathcal{M}_{nn} and even \mathcal{M}_{tr} require exponential time to converge to equilibrium. This example was discovered by two of the authors as a result of their previous work studying tile-based self-assembly models [14] and is of independent interest in this setting. We use a reduction from biased permutations to multiple particle ASEP configurations with n zeros and n ones. The resulting ASEP configurations are in bijection with staircase walks [10], which are paths on the Cartesian lattice from $(0, n)$ to $(n, 0)$ that always go to the right or down (see Figure 1b). In [10], Greenberg et al. examined the Markov chain which attempts to swap a neighboring $(0, 1)$ pair, which essentially adds or removes a unit square from the region below the walk, with probability depending on the position of that unit square. The probability of each walk w is proportional to $\prod_{xy < w} \lambda_{xy}$, where the bias $\lambda_{xy} \geq 1/2$ is assigned to the square at (x, y) and $xy < w$ whenever the square at (x, y) lies underneath the walk w . We show that there are settings of the $\{\lambda_{xy}\}$ which cause the chain to be slowly mixing from any starting configuration (or walk). In particular, we show that at stationarity the most likely configurations will be concentrated near the diagonal from $(0, n)$ to $(n, 0)$ (the high entropy, low energy states) or they will extend close to the point (n, n) (the high energy, low entropy states) but it will be unlikely to move between these sets of states because there is a bottleneck that has both low energy and low entropy. Finally, we use the reduction from biased permutations to biased lattice paths to produce a positively biased set of probabilities \mathbf{P} for which \mathcal{M}_{nn} also requires exponential time to mix.

Suppose, for ease of notation, that we are sampling permutations with $2n$ entries (having an odd number of elements will not cause qualitatively different behavior). Let $M = 2n^{2/3}$, $0 < \delta < \frac{1}{2}$ be a constant, $\epsilon = 1/n^2$. For $i < j \leq n$ or $n < i < j$, $p_{i,j} = 1$, ensuring that once the elements $1, 2, \dots, n$ get in order, they stay in order (and similarly for the elements $n+1, n+2, \dots, 2n$). The $p_{i,j}$ values for $i \leq n < j$ are defined as follows (see Figure 1a):

$$p_{i,j} = \begin{cases} 1 - \delta & \text{if } i + 2n - j + 1 \geq n + M; \\ \frac{1}{2} + \epsilon & \text{otherwise.} \end{cases}$$

We identify sets S_1, S_2, S_3 such that $\pi(S_2)$ is exponentially smaller than both $\pi(S_1)$ and $\pi(S_3)$, but to get between S_1 and S_3 , \mathcal{M}_{nn} and \mathcal{M}_{tr} must pass through S_2 , the cut. Since the smallest (largest) n elements of the biased permutation never change or-

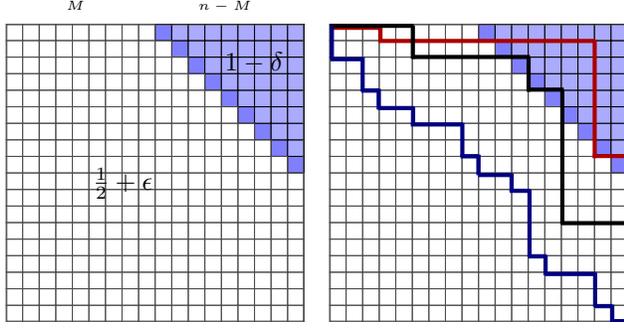


Figure 1: (a) Fluctuating bias with exponential mixing time. (b) Staircase walks in S_1, S_2 , and S_3 .

der once they get put into increasing order, permutations with these elements out of order have zero stationary probability. Hence they do not contribute to $\pi(S_1), \pi(S_2)$, or $\pi(S_3)$. Hence we can represent the smallest n numbers as ones and the largest n numbers as zeros, assuming that within each class the elements are in increasing order. Given a permutation σ , let $f(\sigma)$ be a sequence of ones and zeros, where $f(\sigma)_i = 1$ if $i \leq n$ and 0 otherwise. Then if σ is such that elements $1, 2, \dots, n$ and elements $n+1, n+2, \dots, 2n$ are each in order, $f(\sigma)$ maps σ uniquely to a staircase walk. For example, the permutation $\sigma = (5, 1, 7, 8, 4, 3, 6, 2)$ maps to $f(\sigma) = (0, 1, 0, 0, 1, 1, 0, 1)$. The probability that an adjacent 1 and a 0 swap in \mathcal{M}_{nn} then depends on how many ones and zeros occur before that point in the permutation. Specifically, if element i is a 0 and element $i+1$ is a 1 then we swap them with probability $\frac{1}{2} + \epsilon$ if the number of ones occurring before position x plus the number of zeros occurring after $i+1$ is less than $n+M-1$. Otherwise, they swap with probability $1 - \delta$. Equivalently, the probability of adding a unit square at position $v = (x, y)$, which is called the *bias at $v = (x, y)$* , is $\frac{1}{2} + \epsilon$ if $x+y \leq n+M$, and $1 - \delta$ otherwise; see Figure 1b. We will show that in this case, the Markov chain is slow. The idea is that in the stationary distribution, there is a good chance that the ones and zeros will be well-mixed, since this is a high entropy situation. However, the identity permutation also has high weight, and the parameters are chosen so that the entropy of the well-mixed permutations balances with the energy of the maximum (identity) permutation, and that to get between them is not very likely (low entropy and low energy). We prove that for the set \mathbf{P} defined above, \mathcal{M}_{nn} and \mathcal{M}_{tr} have a bad cut. Then we use the *conductance* to prove \mathcal{M}_{nn} and \mathcal{M}_{tr} are slowly mixing. For an ergodic Markov chain with distribution π , the

conductance is defined as

$$\Phi = \min_{\substack{S \subseteq \Omega \\ \pi(S) \leq 1/2}} \sum_{s_1 \in S, s_2 \in \bar{S}} \pi(s_1) \mathcal{P}(s_1, s_2) / \pi(S).$$

We will show that the bad cut (S_1, S_2, S_3) implies that Φ is exponentially small. The following theorem relates the conductance and mixing time (see, e.g., [11, 16]).

THEOREM 3.1. *For any Markov chain with conductance Φ and mixing time $\tau(\epsilon)$, for all $\epsilon > 0$ we have*

$$\tau(\epsilon) \geq \left(\frac{1}{4\Phi} - 1/2 \right) \log \left(\frac{1}{2\epsilon} \right).$$

We are now ready to prove the main theorem of the section.

THEOREM 3.2. *There exists a positively biased preference set \mathbf{P} for which the mixing time $\tau(\epsilon)$ of the Markov chain \mathcal{M}_{nn} satisfies*

$$\tau(\epsilon) \geq \frac{1}{2} \left(e^{n^{1/3}-1} - 1 \right) \log \left(\frac{1}{2\epsilon} \right).$$

Proof. For a staircase walk σ , define the *height* of σ_i as $\sum_{j \leq i} \sigma_j$, and let $\max(\sigma)$ be the maximum height of σ_i over all $1 \leq i \leq 2n$. Let S_1 be the set of configurations σ such that $\max(\sigma) < n+M$, S_2 the set of configurations such that $\max(\sigma) = n+M$, and S_3 the set of configurations such that $\max(\sigma) > n+M$. That is, S_1 is the set of configurations that never reach the dark blue diagonal in Figure 1b, S_2 is the set whose maximum peak is on the dark blue line, and S_3 is the set which crosses that line and contains squares in the light blue triangle. Define $\gamma = (1/2 + \epsilon)/(1/2 - \epsilon)$, which is the ratio of two configurations that differ by swapping a $(0, 1)$ pair with probability $\frac{1}{2} + \epsilon$. First we notice that since the maximal tiling is in S_3 ,

$$\pi(S_3) \geq \frac{1}{Z} \gamma^{n^2 - \frac{(n-M)^2}{2}} (\delta^{-1} - 1)^{\frac{(n-M)^2}{2}}.$$

Also, $\pi(S_1) = \frac{1}{Z} \sum_{\sigma \in S_1} \gamma^{A(\sigma)}$, where $A(\sigma)$ is the number of unit squares below σ . We have that

$$\begin{aligned} \pi(S_1) &= \frac{1}{Z} \sum_{\sigma \in S_1} \gamma^{A(\sigma)} \leq \frac{1}{Z} \sum_{\sigma \in S_1} \gamma^{n^2 - \frac{(n-M)^2}{2}} \\ &\leq \frac{1}{Z} \binom{2n}{n} \gamma^{n^2 - \frac{(n-M)^2}{2}} \\ &\leq \frac{1}{Z} (2e)^n \gamma^{n^2 - \frac{(n-M)^2}{2}} \\ &\leq \frac{1}{Z} \gamma^{n^2 - \frac{(n-M)^2}{2}} (\delta^{-1} - 1)^{\frac{(n-M)^2}{2}} \\ &\leq \pi(S_3) \end{aligned}$$

for large enough n , since $1/\delta > 2$ is a constant. Hence $\pi(S_1) \leq \pi(S_3)$. We will show that $\pi(S_2)$ is exponentially small in comparison to $\pi(S_1)$ (and hence also to $\pi(S_3)$).

$$\pi(S_2) = \frac{1}{Z} \sum_{\sigma \in S_2} \gamma^{A(\sigma)} \leq \frac{\gamma^{n^2} |S_2|}{Z}.$$

We bound $|S_2|$ as follows. The unbiased Markov chain is equivalent to a simple random walk $W_{2n} = X_1 + X_2 + \dots + X_{2n} = 0$, where $X_i \in \{+1, -1\}$ and where a $+1$ represents a step to the right and a -1 represents a step down. We call this random walk *tethered* since it is required to end at 0 after $2n$ steps. Compare walk W_{2n} with the untethered simple random walk $W'_{2n} = X'_1 + X'_2 + \dots + X'_{2n}$.

$$\begin{aligned} P\left(\max_{1 \leq t \leq 2n} W_t \geq M\right) &= P\left(\max_{1 \leq t \leq 2n} W'_t \geq M \mid W'_{2n} = 0\right) \\ &= \frac{P(\max_{1 \leq t \leq 2n} W'_t \geq M)}{P(W'_{2n} = 0)} \\ &= \frac{2^{2n}}{\binom{2n}{n}} P\left(\max_{1 \leq t \leq 2n} W'_t \geq M\right) \\ &\approx \sqrt{\pi n} P\left(\max_{1 \leq t \leq 2n} W'_t \geq M\right). \end{aligned}$$

Since the $\{X'_i\}$ are independent, we can use Chernoff bounds to see that

$$P\left(\max_{1 \leq t \leq 2n} W'_t \geq M\right) \leq 2n P(W'_{2n} \geq M) \leq 2ne^{-\frac{M^2}{2n}}.$$

Together these show that

$$P\left(\max_{1 \leq t \leq 2n} W_t \geq M\right) < e^{-n^{1/3}},$$

by definition of M . Therefore we have

$$\begin{aligned} \pi(S_2) &\leq \frac{1}{Z} \gamma^{n^2} |S_2| \leq \frac{1}{Z} \binom{2n}{n} e^{-n^{1/3}} \\ &\leq \frac{1}{Z} \binom{2n}{n} e^{-n^{1/3}+1} (1 - e^{-n^{1/3}}) \\ &\leq \frac{1}{Z} |S_1| e^{-n^{1/3}+1} \\ &\leq e^{-n^{1/3}+1} \pi(S_1), \end{aligned}$$

as desired. Thus, the conductance satisfies

$$\begin{aligned} \Phi &\leq \sum_{x \in S_1} \frac{\pi(x)}{\pi(S_1)} \sum_{y \in S_2} P(x, y) \\ &\leq \sum_{x \in S_1} \frac{\pi(x)}{\pi(S_1)} \pi(S_2) \\ &\leq e^{-n^{1/3}+1} \pi(S_1) \leq \frac{e^{-n^{1/3}+1}}{2}. \end{aligned}$$

Hence, by Theorem 3.1, $\tau(\epsilon)$, the mixing time of \mathcal{M}_{nn} satisfies

$$\tau(\epsilon) \geq \frac{1}{2} \left(e^{n^{1/3}-1} - 1 \right) \log \left(\frac{1}{2\epsilon} \right).$$

□

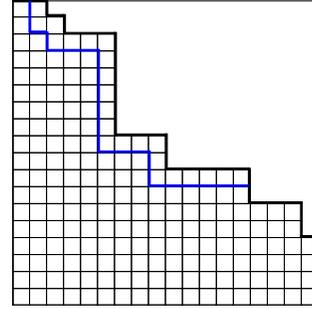


Figure 2: A move that swaps an arbitrary $(1, 0)$ pair.

In fact, this proof can be extended to the more general Markov chain where we can swap any 1 with any 0, as long as we maintain the correct stationary distribution. This is easy to see, because any move that swaps a single 1 with a single 0 can only change the maximum height by at most 2 (see Figure 2). If we expand S_2 to include all configurations with maximum height $n + M$ or $n + M + 1$, $\pi(S_2)$ is still exponentially smaller than $\pi(S_1) \leq \pi(S_3)$. Hence the Markov chain that swaps an arbitrary $(1, 0)$ pair still takes exponential time to converge.

4 Choose Your Weapon

Despite the slow mixing example outlined in the previous section, there are many cases for which the chain will be rapidly mixing. We define two new classes for which we can rigorously demonstrate this and we provide the proofs in the next two sections.

For the first class, imagine a community of n people, each with a unique combative talent. Each member has his or her weapon of choice, and a competition with any other member of the community using this weapon affords that person a fixed advantage. When two people are chosen to compete, they each prefer using their own weapon of choice, so we resolve this by letting the person with the higher rank (e.g., age, seniority, etc.) choose the weapon they both will use. At any point in time our competitors are ordered and nearest neighbors are randomly selected to compete, where upon the winner is moved in front of the loser in the ordering.

To formalize the “Choose Your Weapon” scenario, we are given $1/2 \leq r_1, r_2, \dots, r_{n-1} \leq 1$ and the set \mathbf{P} satisfies $p_{i,j} = r_i$, if $i < j$ and $p_{i,j} = 1 - p_{j,i}$ if $j < i$.

The moves of the nearest neighbor Markov chain \mathcal{M}_{nn} formalize the competitions, and our goal is to bound the mixing rate of this chain. Notice that this class includes the constant bias case studied by Benjamini et al. as a special case, and indeed our analysis yields an independent and simpler proof that the nearest neighbor Markov chain \mathcal{M}_{nn} is rapidly mixing in that context.

We shall show that the chain \mathcal{M}_{nn} is always rapidly mixing for probabilities \mathbf{P} defined in this way. Our proof relies on a bijection between permutations and *Inversion Tables* [12, 17] that, for each element i , record how many elements $j > i$ come before i in the permutation. We consider a Markov chain \mathcal{M}_{inv} that simply increments or decrements a single element of the inversion table in each step; using the bijection with permutations this corresponds to transpositions of elements that are not necessarily nearest neighbors to the Markov chain \mathcal{M}_{nn} . Remarkably, this allows \mathcal{M}_{inv} to decompose into a product of simple one-dimensional random walks and bounding the convergence time is very straightforward. Finally, we use comparison techniques [5, 15] to bound the mixing time of the nearest neighbor chain \mathcal{M}_{nn} .

4.1 The inversion table representation. The Markov chain \mathcal{M}_{inv} acts on the *inversion table* for the permutation [12, 17], which has an entry for each $i \in [n]$ counting the number of inversions involving i ; that is, the number of values $j > i$ where j comes before i in the permutation (see Figure 3). It is easy to see that the i th element of the inversion table is an integer between 0 and $n - i$. In fact, the function I is a bijection between the set of permutations and the set \mathcal{I} of all possible inversion tables (all sequences $X = (x_1, x_2, \dots, x_n)$ where $0 \leq x_i \leq n - i$ for all $i \in [n]$). To see this, we will construct a permutation from any inversion table $X \in \mathcal{I}$. Place the element 1 in the $(x_1 + 1)$ st position of the permutation. Next, there are $n - 1$ slots remaining. Among these, place the element 2 in the $(x_2 + 1)$ st position remaining (ignoring the slot already filled by 1). Continuing, after placing $i - 1$ elements into the permutation, there are $n - i + 1$ slots remaining, and we place the element i into the $(x_i + 1)$ st position among the remaining slots. This proves that I is a bijection from S_n to \mathcal{I} .

Given this bijection, a natural algorithm for sampling permutations is the following local Markov chain on inversion tables: select a position $i \in [n]$ and attempt

$$\begin{aligned} \sigma &= 8 \ 1 \ 5 \ 3 \ 7 \ 4 \ 6 \ 2 \\ I(\sigma) &= 1 \ 7 \ 2 \ 3 \ 1 \ 2 \ 1 \ 0 \end{aligned}$$

Figure 3: The inversion table for a permutation.

to either add one or subtract one from x_i , according to the appropriate probabilities. This amounts to adding or removing an inversion involving i without affecting the number of inversions involving any other integer, and is achieved by swapping the element i with an element $j > i$ such that every element in between is smaller than both i and j . If i moves ahead of j , this move happens with probability $p_{i,j}$ because for each k originally between i and j , $p_{k,i} = r_k = p_{k,j}$, so the net effect of the move is neutral. The detailed balance condition ensures that π is the correct stationary distribution. Formally, the Markov chain \mathcal{M}_{nn} is defined as follows.

The Inversion Markov chain \mathcal{M}_{inv}

Starting at any permutation σ_0 , repeat:

- Select $(i, b) \in [n] \times \{-1, +1\}$ u.a.r.
 - If $b = +1$ let j be the first element after i in σ_t such that $j > i$. With probability $p_{j,i}$, obtain σ_{t+1} from σ_t by swapping i and j .
 - If $b = -1$ let j be the last element before i in σ_t such that $j > i$. With probability $p_{i,j}$, obtain σ_{t+1} from σ_t by swapping i and j .
- Otherwise, $\sigma_{t+1} = \sigma_t$.

This Markov chain contains the moves of \mathcal{M}_{nn} (and therefore also connects the state space). Although elements can jump across several elements, it is still fairly local compared with the general transposition chain \mathcal{M}_{tr} which has $\binom{n}{2}$ choices at every step, since \mathcal{M}_{inv} has at most $2n$.

4.2 Rapid mixing of \mathcal{M}_{inv} . The inversion Markov chain \mathcal{M}_{inv} can be viewed as a product of n independent processes. The i th process is a one-dimensional random walk bounded between 0 and $n - i$ that moves up by one with probability r_i and down by one with probability $1 - r_i$; its mixing time is $O(n^2 \log n)$, unless r_i is bounded away from $1/2$, in which case its mixing time is $O(n)$. We make moves in each chain with probability $1/n$, since we update one random walk at a time. The main tool we use for proving rapid mixing of \mathcal{M}_{inv} is the path coupling theorem. We use the following version due to Dyer and Greenhill [7].

THEOREM 4.1. (PATH COUPLING) *Let d be an integer-valued metric defined on $\Omega \times \Omega$ which takes values in $\{0, \dots, B\}$. Let U be a subset of $\Omega \times \Omega$ such that for all $(X_t, Y_t) \in \Omega \times \Omega$ there exists a path $X_t = Z_0, Z_1, \dots, Z_r = Y_t$ between X_t and Y_t such that $(Z_i, Z_{i+1}) \in U$ for*

$0 \leq i < r$ and

$$\sum_{i=0}^{r-1} d(Z_i, Z_{i+1}) = d(X_0, X_t).$$

Let M be a Markov chain on Ω with transition matrix P and mixing time $\tau(\epsilon)$. Consider any random function $f : \Omega \rightarrow \Omega$ such that $P[f(X) = Y] = P(X, Y)$ for all $X, Y \in \Omega$, and define a coupling of the Markov chain by $(X_t, Y_t) \rightarrow (X_{t+1}, Y_{t+1}) = (f(X_t), f(Y_t))$. If there exists $\beta < 1$ such that

$$E[d(X_{t+1}, Y_{t+1})] \leq \beta d(X_t, Y_t),$$

for all $(X_t, Y_t) \in U$, then the mixing time of M satisfies

$$\tau(\epsilon) \leq \frac{\ln(B\epsilon^{-1})}{1 - \beta}.$$

Now we are ready to prove the following theorem, bounding the mixing time of \mathcal{M}_{inv} .

THEOREM 4.2. *The mixing time of \mathcal{M}_{inv} on the uniform distribution over permutations on n elements satisfies $\tau(\epsilon) = O(n^3 \log(n\epsilon^{-1}))$.*

Proof. We define a distance metric d on pairs X, Y of inversion tables as a sum over the distances between the entries x_i and y_i , where the distance between x_i and y_i is defined by $\sum_{j=x_i}^{y_i-1} \phi(j)$ if $x_i \leq y_i$, where $\phi(j) = \cos(\pi/3(\frac{j}{n} - \frac{1}{2}))$, for $j = 1, 2, \dots, n-1$. The function ϕ is log-concave and symmetric across $n/2$. The distance d is

$$d(X, Y) = \sum_{i=1}^n \sum_{j=\min\{x_i, y_i\}}^{\max\{x_i, y_i\}-1} \phi(j).$$

We use the natural coupling on inversion tables where we choose the same element i for both $X = X_t$ and $Y = Y_t$ for each step t . A pair of inversion tables (X, Y) is in $U \subset \mathcal{I} \times \mathcal{I}$ if Y can be obtained from X by adding or subtracting 1 from a single x_i . Consider such a pair, and without loss of generality, suppose $y_i = x_i + 1$. We will show that the expected change in distance after one step of \mathcal{M}_{inv} is always negative. The result then follows from Theorem 4.1. Every move happens with probability $1/(4n)$. Notice that there are two moves that decrease the distance; namely, choosing i and subtracting 1 from y_i or adding 1 to x_i , and for each of these, the change in distance is $-\phi(x_i)$. There are also two moves that increase the distance. Adding 1 to y_i results in an increase of $\phi(y_i) = \phi(x_i + 1)$, and subtracting 1 from x_i results in a decrease of $\phi(x_i - 1)$. Notice that all other moves (those involving a different

one-dimensional process) are neutral, because they occur with the same probability in both chains. Hence the expected change in distance is $E[d(X_{t+1}, Y_{t+1}) - d(X_t, Y_t)] = \frac{1}{4n}(-2\phi(x_i) + \phi(x_i - 1) + \phi(x_i + 1))$. Using the trigonometric identity $\cos(\alpha + \gamma) + \cos(\alpha - \gamma) = 2\cos(\alpha)\cos(\gamma)$, we see that

$$\begin{aligned} \phi(x_i - 1) + \phi(x_i + 1) &= 2\phi(x_i)\cos(\pi/(3n)) \\ &\leq 2\phi(x_i)\left(1 - \frac{(\pi/3)^2}{2n^2}\right), \end{aligned}$$

see Wilson[18] for more details. Thus, since $d(X_t, Y_t) = \phi(x_i) \geq \cos(\pi/3) = 1/2$,

$$\begin{aligned} E[d(X_{t+1}, Y_{t+1})] &\leq d(X_t, Y_t) + \frac{1}{4n}2\phi(x_i)\left(-1 + 1 - \frac{(\pi/3)^2}{2n^2}\right) \\ &\leq d(X_t, Y_t)\left(1 - \frac{\phi(x_i)\pi^2}{18n^3}\right) \\ &\leq d(X_t, Y_t)\left(1 - \frac{\pi^2}{36n^3}\right). \end{aligned}$$

Moreover, the largest distance between any two inversion tables is

$$d(X_t, Y_t) = \sum_{i=1}^n \sum_{j=1}^{n-i-1} \phi(j) \leq \sum_{i=1}^n n - i - 1 = \frac{n(n-3)}{2}.$$

Hence by Theorem 4.1,

$$\tau(\epsilon) \leq \frac{36n^3}{\pi^2} \ln(n^2\epsilon^{-1}) = O(n^3 \log(n\epsilon^{-1})).$$

□

REMARK 4.1. *The same proof also applies to the case where the probability of swapping i and j depends on the object with lower rank (i.e., we are given r_2, \dots, r_n and we let $p_{i,j} = r_j$ for all $i < j$). This case is related to a variant of the MA1 list update algorithm, where if a record is requested, we try to move the associated record x ahead of its immediate predecessor in the list, if it exists. If it has higher rank than its predecessor, then it always succeeds, while if its rank is lower we move it ahead with probability $f_x = r_x/(1 + r_x) \leq 1$.*

The case that some of the $r_i = 1/2$ also mixes in time $O(n^3 \log n)$ by the same proof. However, when each r_i is bounded away from $1/2$, we have the following theorem.

THEOREM 4.3. *Let $1/2 < r_1, r_2, \dots, r_{n-1} < 1$ be constants. If $\mathbf{P} = \{p_{i,j} = r_{\min\{i,j\}}\}$, then the mixing time of \mathcal{M}_{inv} satisfies*

$$\tau(\epsilon) = O(n^2 \ln(n\epsilon^{-1})).$$

Proof. As above, we use path coupling. The set U is defined in the same way, but we will use a different distance metric d . Let $\alpha_i = 1/(2(1 - r_i))$ and define

$$d(X, Y) = \sum_{i=1}^n \sum_{j=\min\{x_i, y_i\}}^{\max\{x_i, y_i\}-1} \alpha_i^j.$$

Let $(X, Y) \in U$ and suppose Y is obtained from X by adding 1 to x_i . Then as before, any move of \mathcal{M}_{inv} whose smaller index is not i succeeds or fails with the same probability in X and Y . There are two moves that decrease the distance: adding 1 to x_i , which happens with probability $(1 - r_i)/(4n)$, or subtracting 1 from y_i , which happens with probability $r_i/(4n)$. Both of these moves decrease the distance by $\alpha_i^{x_i}$. On the other hand, \mathcal{M}_{inv} proposes adding 1 to y_i with probability $(1 - r_i)/(4n)$, which increases the distance by $\alpha_i^{x_i+1}$, and \mathcal{M}_{inv} proposes subtracting 1 from x_i and succeeds with probability $r_i/(4n)$, increasing the distance by $\alpha_i^{x_i-1}$. Thus the expected change in distance is

$$\begin{aligned} E[d(X_{t+1}, Y_{t+1}) - d(X_t, Y_t)] &= \frac{1}{4n} (-\alpha_i^{x_i} + (1 - r_i)\alpha_i^{x_i+1} + r_i\alpha_i^{x_i-1}) \\ &= \frac{\alpha_i^{x_i-1}}{4n} (-\alpha_i + (1 - r_i)\alpha_i^2 + r_i) \\ &= \frac{\alpha_i^{x_i-1}}{4n} \left(r_i - \frac{1}{4(1 - r_i)} \right) \\ &= \frac{\alpha_i^{x_i}}{2n} \cdot \frac{-(2r_i - 1)^2}{4} \\ &= \frac{-d_t(2r_i - 1)^2}{8n}, \end{aligned}$$

since $d_t = \alpha_i^{x_i}$. Hence $E[d_{t+1}] \leq d_t(1 - (2r_i - 1)^2/(8n))$. Moreover, the maximum distance between any two inversion tables is

$$B = \sum_{i=1}^{n-1} 1 + \alpha_i + \dots + \alpha_i^{n-i} = \sum_{i=1}^{n-1} \frac{\alpha_i^n - 1}{\alpha_i - 1} = O(n\alpha_{\max}^n),$$

where $\alpha_{\max} = \max_i \{\alpha_i\} = \frac{1}{2(1 - r_{\max})}$. Hence

$$\ln(B\epsilon^{-1}) = O(n \log(n\epsilon^{-1})),$$

so by Theorem 4.1, we have

$$\tau(\epsilon) = O(n^2 \log(n\epsilon^{-1})).$$

4.3 Comparing \mathcal{M}_{inv} with \mathcal{M}_{nn} . The comparison method can be used to infer the mixing time of one chain given the mixing time of another, similar chain. If P' and P are the transition matrices of two reversible Markov chains on the same state space Ω with the same stationary distribution π , the comparison method [5, 15] allows us to relate the mixing times of these two chains. Let $E(P) = \{(\sigma, \beta) : P(\sigma, \beta) > 0\}$ and $E(P') = \{(\sigma, \beta) : P'(\sigma, \beta) > 0\}$ denote the sets of edges of the two graphs, viewed as directed graphs. For each σ, β with $P'(\sigma, \beta) > 0$, define a path $\gamma_{\sigma\beta}$ using a sequence of states $\sigma = \sigma_0, \sigma_1, \dots, \sigma_k = \beta$ with $P(\sigma_i, \sigma_{i+1}) > 0$, and let $|\gamma_{\sigma\beta}|$ denote the length of the path. Let $\Gamma(v, \omega) = \{(\sigma, \beta) \in E(P') : (v, \omega) \in \gamma_{\sigma\beta}\}$ be the set of paths that use the transition (v, ω) of P . Finally, let $\pi_* = \min_{\rho \in \Omega} \pi(\rho)$ and define

$$A = \max_{(v, \omega) \in E(P)} \frac{1}{\pi(v)P(v, \omega)} \sum_{\Gamma(v, \omega)} |\gamma_{\sigma\beta}| \pi(\sigma) P'(\sigma, \beta).$$

We use the following formulation of the comparison method [15].

THEOREM 4.4. *Given two Markov chains each with stationary distribution π , transition matrices P and P' and mixing times $\tau(\epsilon)$ and $\tau'(\epsilon)$, respectively. Define A and π_* as above, then for $0 < \epsilon < 1$, we have*

$$\tau(\epsilon) \leq \frac{4 \log(1/(\epsilon\pi_*))}{\log(1/2\epsilon)} A \tau'(\epsilon).$$

Here, we use 4.4 to infer that \mathcal{M}_{nn} is fast from the mixing time of \mathcal{M}_{inv} . We assume that each $p_{i,j}$ is a constant less than 1; this is to ensure a good comparison between the spectral gap and the mixing time. If this condition is not satisfied, then the proofs still go through and will give a bound on the spectral gap, but will not provide a good bound on the mixing time.

THEOREM 4.5. *Let $1/2 < r_1, r_2, \dots, r_{n-1} < 1$ be constants. If $\mathbf{P} = \{p_{i,j} = r_{\min\{i,j\}}\}$, then the mixing time of \mathcal{M}_{nn} satisfies*

$$\tau(\epsilon) = O(n^7 \log(n/\epsilon)).$$

Here we are using the bound from Theorem 4.2, and if each $p_{i,j}$ is bounded away from $1/2$ then we would get a better bound using Theorem 4.3.

Proof. In order to apply Theorem 4.4, we need to define, for any transition $e = (\sigma, \beta)$ of the Markov chain \mathcal{M}_{inv} , a sequence of transitions of \mathcal{M}_{nn} . Let e be a transition of \mathcal{M}_{inv} which performs a transposition on elements $\sigma(i)$ and $\sigma(j)$, where $i < j$. Recall \mathcal{M}_{inv} can only swap $\sigma(i)$ and $\sigma(j)$ if all the elements between them are

□

smaller than both $\sigma(i)$ and $\sigma(j)$. To obtain a sufficient bound on the congestion along each edge, we ensure that in each step of the path, we do not decrease the weight of the configuration. This is easy to do; in the first stage, move $\sigma(i)$ to the right, one step at a time, until it swaps with $\sigma(j)$. This removes an inversion of the type $(\sigma(i), \sigma(k))$ for every $i < k < j$, so clearly we have not decreased the weight of the configuration at any step. Next, move $\sigma(j)$ to the left, one step at a time, until it reaches position i . This completes the move e , and at each step, we are adding back an inversion of the type $(\sigma(j), \sigma(k))$ for some $i < k < j$. Since $\sigma(k) = \min\{\sigma(j), \sigma(k)\} = \min\{\sigma(i), \sigma(k)\}$, we have $p_{\sigma(i), \sigma(k)} = p_{\sigma(j), \sigma(k)}$ for every $i < k < j$, so in this stage we restore all the inversions destroyed in the first stage, for a net change of $p_{\sigma(i), \sigma(j)}$.

Given a transition (v, ω) of \mathcal{M}_{nn} we must upper bound the number of canonical paths $\gamma_{\sigma\beta}$ that use this edge, which we do by bounding the amount of information needed in addition to (v, ω) to determine σ and β uniquely. will provide a bound on the number of paths which use (v, ω) . For moves in the first stage, all we need to remember is $\sigma(i)$, because we know $\sigma(j)$ (it is the element moving forward). We also need to remember where $\sigma(j)$ came from. Given this information along with v and ω we can uniquely recover (σ, β) . Thus there are at most n^2 paths which use any edge (v, ω) . Also, notice that the maximum length of any path is $2n$.

Next we bound the quantity A which is needed to apply Theorem 4.4. Recall that we have guaranteed that $\pi(\sigma) \leq \max\{\pi(v), \pi(\omega)\}$. Assume first that $\pi(\sigma) \leq \pi(v)$. Then

$$\begin{aligned} A &= \max_{(v, \omega) \in E(P)} \left\{ \frac{1}{\pi(v)P(v, \omega)} \sum_{\Gamma(v, \omega)} |\gamma_{\sigma\beta}| \pi(\sigma) P'(\sigma, \beta) \right\} \\ &\leq \max_{(v, \omega) \in E(P)} \sum_{\Gamma(v, \omega)} 2n \frac{P'(\sigma, \beta)}{P(v, \omega)} \\ &\leq \max_{(v, \omega) \in E(P)} \sum_{\Gamma(v, \omega)} 2n \frac{1/(2n)}{\frac{\lambda}{(1+\lambda)(n-1)}} = O(n^3). \end{aligned}$$

If, on the other hand, $\pi(\sigma) \leq \pi(\omega)$, then we use detailed

balance to obtain:

$$\begin{aligned} A &= \max_{(v, \omega) \in E(P)} \left\{ \frac{1}{\pi(v)P(v, \omega)} \sum_{\Gamma(v, \omega)} |\gamma_{\sigma\beta}| \pi(\sigma) P'(\sigma, \beta) \right\} \\ &= \max_{(v, \omega) \in E(P)} \left\{ \frac{1}{\pi(\omega)P(\omega, v)} \sum_{\Gamma(v, \omega)} |\gamma_{\sigma\beta}| \pi(\sigma) P'(\sigma, \beta) \right\} \\ &\leq \max_{(v, \omega) \in E(P)} \sum_{\Gamma(v, \omega)} 2n \frac{P'(\sigma, \beta)}{P(\omega, v)} \\ &\leq \max_{(v, \omega) \in E(P)} \sum_{\Gamma(v, \omega)} 2n \frac{1/(2n)}{\frac{\lambda}{(1+\lambda)(n-1)}} = O(n^3). \end{aligned}$$

In either case, we have $A = O(n^3)$. Let $\lambda = \min_{i < j} p_{j,i}/p_{i,j}$. Then $\pi_* = \min_{\rho \in \Omega} \pi(\rho) \geq \lambda^{\binom{n}{2}}/n!$, so $\log(1/(\epsilon\pi_*)) = O(n^2 \log \epsilon^{-1})$, since each $p_{i,j}$ is bounded away from 1 implying that λ is a positive constant. Appealing to Theorem 4.4 we have that

$$\tau_{nn} = O(n^7 \log(n/\epsilon)).$$

□

5 League Hierarchies

We now introduce a second general class of input probabilities \mathbf{P} for which we show \mathcal{M}_{nn} is always rapidly mixing. Imagine a sporting franchise consisting of an A-league with stronger players and a B-league with weaker players. We assume that any player from the A-league has a fixed advantage over any player from the B-league, representing his or her probability of winning in a matchup. Within each of these leagues we have tier-1 and tier-2 players, where again a player from the stronger tier has a fixed probability of winning a competition against a tier-2 player. Likewise for the tiers in the other league, but of course the fixed advantage there can be different. This partition of each tier into stronger and weaker players continues recursively. To formalize the class of ‘‘League Hierarchies,’’ let T be a proper rooted binary tree with n leaf nodes, labeled $1, \dots, n$ in sorted order. Each non-leaf node v of this tree is labeled with a value $\frac{1}{2} \leq q_v \leq 1$. For $i, j \in [n]$, let $i \wedge j$ be the lowest common ancestor of the leaves labeled i and j . We say that \mathbf{P} has *league structure* T if $p_{i,j} = q_{i \wedge j}$. For example, Figure 4a shows a set \mathbf{P} such that $p_{1,4} = .8$, $p_{4,9} = .9$, and $p_{5,8} = .7$. We define matches by pairing up adjacent players in the current ranking and then we promote the winners, thus simulating \mathcal{M}_{nn} .

To show \mathcal{M}_{nn} is rapidly mixing for any input probabilities in the League Hierarchy class, we introduce a new combinatorial representation of each permutation that will be useful for the proofs. This representation

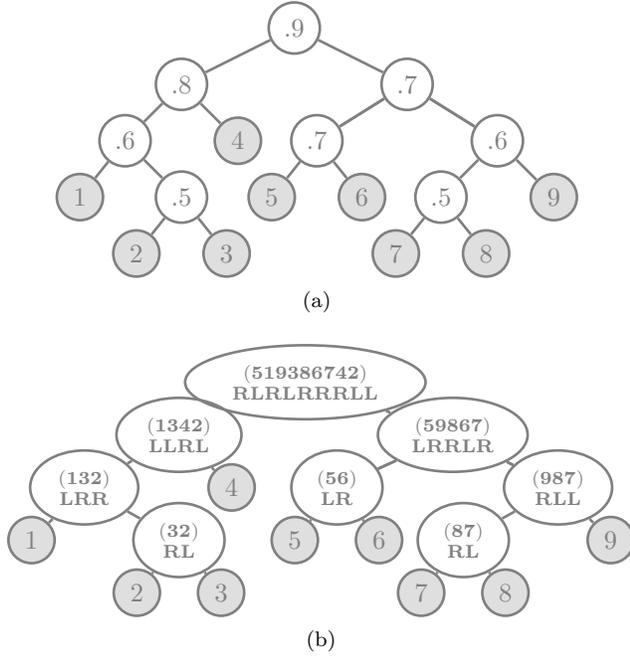


Figure 4: A set \mathbf{P} with league structure, and the corresponding tree-encoding of the permutation 519386742.

associates a bit string b_v to each node v of a binary tree with n leaves. Specifically, $b_v \in \{L, R\}^{\ell_v}$ where ℓ_v is the number of leaves in t_v , the subtree rooted at v , and for each element i of the sub-permutation corresponding to the leaves of t_v , $b_v(i)$ records whether i lies under the left or the right branch of v (see Figure 4b). The set of these bit strings is in bijection with the permutations. We consider a chain $\mathcal{M}_{tree}(T)$ that allows transpositions when they correspond to a nearest neighbor transposition in exactly one of the bit strings. Thus, the mixing time of $\mathcal{M}_{tree}(T)$ decomposes into a product of $n-1$ ASEP chains and we can conclude that the chain $\mathcal{M}_{tree}(T)$ is rapidly mixing using results in the constant bias case [1, 10]. Again, we use comparison techniques to conclude that \mathcal{M}_{nn} is also rapidly mixing when we have weak monotonicity, although we show that $\mathcal{M}_{tree}(T)$ is always rapidly mixing.

5.1 The Markov chain $\mathcal{M}_{tree}(T)$. We define the Markov chain $\mathcal{M}_{tree}(T)$ over permutations, given set \mathbf{P} with league structure T .

The Markov chain $\mathcal{M}_{tree}(T)$

Starting at any permutation σ_0 , repeat:

- Select distinct $a, b \in [n]$ with $a < b$ u.a.r.
- If every number between a and b in the permutation σ_t is not a descendant in T

of $a \wedge b$, obtain σ_{t+1} from σ_t by placing a, b in order with probability $p_{a,b}$, and out of order with probability $1 - p_{a,b}$, leaving all elements between them fixed.

- Otherwise, $\sigma_{t+1} = \sigma_t$.

First, we show that this Markov chain samples from the same distribution as \mathcal{M}_{nn} .

LEMMA 5.1. *The Markov chain $\mathcal{M}_{tree}(T)$ has the same stationary distribution as \mathcal{M}_{nn} .*

Proof. Let π be the stationary distribution of \mathcal{M}_{nn} , and let σ_1, σ_2 be a transition in $\mathcal{M}_{tree}(T)$. It suffices to show that the detailed balance condition holds for this transition with the stationary distribution π .

Recall that we may express $\pi(\sigma) = \prod_{i,j|i <_{\sigma} j} p_{i,j} / Z$ where $Z = \sum_{\sigma \in \Omega} \prod_{i,j|i <_{\sigma} j} p_{i,j}$. The transition σ_1, σ_2 transposes some two elements $a <_{\sigma_1} b$, where every element between a, b in σ_i is not a descendant of $a \wedge b$ in T . Let x_1, \dots, x_k be those elements. Thus, the path from a or b to x_i in T must pass through $a \wedge b$ and go to another part of the tree. For every such element x_i , $a \wedge x_i = (a \wedge b) \wedge x_i = b \wedge x_i$.

From the observation, we see from the league structure that $p_{a,x_i} = p_{b,x_i}$ for every x_i between a and b . Also, we see that either both $a < x_i, b < x_i$ or $a > x_i, b > x_i$, since all numbers between a, b are necessarily descendants of $a \wedge b$.

Therefore,

$$\frac{\pi(\sigma_1)}{\pi(\sigma_2)} = \frac{p_{a,b} \prod_i p_{a,x_i}}{p_{b,a} \prod_i p_{b,x_i}} = \frac{p_{a,b}}{p_{b,a}}.$$

This is exactly the ratio of the transition probabilities in $\mathcal{M}_{tree}(T)$, thus $\mathcal{M}_{tree}(T)$ also has stationary distribution π . \square

The key to the proof that $\mathcal{M}_{tree}(T)$ is rapidly mixing is again to decompose the chain into $n-1$ independent Markov chains, $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_{n-1}$, one for each non-leaf node of the tree T . We introduce an alternate representation of a permutation as a set of binary strings arranged like the tree T . We use the characters L and R for our binary representation instead of 0 and 1 for convenience. For each non-leaf node v in the tree T , let $L(v)$ be its left descendants, and $R(v)$ be its right descendants. We now do the following:

- For each non-leaf node v do the following:
 - List each descendant x of v in the order we encounter them in the permutation σ . These are parenthesized in Figure 4b.

- For each listed element x , write a L if $x \in L(v)$ and a R if $x \in R(v)$. This is the final encoding in Figure 4b.

We see that any σ will lead to an assignment of binary strings at each non-leaf node v with $L(v)$ L 's and $R(v)$ R 's. This is a bijection between the set of permutations and the set of assignments of binary strings to the tree T . Given any such assignment of binary strings, we can recursively reconstruct the permutation σ as follows:

- For each leaf node i , let its string be the string “ i ”.
- For any node n with binary string b ,
 - Determine the strings of its two children. Call these s_L, s_R .
 - Interleave the elements of s_L with s_L , choosing an element of s_L for each L in b , and an element of s_R for each R .

With this bijection, we first analyze $\mathcal{M}_{tree}(T)$'s behavior over tree representations and later extend this analysis to permutations. The Markov chain $\mathcal{M}_{tree}(T)$, when proposing a swap of the elements a and b , will only attempt to swap them if a, b correspond to some adjacent L and R in the string associated with $a \wedge b$. Swapping a and b does not affect any other string, so each non-leaf node v represents an independent exclusion process with $L(v)$ L 's and $R(v)$ R 's. These exclusion processes have been well-studied [3, 18, 1, 10]. We use the following bounds on the mixing times of the symmetric and asymmetric simple exclusion processes.

THEOREM 5.1. *Let \mathcal{M} be the exclusion process with parameter p on k_L L 's and k_R R 's, where $k = k_L + k_R$. Then*

1. if $p = 1/2$, $\tau(\epsilon) = O(k^3 \log(k_L k_R / \epsilon))$. [3, 18]
2. if $p > 1/2$, then $\tau(\epsilon) = O(k(\min\{k_L, k_R\} + \log k) \log(\epsilon^{-1})) = O(k^2 \log(\epsilon^{-1}))$. [10]

The bounds in Theorem 5.1 refer to the exclusion process which selects a position at random and swaps the two elements in that position with the appropriate probability. However, our process selects arbitrary pairs (i, j) consisting of a single L and a single R . Since we only swap (i, j) if they are neighboring, this may slow down the chain by a factor of at most k .

Since each exclusion process \mathcal{M}_i operates independently, the overall mixing time will be roughly n times the mixing time of each piece, slowed down by the inverse probability of selecting that process. Each \mathcal{M}_i has a different size, and a different mixing time relative to

its size. To employ the bounds from Theorem 5.1, we need the following theorem, which relates the mixing time of a product of independent Markov chains to the mixing time of each component. Similar results have been proved before in other settings [1, 2]. The proof is elementary, and is in the appendix.

THEOREM 5.2. *Suppose the Markov chain \mathcal{M} is a product of N independent Markov chains $\{\mathcal{M}_i\}$, where \mathcal{M} updates each \mathcal{M}_i with probability p_i , and $\sum_i p_i = 1$. If $\tau_i(\epsilon)$ is the mixing time for \mathcal{M}_i and $\tau_i(\epsilon) \geq 4 \ln \epsilon$ for each i , then*

$$\tau(\epsilon) \leq \max_{i=1,2,\dots,N} \frac{2}{p_i} \tau_i\left(\frac{\epsilon}{2N}\right).$$

Finally, we can prove that $\mathcal{M}_{tree}(T)$ is rapidly mixing.

THEOREM 5.3. *If \mathbf{P} has league structure T , then the mixing time of $\mathcal{M}_{tree}(T)$ under \mathbf{P} satisfies*

$$\tau(\epsilon) = O(n^5 \log(n/\epsilon)).$$

If \mathbf{P} is such that each $q_i > 1/2$ is a constant, then

$$\tau(\epsilon) = O(n^3 \log^2(n/\epsilon)).$$

Proof. In order to apply Theorem 5.2 to the Markov chain $\mathcal{M}_{tree}(T)$, we note that for a node with k_L L 's and k_R R 's ($k = k_L + k_R$), the probability of selecting that node is $\frac{k_L k_R}{\binom{n}{2}}$. Since $M = n - 1$, Theorem 5.2 implies

$$\tau(\epsilon) \leq \frac{n(n-1)}{k_L k_R} k^4 \ln(2nk_L k_R / \epsilon) = O(n^5 \log(n/\epsilon)).$$

Of course, if all of the chains have probabilities that are bounded away from $1/2$, then we can use the second bound from Theorem 5.1 to obtain

$$\begin{aligned} \tau(\epsilon) &\leq \frac{n(n-1)}{k_L k_R} k^2 (\min\{k_L, k_R\} + \log k) \log(2n/\epsilon) \\ &\leq \frac{n(n-1)k^2}{\max\{k_L, k_R\}} \left(1 + \frac{\log k}{\min\{k_L, k_R\}}\right) \log(2n/\epsilon). \end{aligned}$$

There are two cases to consider. Let $c > 0$.

If $\min\{k_L, k_R\} \geq c \log k$ then

$$\tau(\epsilon) \leq \frac{n(n-1)k^2}{k/2} (1+c) \log(2n/\epsilon) = O(n^3 \log(n/\epsilon)).$$

Otherwise, $\max\{k_L, k_R\} > k - c \log k$, so

$$\begin{aligned} \tau(\epsilon) &\leq \frac{n(n-1)k^2}{k - c \log k} (1 + \log k) \log(2n/\epsilon) \\ &= \frac{n(n-1)k}{1 - \frac{c \log k}{k}} (1 + \log k) \log(2n/\epsilon) \\ &= O(n^3 \log^2(n) \log(\epsilon^{-1})). \end{aligned}$$

□

5.2 Comparing $\mathcal{M}_{tree}(T)$ with \mathcal{M}_{nn} . Next, we show that \mathcal{M}_{nn} is rapidly mixing when \mathbf{P} has league structure and is *weakly monotone*:

DEFINITION 5.1. *The set \mathbf{P} is weakly monotone if properties 1 and either 2 or 3 are satisfied.*

1. $p_{i,j} \geq 1/2$ for all $1 \leq i < j \leq n$, and
2. $p_{i,j+1} \geq p_{i,j}$ for all $1 \leq i < j \leq n-1$ or
3. $p_{i-1,j} \geq p_{i,j}$ for all $2 \leq i < j \leq n$.

We note that if \mathbf{P} satisfies all three properties then it is *monotone*, as defined by Jim Fill [9].

The comparison proof in this setting is similar to the comparison proof in Section 4.3, except we allow elements between $\sigma(i)$ and $\sigma(j)$ that are larger or smaller than both i and j . This poses a problem, because we may not be able to move $\sigma(j)$ towards $\sigma(i)$ without greatly decreasing the weight. However, we can resolve this if \mathbf{P} is weakly monotone. Specifically, we prove the following theorem.

THEOREM 5.4. *If \mathbf{P} has league structure, is weakly monotone and every $p_{ij} \in \mathbf{P}$ is a constant less than 1, then the mixing time of \mathcal{M}_{nn} satisfies*

$$\tau_{nn}(\epsilon) = O(n^9 \log(n/\epsilon)).$$

Again, we are assuming the worst case bound on the mixing time of $\mathcal{M}_{tree}(T)$ given in Theorem 5.3, and if each $p_{i,j}$ is bounded away from 1/2 then we would get a better bound.

Proof. Throughout this proof we assume that \mathbf{P} satisfies properties 1 and 2 of the weakly monotone definition. If instead \mathbf{P} satisfies property 3, then the proof is very similar. In order to apply Theorem 4.4 to relate the mixing time of \mathcal{M}_{nn} to the mixing time of $\mathcal{M}_{tree}(T)$ we need to define for each transition of $\mathcal{M}_{tree}(T)$ a canonical path using transitions of \mathcal{M}_{nn} . Let $e = (\sigma, \beta)$ be a transition of $\mathcal{M}_{tree}(T)$ which performs a transposition on elements $\sigma(i)$ and $\sigma(j)$. If there are no elements between $\sigma(i)$ and $\sigma(j)$ then e is already a transition of \mathcal{M}_{nn} and we are done. Otherwise, σ contains the string $\sigma(i), \sigma(i+1), \dots, \sigma(j-1), \sigma(j)$ and β contains $\sigma(j), \sigma(i+1), \dots, \sigma(j-1), \sigma(i)$. From the definition of $\mathcal{M}_{tree}(T)$ we know that for each $\sigma(k)$, $k \in [i+1, j-1]$, either $\sigma(k) > \sigma(i), \sigma(j)$ or $\sigma(k) < \sigma(i), \sigma(j)$. Define $S = \{\sigma(k) : \sigma_k < \sigma(i), \sigma(j)\}$ and $B = \{\sigma(k) : \sigma_k > \sigma(i), \sigma(j)\}$. To obtain a good bound on the congestion along each edge we must ensure that the weight of the configurations on the path are not smaller than the weight of σ . Thus, we define three stages in our path from σ to β . In the first, we shift the elements of S to the left, removing an inversion with each element of B . In the second stage we move $\sigma(i)$ next to $\sigma(j)$ and in

the third stage we move $\sigma(j)$ to $\sigma(i)$'s original location. Finally, we shift the elements of S to the right to return them to their original locations. See Figure 5.

Stage 1: At a high-level in this stage we are shifting the elements in S to the left in order to remove an inversion with every element in B . First if $\sigma(j-1) \in B$, shift $\sigma(j)$ to the left until an element from S is immediately to the left of $\sigma(j)$. Next, starting at the right-most element in S and moving left, for each $\sigma(k) \in S$ such that $\sigma(k-1) \in B$, move $\sigma(k)$ to the left one swap at a time until $\sigma(k)$ has an element from S or $\sigma(i)$ on its immediate left (see Figure 6a). Notice that for each element $\sigma(l) \in B$ we have removed exactly one $(\sigma(l), \sigma(k))$ inversion where $\sigma(k) \in S$.

Stage 2: Next perform a series of nearest neighbor swaps to move $\sigma(i)$ to the right until it is in the original position occupied by $\sigma(j)$ in σ (see Figure 6b). While we have created an $(\sigma(k), \sigma(i))$ inversion for each element $\sigma(k) \in B$ the weight has not decreased from the original weight because in Stage 1 we removed an $(\sigma(k), \sigma(l))$ inversion (or an $(\sigma(k), \sigma(j))$ inversion) and $(\sigma(k), \sigma(l)) > (\sigma(k), \sigma(j))$ and $(\sigma(k), \sigma(j)) = (\sigma(k), \sigma(i))$ because the \mathbf{P} are weakly monotone. For each $\sigma(k) \in S$ we also removed a $(\sigma(k), \sigma(j))$ inversion.

Stage 3: Perform a series of nearest neighbor swaps to move $\sigma(j)$ to the left until it is in the same position $\sigma(i)$ was originally. While we created an $(\sigma(k), \sigma(j))$ inversion for each $\sigma(k) \in S$, these inversions have the same weight as the $(\sigma(i), \sigma(k))$ inversion we removed in Stage 2. In addition we have removed an $(\sigma(l), \sigma(j))$ inversion for each $\sigma(l) \in B$.

Stage 4: Finally we want to return the elements in S and B to their original position. Starting with the left-most element in S that was moved in Stage 1, perform the nearest neighbor swaps to the right necessary to return it to its original position. It's clear from the definition of the stages that the weight of a configuration never decreases below the weight of $\min(\pi(\sigma), \pi(\beta))$.

Given a transition (ν, ω) of \mathcal{M}_{nn} we must upper bound the number of canonical paths $\gamma_{\sigma\beta}$ that use this edge. Thus, we analyze the amount of information

5	8	9	<u>2</u>	10	<u>3</u>	<u>4</u>	<u>1</u>	7
5	<u>2</u>	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	7
<u>2</u>	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	5	7
<u>2</u>	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	7	5
7	<u>2</u>	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	5
7	8	9	<u>2</u>	10	<u>3</u>	<u>4</u>	<u>1</u>	5

Figure 5: The canonical path for transposing 5 and 7. Notice that the elements in S are underlined.

5	8	9	<u>2</u>	10	<u>3</u>	<u>4</u>	<u>1</u>	7
5	8	9	<u>2</u>	<u>3</u>	10	<u>4</u>	<u>1</u>	7
5	8	<u>2</u>	9	<u>3</u>	10	<u>4</u>	<u>1</u>	7
5	<u>2</u>	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	7

(a)

5	<u>2</u>	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	7
<u>2</u>	5	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	7
<u>2</u>	8	5	9	<u>3</u>	10	<u>4</u>	<u>1</u>	7
				\vdots				
<u>2</u>	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	5	7
<u>2</u>	8	9	<u>3</u>	10	<u>4</u>	<u>1</u>	7	5

(b)

Figure 6: Stages 1 and 2 of the canonical path for transposing 5 and 7.

needed in addition to (z, w) to determine σ and β uniquely. First we record whether (σ, β) is already a nearest neighbor transition or which stage we are in. Next for any of the 4 stages we record the original location of $\sigma(i)$ and $\sigma(j)$. Given this information, along with v and ω , we can uniquely recover (σ, β) . Hence, there are at most $4n^2$ paths through any edge (v, ω) . Also, note that the maximum length of any path is $4n$.

Next we bound the quantity A which is needed to apply Theorem 4.4. Recall that we have guaranteed that $\pi(\sigma) \leq \max\{\pi(v), \pi(\omega)\}$. Assume that $\pi(\sigma) \leq \pi(v)$. Let $\lambda = \min_{i < j} p_{j,i}/p_{i,j}$. Then

$$\begin{aligned}
A &= \max_{(v, \omega) \in E(P)} \left\{ \frac{1}{\pi(v)P(v, \omega)} \sum_{\Gamma(v, \omega)} |\gamma_{\sigma\beta}| \pi(\sigma) P'(\sigma, \beta) \right\} \\
&\leq \max_{(v, \omega) \in E(P)} \sum_{\Gamma(v, \omega)} 2n \frac{P'(\sigma, \beta)}{P(v, \omega)} \\
&\leq \max_{(v, \omega) \in E(P)} \sum_{\Gamma(v, \omega)} 2n \frac{1/\binom{n}{2}}{\lambda^{(1+\lambda)(n-1)}} = O(n^2).
\end{aligned}$$

If, on the other hand, $\pi(\sigma) \leq \pi(\omega)$, then we use detailed

balance to obtain:

$$\begin{aligned}
A &= \max_{(v, \omega) \in E(P)} \left\{ \frac{1}{\pi(v)P(v, \omega)} \sum_{\Gamma(v, \omega)} |\gamma_{\sigma\beta}| \pi(\sigma) P'(\sigma, \beta) \right\} \\
&= \max_{(v, \omega) \in E(P)} \left\{ \frac{1}{\pi(\omega)P(\omega, v)} \sum_{\Gamma(v, \omega)} |\gamma_{\sigma\beta}| \pi(\sigma) P'(\sigma, \beta) \right\} \\
&\leq \max_{(v, \omega) \in E(P)} \sum_{\Gamma(v, \omega)} 2n \frac{P'(\sigma, \beta)}{P(\omega, v)} \\
&\leq \max_{(v, \omega) \in E(P)} \sum_{\Gamma(v, \omega)} 2n \frac{1/\binom{n}{2}}{\lambda^{(1+\lambda)(n-1)}} = O(n^2).
\end{aligned}$$

In either case, we have $A = O(n^2)$. Then $\pi_* = \min_{\rho \in \Omega} \pi(\rho) \geq \lambda^{\binom{n}{2}}/n!$, so $\log(1/(\epsilon\pi_*)) = O(n^2 \log \epsilon^{-1})$, as above. Applying Theorem 4.4 proves that $\tau_{nn} = O(n^9 \log(n/\epsilon))$. \square

REMARK 5.1. *By repeating Stage 1 of the path a constant number of times, it is possible to relax the weakly monotone condition slightly if we are satisfied with a polynomial bound on the mixing time.*

6 Conclusions

In this paper, we introduced new classes of positively biased probability distributions \mathbf{P} for which the nearest neighbor transposition chain is provably rapidly mixing, the choose your weapon and league hierarchy classes. Both classes represent linear families of input parameters defining \mathbf{P} , greatly generalizing the single parameter constant bias case. The only case in which we know the chain is rapidly mixing for a quadratic class is when all of the $p_{i,j}$ are 0 or 1 and the problem reduces to linear extensions of a partial order. It would be interesting to bound the mixing rate for other quadratic families.

It is also worth mentioning that our counterexample from Section 3 showing that there exist positively biased distributions for which the chain mixes slowly does not satisfy the monotonicity condition in Fill's conjecture, so this conjecture certainly is worthy of further consideration. Moreover, the new classes where we can show the chain always converges quickly do not necessarily satisfy monotonicity, so there may be another condition that characterizes a larger class of distributions \mathbf{P} for which the chain always converges quickly.

Acknowledgments. We thank Jim Fill for sharing his earlier work on this problem and for several useful conversations.

References

- [1] I. Benjamini, N. Berger, C. Hoffman, and E. Mossel. Mixing times of the biased card shuffling and the asymmetric exclusion process. *Trans. Amer. Math. Soc.*, 2005.
- [2] N. Bhatnagar and D. Randall. Torpid mixing of simulated tempering on the potts model. In *Proceedings of the 15th ACM/SIAM Symposium on Discrete Algorithms*, SODA '04, pages 478–487, 2004.
- [3] R. Bubley and M. Dyer. Faster random generation of linear extensions. In *Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, SODA '98, 1998.
- [4] P. Diaconis and L. Saloff-Coste. Comparison techniques for random walks on finite groups. *The Annals of Applied Probability*, 21:2131–2156, 1993.
- [5] P. Diaconis and L. Saloff-Coste. Comparison theorems for reversible markov chains. *The Annals of Applied Probability*, 3:696–730, 1993.
- [6] P. Diaconis and M. Shahshahani. Generating a random permutation with random transpositions. *Probability Theory and Related Fields*, 57:159–179, 1981.
- [7] M. Dyer and C. Greenhill. A more rapidly mixing markov chain for graph colorings. *Random Structures & Algorithms*, 13:285–317, 1998.
- [8] J. Fill. Background on the gap problem. *Unpublished manuscript*, 2003.
- [9] J. Fill. An interesting spectral gap problem. *Unpublished manuscript*, 2003.
- [10] S. Greenberg, A. Pascoe, and D. Randall. Sampling biased lattice configurations using exponential metrics. In *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, 2009.
- [11] M. Jerrum and A. Sinclair. Approximate counting, uniform generation and rapidly mixing markov chains. *Information and Computation*, 82:93–133, 1989.
- [12] D. Knuth. *The Art of Computer Programming*, volume 3: Sorting and Searching. Addison Wesley, 1973.
- [13] D. Levin, Y. Peres, and E. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.
- [14] A. Pascoe and D. Randall. Self-assembly and convergence rates of heterogenous reversible growth processes. In *Foundations of Nanoscience*, 2009.
- [15] D. Randall and P. Tetali. Analyzing glauber dynamics by comparison of Markov chains. *Journal of Mathematical Physics*, 41:1598–1615, 2000.
- [16] A. Sinclair. *Algorithms for random generation and counting*. Progress in theoretical computer science. Birkhäuser, 1993.
- [17] S. Turrini. Optimization in permutation spaces. *Western Research Laboratory Research Report*, 1996.
- [18] D. Wilson. Mixing times of lozenge tiling and card shuffling markov chains. *The Annals of Applied Probability*, 1:274–325, 2004.

A Appendix: Bounding the Mixing Time of a Product of Markov Chains (Theorem 5.2)

In Sections 4 and 5 we use combinatorial bijections to express the mixing time of \mathcal{M}_{nn} as a product of independent, smaller Markov chains. Our bounds on the mixing rate rely on relating the mixing time of the larger and smaller chains. While there exist many results relating the mixing time of a product of Markov chains (see, for example, references in [1, 2]), these assume that the smaller chains defining the product are of comparable size. These theorems would yield weaker results in our case where the smaller Markov chains can be of vastly different size, so we include a proof of the more tailored theorem here for completeness.

We now prove Theorem 5.2 from Section 5, which states that if the Markov chain \mathcal{M} is a product of M independent Markov chains $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_M$, each with mixing time $\tau_i(\epsilon)$, and \mathcal{M} updates \mathcal{M}_i with probability p_i , then the mixing time of \mathcal{M} is

$$\tau(\epsilon) \leq \max_{i=1,2,\dots,M} \max \left\{ \frac{2}{p_i} \tau_i \left(\frac{\epsilon}{4M} \right), \frac{8}{p_i} \ln \left(\frac{\epsilon}{8M} \right) \right\}.$$

In particular, if each $\tau_i(\epsilon) \geq 4 \ln(\epsilon)$ then

$$\tau(\epsilon) \leq \max_{i=1,2,\dots,M} \frac{2}{p_i} \tau_i \left(\frac{\epsilon}{4M} \right).$$

Proof. Suppose the Markov chain \mathcal{M} has transition matrix P , and each \mathcal{M}_i has transition matrix P_i and state space Ω_i . Let $B_i = p_i P_i + (1 - p_i)I$, where I is the identity matrix of the same size as P_i , be the transition matrix of \mathcal{M}_i , slowed down by the probability p_i of selecting \mathcal{M}_i . First we show that the total variation distance satisfies

$$1 + 2d_{tv}(P^t, \pi) \leq \prod_i (1 + 2d_{tv}(B_i^t, \pi_i)).$$

To show this, notice that for $x = (x_1, x_2, \dots, x_M), y = (y_1, y_2, \dots, y_M) \in \Omega$, $P^t(x, y) = \prod_i B_i^t(x_i, y_i)$. Let $\epsilon_i(x_i, y_i) = B_i^t(x_i, y_i) - \pi_i(y_i)$ and for any $x_i \in \Omega_i$,

$$\epsilon_i(x_i) = \sum_{y_i \in \Omega_i} |\epsilon_i(x_i, y_i)| \leq 2d_{tv}(B_i^t, \pi_i).$$

Then,

$$\begin{aligned}
d_{tv}(P^t, \pi) &= \max_{x \in \Omega} \frac{1}{2} \sum_{y \in \Omega} |P^t(x, y) - \pi(y)| \\
&= \max_{x \in \Omega} \frac{1}{2} \sum_{y \in \Omega} \left| \prod_i B_i^t(x_i, y_i) - \prod_i \pi_i(y_i) \right| \\
&= \max_{x \in \Omega} \frac{1}{2} \sum_{y \in \Omega} \left| \prod_i (\epsilon_i(x_i, y_i) + \pi_i(y_i)) - \prod_i \pi_i(y_i) \right| \\
&= \max_{x \in \Omega} \frac{1}{2} \sum_{y \in \Omega} \left| \sum_{S \subseteq [M], S \neq \emptyset} \prod_{i \in S} \epsilon_i(x_i, y_i) \prod_{i \notin S} \pi_i(y_i) \right| \\
&\leq \max_{x \in \Omega} \frac{1}{2} \sum_{y \in \Omega} \sum_{S \subseteq [M], S \neq \emptyset} \prod_{i \in S} |\epsilon_i(x_i, y_i)| \prod_{i \notin S} |\pi_i(y_i)| \\
&= \max_{x \in \Omega} \frac{1}{2} \sum_{S \subseteq [M], S \neq \emptyset} \prod_{i \in S} \sum_{y_i \in \Omega_i} |\epsilon_i(x_i, y_i)| \prod_{i \notin S} \sum_{y_i \in \Omega_i} |\pi_i(y_i)| d_{tv}(B_i^t, \pi_i) \\
&= \max_{x \in \Omega} \frac{1}{2} \sum_{S \subseteq [M], S \neq \emptyset} \prod_{i \in S} \epsilon_i(x_i) \prod_{i \notin S} 1 \\
&= \max_{x \in \Omega} \frac{1}{2} \prod_i (1 + \epsilon_i(x_i)) - 1/2 \\
&\leq \frac{1}{2} \prod_i (1 + 2d_{tv}(B_i^t, \pi_i)) - 1/2,
\end{aligned}$$

as desired. Thus to show $d_{tv}(P^t, \pi) \leq \epsilon$, it suffices to show $d_{tv}(B_i^t, \pi_i) \leq \epsilon/(2M)$ for each i , as

$$\begin{aligned}
1 + 2d_{tv}(P^t, \pi) &\leq \prod_i (1 + 2d_{tv}(B_i^t, \pi_i)) \\
&\leq \prod_i (1 + 2\epsilon/(2M)) \\
&\leq e^\epsilon \leq 1 + 2\epsilon.
\end{aligned}$$

Hence it suffices to show $d_{tv}(B_i^t, \pi_i) \leq \epsilon/(2M)$ for each i .

Let $q_i = 1 - p_i$. Since

$$B_i^t = (p_i P_i + q_i I)^t = \sum_{j=0}^t \binom{t}{j} p_i^j q_i^{t-j} P_i^j I,$$

we have

$$\begin{aligned}
d_{tv}(B_i^t, \pi_i) &= \max_{x_i \in \Omega_i} \frac{1}{2} \sum_{y_i \in \Omega_i} |B_i^t(x_i, y_i) - \pi_i(y_i)| \\
&= \max_{x_i \in \Omega_i} \frac{1}{2} \sum_{y_i \in \Omega_i} \left| \sum_{j=0}^t \binom{t}{j} p_i^j q_i^{t-j} P_i^j(x_i, y_i) - \pi_i(y_i) \right|
\end{aligned}$$

$$\begin{aligned}
&= \max_{x_i \in \Omega_i} \frac{1}{2} \sum_{y_i \in \Omega_i} \left| \sum_{j=0}^t \binom{t}{j} p_i^j q_i^{t-j} (P_i^j(x_i, y_i) - \pi_i(y_i)) \right| \\
&\leq \max_{x_i \in \Omega_i} \frac{1}{2} \sum_{y_i \in \Omega_i} \sum_{j=0}^t \binom{t}{j} p_i^j q_i^{t-j} |P_i^j(x_i, y_i) - \pi_i(y_i)| \\
&= \sum_{j=0}^t \binom{t}{j} p_i^j q_i^{t-j} \max_{x_i \in \Omega_i} \frac{1}{2} \sum_{y_i \in \Omega_i} |P_i^j(x_i, y_i) - \pi_i(y_i)| \\
&= \sum_{j=0}^t \binom{t}{j} p_i^j q_i^{t-j} d_{tv}(P_i^j, \pi_i).
\end{aligned}$$

Let $t_i = \tau_i(\epsilon/(4M))$. Now, for $j \geq t_i = \tau_i(\epsilon/(4M))$, we have that $d_{tv}(P_i^j, \pi_i) < \epsilon/(4M)$. For all j , we have $d_{tv}(P_i^j, \pi_i) \leq 2$, so if X is a binomial random variable with parameters t and p_i with $q_i = 1 - p_i$, we have

$$\begin{aligned}
&\leq \sum_{j=0}^t \binom{t}{j} p_i^j q_i^{t-j} d_{tv}(P_i^j, \pi_i) \\
&= \sum_{j=0}^{t_i-1} \binom{t}{j} p_i^j q_i^{t-j} d_{tv}(P_i^j, \pi_i) + \sum_{j=t_i}^t \binom{t}{j} p_i^j q_i^{t-j} d_{tv}(P_i^j, \pi_i) \\
&< 2 \sum_{j=0}^{t_i-1} \binom{t}{j} p_i^j q_i^{t-j} + \sum_{j=t_i}^t \binom{t}{j} p_i^j q_i^{t-j} \epsilon/(2M) \\
&= 2P(X < t_i) + \epsilon/(2M).
\end{aligned}$$

By Chernoff bounds, $P(X < (1 - \delta)tp_i) \leq e^{-tp_i\delta^2/2}$. Setting $\delta = 1 - t_i/(tp_i)$, then for all $t > 2t_i/p_i$, $\delta^2 \geq 1/4$ and we have

$$P(X < t_i) \leq e^{-tp_i\delta^2/2} \leq e^{-tp_i/8} \leq \epsilon/(8M)$$

as long as $t \geq 8 \ln(\epsilon/(8M))/p_i$. Therefore for $t \geq \max\{8 \ln(\epsilon/(8M))/p_i, 2t_i/p_i\}$,

$$\begin{aligned}
d_{tv}(B_i^t, \pi_i) &= 2P(X < t_i) + \epsilon/(4M) \\
&\leq 2\epsilon/(8M) + \epsilon/(4M) = \epsilon/(2M).
\end{aligned}$$

Hence by time t the total variation distance satisfies $d_{tv}(P^t, \pi) \leq \epsilon$. \square