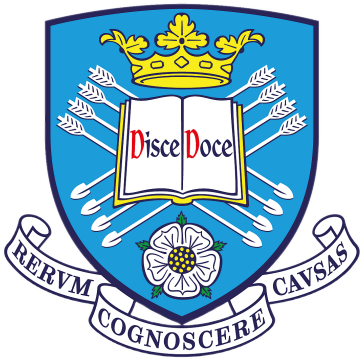
Markov Chains, Mixing Times and Shuffling Cards



Mikey Guinness

May 2026

A dissertation submitted in partial fulfilment of the requirements for the degree of Master of Mathematics

School of Mathematical and Physical Sciences

The University of Sheffield

Preface

The motivation of this dissertation is to explore how Markov chains and mixing times are implemented in the context of shuffling playing cards. This paper outlines results and methods which build to give upper and lower bounds for shuffling, before moving on to analysis of the cut-off phenomenon which occurs between these bounds, and finally finishing with theoretical and computer modelling of more complex spatial card mixing.

1. Markov Chains

1.1. **Basic Properties.** Although it is expected the reader of this paper will understand the basic properties of Markov chains, it will be useful to reinforce some of those properties which are integral to this paper, which will be done in this section.
Markov chains are stochastic processes that evolve over time based on a principle that the future only depends on the present state, independently of past history. More formally, let $(X_t)t \geq 0$ be a sequence of random variables taking values in a finite set Ω, called the state space. The process is called a (discrete-time) Markov chain if, for all times $t \geq 0$ and all states $x_0, x_1, \ldots, x_t, y \in \Omega$ with positive probability,

$$\mathbb{P}(Xt+1 = y \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{t+1} = y \mid X_t = x_t).$$

This is the Markov property of a Markov chain. In the applications for this paper only time-homogeneous Markov chains will be considered. In other words, the transition probabilities are independent of time.
Since the state space Ω is finite, the transition probabilities can be encoded in a matrix

$$P = (P(x, y))x, y \in \Omega,$$

called the transition matrix. Each entry satisfies $P(x, y) \geq 0$, and for each state x

$$\sum y \in \Omega P(x, y) = 1,$$

so each row of P is a probability distribution. This matrix P completely determines the chain's evolution. More importantly, moving from state x to state y in t steps is given by the (x, y)-entry of the matrix power P^t, denoted

$$P^t(x, y) = \mathbb{P}(X_t = y \mid X_0 = x).$$

This is a powerful result, as matrix multiplication can be used to calculate transitions over multiple steps.

In terms of card shuffling with a deck of n cards, each possible way of ordering the cards is a permutation, so the state space of the deck is the symmetric group S_n, if each card is represented using the integers 1 through n. A shuffling procedure is a Markov chain defined by specifying how one permutation moves to the next at each step.

Example 1.1. *Consider a simple Markov chain with state space* $\Omega = \{1, 2, 3\}$ *and transition matrix*

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \end{pmatrix}.$$

If the chain starts at state 1*, then at the next step it must move to state* 2*. From state* 2*, it moves to either* 1 *or* 3 *with equal probability.*

This example depicts how the transition matrix determines the evolution of the chain. After two steps, starting from 1*, the chain may return to* 1 *or move to* 3*, each with probability* 1/2*.*

1.2. **Stationary Distribution.** Another key concept needed for this paper is stationary distributions. A probability distribution π on Ω is said to be stationary if it satisfies

$$\pi = \pi P,$$

that is,

$$\pi(y) = \sum_{x \in \Omega} \pi(x) P(x, y) \quad \text{for all } y \in \Omega.$$

If the chain starts with distribution π, or reaches π, it will remain at that distribution at all future times. In the case of card shuffling, it is natural to ask whether the stationary distribution is uniform on S_n. This is true for many standard shuffling procedures, but not for all possible Markov chains on S_n.

Now a general result which explains why the uniform distribution arises in many natural examples.

Proposition 1.2. *Let G be a finite group, and let μ be a probability distribution on G. Consider the Markov chain defined by*

$$X_{t+1} = Y_{t+1} X_t,$$

where $Y_1, Y_2, \ldots$ are independent random variables with common distribution μ. Then the uniform distribution on G is stationary.

Proof. Let π denote the uniform distribution on G, so that $\pi(g) = 1/|G|$ for all $g \in G$. The transition matrix is

$$P(g, h) = \mathbb{P}(Yg = h) = \mu(hg^{-1}).$$

Then

$$(\pi P)(h) = \sum_{g \in G} \pi(g) P(g, h) = \frac{1}{|G|} \sum_{g \in G} \mu(hg^{-1}).$$

As g ranges over G, the element hg^{-1} also ranges over all of G. Hence

$$\sum_{g \in G} \mu(hg^{-1}) = \sum_{z \in G} \mu(z) = 1,$$

and so

$$(\pi P)(h) = \frac{1}{|G|} = \pi(h).$$

□

For card shuffling, the group is S_n, the symmetric group. Many natural shuffles, including riffle shuffles, can be viewed as random walks on S_n of this form. For such shuffles, the stationary distribution is therefore uniform.

1.3. **Returning to Shuffling.** Returning to the motivation of the paper, the example of card shuffling. Let S_n denote the set of all permutations of n cards. A shuffling procedure defines a Markov chain (X_t) on S_n, where X_t represents the ordering of the deck after t shuffles.

The Markov property holds because the outcome of the next shuffle depends only on the deck's current order. As mentioned prior, the stationary distribution for this chain is the uniform distribution on S_n.

The next key question is then, how many steps are required for the distribution of X_t to be "close" to uniform?

2. Total Variation Distance And Mixing Times

2.1. **Total Variation Distance.** If the desire is to explore and measure the rate at which a Markov chain approaches its stationary distribution, a way to measure the distance between probability distributions is needed. The standard metric used in this context is the total variation distance.

Let μ and ν be two probability distributions on a finite state space Ω. The total variation distance between μ and ν is defined by

$$|\mu - \nu|\text{TV} = \frac{1}{2}\sum x \in \Omega|\mu(x) - \nu(x)|.$$

An equivalent and often more intuitive formula is

$$|\mu - \nu|\text{TV} = \max A \subseteq \Omega|\mu(A) - \nu(A)|.$$

This formula shows that total variation distance measures the largest possible difference in probability that the two distributions assign to the same event. Being specific, it quantifies how well one can distinguish between μ and ν based on a single observation.

The total variation distance is always between 0 and 1. It is equal to 0 if and only if $\mu = \nu$, and it is equal to 1 when the distributions are completely disjoint.

Example 2.1. *Let* $\Omega = \{1, 2, 3, 4\}$*, and define*

$$\mu = (1, 0, 0, 0), \quad \nu = (0, 1, 0, 0).$$

Then

$$\|\mu - \nu\|_{\text{TV}} = 1,$$

so the distributions are completely distinguishable.

Example 2.2. *Let*

$$\mu = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right), \quad \nu = \left(\frac{3}{10}, \frac{1}{5}, \frac{1}{4}, \frac{1}{4}\right).$$

Then

$$\|\mu - \nu\|_{\text{TV}} = \frac{1}{20},$$

so the distributions are very close.

2.2. **Stationarity.** Recall that a probability distribution π on the state space Ω is called stationary if it satisfies

$$\pi = \pi P.$$

The term stationarity means the chain has reached its stationary distribution. It is important to distinguish between being close to stationarity and being exactly in stationarity. Most of the time, the aim is to show that the distribution of the chain is close to π in total variation distance. However, strong stationary times provide a method for identifying a random time at which the chain is exactly in stationarity.

Example 2.3. *Consider the Markov chain on $\Omega = \{0, 1\}$ with transition matrix*

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix},$$

where $0 < \alpha < 1/2$.

This chain remains in its current state with probability $1 - \alpha$ and moves to the other state with probability α. The distribution

$$\pi = \left(\frac{1}{2}, \frac{1}{2}\right)$$

is stationary, since

$$\pi P = \left(\frac{1}{2}, \frac{1}{2}\right)\begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix} = \left(\frac{1}{2}, \frac{1}{2}\right) = \pi.$$

Thus, if the chain starts with distribution π, it remains in this distribution after every step. This example will be continued below when computing the mixing time explicitly.

2.3. **Distance from Stationarity.** Let (X_t) be a Markov chain with transition matrix P and stationary distribution π. For a fixed starting state $x \in \Omega$, define the distance to stationarity at time t by

$$d_x(t) = |P^t(x, \cdot) - \pi|_{\mathrm{TV}},$$

where $P^t(x, \cdot)$ denotes the distribution of the chain at time t given that $X_0 = x$.

Since the convergence rate may depend on the starting state, it makes sense to consider the worst case scenario.

$$d(t) = \max_{x \in \Omega} d_x(t).$$

The function $d(t)$ measures the furthest the chain can be from stationarity after t steps, regardless of the initial permutation or distribution.

2.4. **Mixing Times.** Now, moving on to a main quantity of interest.

For $\varepsilon > 0$, the mixing time of the chain is defined by

$$t_{\mathrm{mix}}(\varepsilon) = \min\{t \geq 0 : d(t) \leq \varepsilon\}.$$

Simply put, $t_{\mathrm{mix}}(\varepsilon)$ is the smallest time at which the distribution of the chain is within ε of the stationary distribution, uniformly over all starting states.

In many applications, one fixes a particular value of ε for comparisons etc. For most card shuffling usages, the consensus is to select $\varepsilon = 1/4$. The precise choice of ε is usually not important, as changing it only affects the mixing times by a constant factor, but a value of $1/4$ is standard and works well.

Following, the aforementioned distance function $d(t)$ satisfies several useful properties.

First, $d(t)$ is non-increasing in t. This reflects the fact that the Markov chain becomes progressively closer to stationarity over time.

Second, the distance function satisfies the sub-multiplicative property:

$$d(s+t) \leq 2d(s)d(t).$$

This inequality plays an important role in bounding mixing times, as it allows control over long-term behaviour by understanding shorter time scales.

These properties ensure that convergence to stationarity occurs in a regular and predictable manner.

Example 2.4. *Continue the two-state chain from Example 2.3:*

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{pmatrix}, \qquad 0 < \alpha < 1/2.$$

The stationary distribution is

$$\pi = \left(\frac{1}{2}, \frac{1}{2}\right).$$

Starting from state 0*, one can compute*

$$P^t(0,\cdot) = \left(\frac{1}{2} + \frac{1}{2}(1-2\alpha)^t, \frac{1}{2} - \frac{1}{2}(1-2\alpha)^t\right).$$

Therefore

$$\|P^t(0,\cdot) - \pi\|_{\mathrm{TV}} = \frac{1}{2}(1-2\alpha)^t.$$

By symmetry, the same expression is obtained when starting from state 1*, so*

$$d(t) = \frac{1}{2}(1-2\alpha)^t.$$

Hence the mixing time is the smallest t *such that*

$$\frac{1}{2}(1-2\alpha)^t \leq \varepsilon.$$

Equivalently,

$$t_{\mathrm{mix}}(\varepsilon) = \left\lceil \frac{\log(1/(2\varepsilon))}{-\log(1-2\alpha)} \right\rceil.$$

In particular, for the standard choice $\varepsilon = 1/4$*,*

$$t_{\mathrm{mix}} = \left\lceil \frac{\log 2}{-\log(1-2\alpha)} \right\rceil.$$

This example shows explicitly that if α *is small, then the chain mixes slowly. Indeed, using*

$$-\log(1-2\alpha) \sim 2\alpha$$

for small α, one obtains

$$t_{\text{mix}} \sim \frac{\log 2}{2\alpha}.$$

2.5. **Motivation for Mixing Times.** Why outline all this? In context with card shuffling, the total variation distance provides a fixed way to measure the "randomness" per se of a deck. With the stationary distribution being a perfectly shuffled deck, and with $d(t)$ showing how close to this the deck is, then the mixing time can be interpreted as the number of shuffles for the deck to be "shuffled" adequately. This now gives the next aim, to provide bounds for mixing times for specific shuffles.

3. The Top-to-Random Shuffle

Before studying other shuffles, such as the riffle shuffle, it is useful to consider a very simple shuffling Markov chain, the top-to-random shuffle. This example is simpler than the riffle shuffle, and it will later on illustrate the importance of a concept called *strong stationary times*.

3.1. **Definition.** The *top-to-random shuffle* [5] is defined as follows. At each step, remove the top card of the deck and insert it uniformly at random into one of the n possible positions.

Equivalently, if the deck is in some order $\pi \in S_n$, the next state is obtained by removing the first card of π and reinserting it into a uniformly chosen position. This gives a Markov chain on S_n.

3.2. **Basic Properties.** The chain is irreducible. Indeed, by repeatedly moving top cards into suitable positions, it is possible to build any desired ordering of the deck.

The chain is also aperiodic, since there is a positive probability that the top card is reinserted into the top position, in which case the deck does not change.

The stationary distribution is the uniform distribution on S_n. This is consistent with the general group argument given earlier that natural shuffling chains on S_n usually preserve the uniform distribution.

The top-to-random shuffle provides an example of a Markov chain which mixes relatively slowly. At each step, only one card is moved, so randomness spreads gradually through the deck.

In particular, before the bottom card can reach the top, every card above it must be removed at least once. This suggests a connection with the coupon collector problem, and leads to a mixing time of order $n \log n$.

3.3. **Coupon Collector Problem.** The connection with the coupon collector problem [3] helps explain why the top-to-random shuffle has an $n \log n$ time scale.

In the coupon collector problem, there are n different coupon types, and each new coupon is chosen independently and uniformly from the set of all n types. Let τ be the first time at which all coupon types have appeared. If τ_k denotes the first time at which k distinct coupon types have been collected, then

$$\tau = \tau_1 + (\tau_2 - \tau_1) + \cdots + (\tau_n - \tau_{n-1}).$$

After $k-1$ different coupon types have already been collected, the probability that the next coupon is new is

$$\frac{n-k+1}{n}.$$

Thus $\tau_k - \tau_{k-1}$ is a geometric random variable with mean

$$\frac{n}{n-k+1}.$$

By linearity of expectation,

$$\mathbb{E}[\tau] = \sum_{k=1}^{n} \frac{n}{n-k+1} = n \sum_{j=1}^{n} \frac{1}{j} \sim n \log n.$$

This idea is useful for the top-to-random shuffle because a card near the bottom of the deck cannot be moved until the cards above it have first been removed from the top. In this sense, the shuffle must gradually affect all parts of the deck before the whole deck can be well mixed.

This analogy should not be interpreted as a proof of the top-to-random mixing time, but more to give intuition as to why the logarithmic factor appears.

4. Riffle Shuffle Model

4.1. **Layman's Terms of Riffling.** Our main shuffle the paper will be exploring is the riffle shuffle, arguably the most common style of shuffling a deck of cards. For now, consider a 2-shuffle, which is performed by splitting the deck into two roughly equal piles and then interleaving the two piles together in an alternating pattern to form one deck again.
This is not precise enough of a definition for a mathematical model. A proper model will be introduced, which also accounts for the two real acts of randomness of a riffle. The first is where the deck is cut into two, as this will not always be exactly halfway, but on average will be. The second is in what order the cards are interleaved. On average, this will be one card from each pile in alternating order, but in practice sometimes multiple cards will drop from one pile at a time.

4.2. **The Gilbert-Shannon-Reeds Model.** The main mathematical model for the riffle shuffle is the Gilbert–Shannon–Reeds (GSR) model. This model, discussed in many papers [1, 5], and based on work from Gilbert [4], describes a riffle shuffle in two stages.

First, the deck of n cards is cut into two piles, with the number of cards in the first pile (and thus also the second pile) being chosen according to a binomial distribution:

$$\mathbb{P}(\text{cut at } k) = \frac{\binom{n}{k}}{2^n}, \quad k = 0, 1, \ldots, n.$$

Second, the two piles are interleaved to form a single deck. At each step of the interleaving process, the next card is chosen from one of the two piles, with probability proportional to the number of cards remaining in each pile. That is, if the left pile contains a cards and the right pile contains b cards, then the next card is drawn from the left pile with probability $a/(a+b)$ and from the right pile with probability $b/(a+b)$.

This procedure defines a probability distribution on the set of all permutations of the deck and therefore determines a transition rule for a Markov chain in S_n. Each application of the shuffle corresponds to one step of the chain. This model also passes an intuitive test. It's visible to see that the deck will tend to be cut near the middle, and it feels correct that one would be more likely to drop a card from the pile with more cards in it.

4.3. **The Inverse Model.** The inverse shuffle representation provides a more efficient way to analyse repeated riffle shuffles. It is called the inverse model because it describes the operation obtained by running a riffle shuffle backwards.

In a regular riffle shuffle, the deck is cut into two packets and then those two packets are interleaved. If one is given the final shuffled deck and wants to undo this operation, one can mark each card according to which packet it came from. Pulling all cards from the first packet back together, while preserving their relative order, and doing the same for the second packet, reverses the interleaving.

This reverse operation is exactly what the inverse model represents. Each card is assigned a label, either 0 or 1, indicating which packet it is pulled back into. Then all cards labelled 0 are moved above all cards labelled 1, while preserving the relative order among cards with the same label. This inverse description of riffle shuffling is used both in the standard Markov chain treatment [5] and in the original Bayer-Diaconis analysis.[1]

This process can be extended inductively. After one shuffle, each card has a label of length 1. After a second shuffle, each card receives a second independent bit. Within each group of cards sharing the first bit, reorder according to the second bit. Continuing in this way, after k shuffles each card is assigned a binary string in $\{0,1\}^k$.

The deck is then ordered lexicographically according to these labels, while preserving the original relative order of cards with identical labels.

The position of the cards after k shuffles is determined by sorting them according to these binary strings in lexicographic order, while preserving the original relative order among cards with identical labels.

Thus, after k shuffles, the permutation of the deck is entirely determined by the random labels assigned to the cards. This representation provides a useful way to understand when the deck becomes well mixed. If all cards receive distinct binary strings, then their relative order is determined by independent random labels, and the resulting permutation is uniformly distributed over S_n. To use this however, it is required to prove this gives the same distribution as the GSR model.

The following example illustrates this construction.

4.4. **Equivalence.** In the inverse model as mentioned, assign to each card an independent bit / value in $\{0,1\}$. Let k be the number of cards assigned the value 0. Then k follows a binomial distribution:

$$\mathbb{P}(k) = \frac{\binom{n}{k}}{2^n}.$$

This matches the binomial distribution of the cut size in the GSR model.

Card	Bit 1	Bit 2	Bit 3	Label
1	0	1	0	010
2	1	0	1	101
3	0	0	1	001
4	1	1	0	110
5	0	1	1	011

After assigning binary labels, the cards are sorted in lexicographic order:

$$001 < 010 < 011 < 101 < 110.$$

FIGURE 1. Inverse riffle shuffle representation. Each card is assigned a random binary string, and the deck is ordered lexicographically according to these labels.

Conditioning on the event that exactly k cards receive label 0, these cards form the first pile, while the remaining $n-k$ cards (having been labelled 1) form the second pile. The final ordering is obtained by collating these two piles while preserving the internal order within each pile.

Any such interleaving can be encoded by a sequence $\omega \in \{0,1\}^n$ containing exactly k zeros and $n-k$ ones, where $\omega_i = 0$ if the i-th position in the final deck is occupied by a card from the first pile, and $\omega_i = 1$ otherwise.

In the GSR model, it follows from the analysis of the interleaving procedure that, conditioned on a cut of size k, each such sequence ω occurs with equal probability:

$$\mathbb{P}(\omega \mid k) = \frac{1}{\binom{n}{k}}.$$

Therefore, the probability of obtaining a specific interleaving ω is

$$\mathbb{P}(\omega) = \mathbb{P}(k)\,\mathbb{P}(\omega \mid k) = \frac{\binom{n}{k}}{2^n} \cdot \frac{1}{\binom{n}{k}} = \frac{1}{2^n}.$$

In the inverse shuffle representation, each assignment of n independent bits occurs with probability $\frac{1}{2^n}$. Since each such assignment determines exactly one interleaving, it follows that every interleaving ω also occurs with probability

$$\mathbb{P}(\omega) = \frac{1}{2^n}.$$

The GSR construction and the inverse construction do not produce the same individual physical interleavings. They are different procedures with different sample spaces.

The point is that these two different constructions induce the same probability law once the shuffle is interpreted in the appropriate direction. Specifically, the inverse shuffle gives the distribution of the inverse of a GSR shuffle. This distinction is harmless for mixing, because inversion is a bijection on S_n and the uniform distribution is invariant under inversion. Therefore the total variation distance from uniform is the same for the GSR shuffle and for its inverse.

Thus, although the two constructions are not literally the same operation, they are equivalent for the purpose of analysing convergence to the uniform distribution.

5. Strong Stationary Times

5.1. **Stopping Times.** Let (X_t) be a Markov chain on a finite state space Ω. A random variable T taking values in $0, 1, 2, \ldots$ is called a stopping time if, for each $t \geq 0$, the event $T = t$ depends only on the values of $X_0, X_1, \ldots, X_t$.

A stopping time is a rule that determines when to stop observing the chain based only on the information available up to the present, and not of any future information. One might define T to be the first time the chain visits a particular state, or the first time a certain condition is satisfied.

5.2. **Strong Stationary Times.** Let π be the stationary distribution of the Markov chain. A stopping time T is called a *strong stationary time* if it satisfies the following two properties:

1: X_T has distribution π.
2: X_T is independent of T.

The first condition says that, at the random time T, the chain is exactly in stationarity. This is stronger than the usual mixing-time statement, which only says that the chain is close to stationarity at some deterministic time.

The second condition is what makes the stopping time a *strong* stationary time. It requires that the state X_T is independent of the value of T. Equivalently, even after conditioning on the event $\{T = t\}$, the distribution of X_T is still π.

In the context of shuffling, the independence condition means that knowing when the stopping time occurred will not reveal information about the final ordering of the deck.

5.3. **Example: Top-to-Random Shuffle.** Returning to the top-to-random shuffle to construct a strong stationary time for this chain.

Let the deck initially be in the ordered state

$$1, 2, \ldots, n.$$

Let τ be the first time at which the original bottom card reaches the top of the deck and is then removed and reinserted.

Proposition 5.1. *The stopping time τ is a strong stationary time for the top-to-random shuffle.*

Proof. Show that at time τ, the deck is uniformly distributed, and that this distribution is independent of τ.

Before card n reaches the top, it has not been moved. All other cards have been repeatedly removed from the top and reinserted at random positions.

Consider the block of cards lying below card n. Each time a card is inserted below card n, it is inserted uniformly among the existing cards in that block. By induction, the order of these cards is uniformly random.

At time $\tau - 1$, all other $n - 1$ cards lie below card n, and are therefore in uniform random order.

At time τ, card n is inserted uniformly among these cards, producing a uniformly random permutation of all n cards.

Moreover, this argument does not depend on the value of τ, so the distribution is independent of τ.

Therefore, τ is a strong stationary time. □

5.4. **Bounding Total Variation Distance.** Strong stationary times are particularly useful because they give a direct way to bound the distance to stationarity.

Proposition 5.2. *Let T be a strong stationary time. Then for all $t \geq 0$ and all initial states $x \in \Omega$,*

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbb{P}_x(T > t).$$

This shows that the total variation distance at time t is bounded above by the probability that the stopping time has not yet occurred by time t.

To understand why this is true, it helps to consider the following. If the event $\{T \leq t\}$ has occurred, then the chain has already reached stationarity at some time before or at time t. Because the chain evolves according to the Markov property and π is stationary, it remains in stationarity at time t. Thus, on the event $\{T \leq t\}$, the distribution of X_t is exactly π.

Any deviation from stationarity at time t must therefore come from the event $\{T > t\}$. The total variation distance can thus be controlled by the probability of this event. This reduces the problem of bounding mixing times to a probabilistic problem, meaning one must construct a strong stationary time and estimate the probability that it exceeds a given time t.

5.5. **Strong Stationary Times for the Riffle Shuffle.** The next stage is to apply these ideas to the riffle shuffle, using the inverse shuffle representation developed in the previous section.

As established, after k riffle shuffles, each card is assigned a binary string of length k, with each bit chosen independently and uniformly from $\{0, 1\}$. The position of the cards is determined by sorting these binary strings in lexicographic order, while preserving the relative order of cards with identical labels.

This representation suggests a natural stopping time. Define T to be the first time at which all cards have distinct binary labels. Now claim that T is a strong stationary time.

First, suppose that $T = k$. Then all cards have distinct labels in $\{0, 1\}^k$. Since the labels are independent and uniformly distributed, their relative order is completely random. Sorting the cards according to these labels therefore produces a uniformly random permutation of the deck. Hence,

$$X_T \sim \mathrm{Uniform}(S_n).$$

Second, the value of T records the first time at which all the binary labels become distinct. It depends on whether any two cards have the same label, but not on the relative order of the labels once they are distinct. The event $\{T = t\}$ informs that the labels were not all distinct before time t, but are all distinct at time t.

Conditional on $\{T = t\}$, the n binary labels at time t are distinct. Since the labels were assigned independently and symmetrically to the cards, each possible relative ordering of these n distinct labels is equally likely. Hence X_T has the uniform distribution even after conditioning on the value of T, proving that X_T is independent of T.

Therefore, T is a strong stationary time for the riffle shuffle, as stated. This construction is significant, as it reduces the problem of bounding the mixing time to estimating the probability that two cards share the same label after k shuffles. The desired upper bound for a riffle shuffle follows.[5]

6. The Main Proof

Proposition 6.1. *For the riffle shuffle on an n-card deck,*

$$t_{\text{mix}} \leq 2\log_2(4n/3)$$

for sufficiently large n.

Proof. Let τ be the first time at which all cards have distinct binary labels while using the inverse shuffle representation. Using the previous section, τ is a strong stationary time. Therefore, for every $t \geq 0$,

$$d(t) \leq \mathbb{P}(\tau > t),$$

where $d(t)$ is the total variation distance from stationarity at time t.

It is therefore satisfactory to choose t so that

$$\mathbb{P}(\tau > t) \leq \frac{1}{4}.$$

After t inverse riffle shuffles, each card has an independent label in $\{0,1\}^t$. Also, each card receives an independent label chosen uniformly from a set of size 2^t. The event $\{\tau \leq t\}$ is exactly the event that all n labels are distinct. Hence

$$\mathbb{P}(\tau \leq t) = \prod_{k=0}^{n-1} \left(1 - \frac{k}{2^t}\right).$$

Now set

$$t = 2\log_2(n/c)$$

for some constant $c > 0$. Then

$$2^t = \frac{n^2}{c^2}, \qquad \frac{k}{2^t} = \frac{c^2 k}{n^2}.$$

So

$$\log \mathbb{P}(\tau \leq t) = \sum_{k=0}^{n-1} \log\left(1 - \frac{c^2 k}{n^2}\right).$$

Since $\frac{c^2 k}{n^2} = O(1/n)$ uniformly for $0 \leq k \leq n-1$, it pertains to use the expansion

$$\log(1-x) = -x + O(x^2) \qquad (x \to 0),$$

uniformly in this range. Therefore,

$$\log \mathbb{P}(\tau \leq t) = -\sum_{k=0}^{n-1} \frac{c^2 k}{n^2} + O\left(\sum_{k=0}^{n-1} \frac{k^2}{n^4}\right).$$

Computing the two sums:

$$\sum_{k=0}^{n-1} k = \frac{n(n-1)}{2}, \qquad \sum_{k=0}^{n-1} k^2 = O(n^3).$$

Substituting these gives

$$\log \mathbb{P}(\tau \leq t) = -\frac{c^2}{n^2} \cdot \frac{n(n-1)}{2} + O\left(\frac{n^3}{n^4}\right) = -\frac{c^2}{2} + O\left(\frac{1}{n}\right).$$

Exponentiating,

$$\mathbb{P}(\tau \leq t) = \exp\left(-\frac{c^2}{2} + O\left(\frac{1}{n}\right)\right).$$

Now choose $c = \frac{3}{4}$. (This is chosen as c must be less than $\sqrt{2\log(4/3)}$, which is approximately 0.7585, so $\frac{3}{4}$ is a natural choice). Then

$$t = 2\log_2\left(\frac{n}{3/4}\right) = 2\log_2(4n/3),$$

and

$$\mathbb{P}(\tau \leq t) = \exp\left(-\frac{9}{32} + O\left(\frac{1}{n}\right)\right).$$

Since

$$e^{-9/32} > \frac{3}{4},$$

it follows that for sufficiently large n,

$$\mathbb{P}(\tau \leq t) > \frac{3}{4}.$$

Hence

$$\mathbb{P}(\tau > t) < \frac{1}{4}.$$

Therefore

$$d(t) \leq \mathbb{P}(\tau > t) < \frac{1}{4},$$

so by the definition of mixing time,

$$t_{\text{mix}} \leq t = 2\log_2(4n/3),$$

for sufficiently large n.

For a standard deck of $n = 52$ cards, this bound gives

$$2\log_2(4 \cdot 52/3) \approx 12.2.$$

Since the number of shuffles must be an integer, this corresponds to 13 riffle shuffles being sufficient according to this argument. It is key to remember here that sometimes a deck will be shuffled in less than 13 shuffles, in fact that is more than likely. There is a known result which state that 7 shuffles is all that is needed, the difference in results here is because what is calculated is an upper bound, not an average number of shuffles needed.[5] □

7. Lower Bounds on Mixing Time

The strong stationary time argument in the previous section gives an upper bound on the mixing time. It shows that after a certain number of riffle shuffles, the deck is close to stationarity. To move towards a more complete result and understanding, one also needs lower bounds.

A lower bound shows that before a certain time the chain is still far from stationarity. The standard method is to find a statistic of the chain whose distribution at time t is still noticeably different from its distribution under stationarity.

7.1. **Distinguishing Statistics.** Let (X_t) be a Markov chain on a finite state space Ω with stationary distribution π. Recall that

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} = \max_{A\subseteq\Omega} \left|\mathbb{P}_x(X_t \in A) - \pi(A)\right|.$$

Therefore, to prove a lower bound, it is enough to find an event A such that the probabilities of A under $P^t(x,\cdot)$ and under π are far apart.

Often this event is constructed using a real-valued function

$$f : \Omega \to \mathbb{R}.$$

Such a function is called a distinguishing statistic when it behaves differently at time t from how it behaves under stationarity.

Proposition 7.1. *Suppose that $f : \Omega \to \mathbb{R}$ satisfies*

$$\mathbb{E}_\pi[f] = 0.$$

Assume that, when the chain starts from x,

$$\mathbb{E}_x[f(X_t)] = m,$$

and that

$$\mathrm{Var}_x(f(X_t)) \le \sigma_t^2, \qquad \mathrm{Var}_\pi(f) \le \sigma_\pi^2.$$

If $m > 0$, then

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \ge 1 - \frac{4\sigma_t^2}{m^2} - \frac{4\sigma_\pi^2}{m^2}.$$

Proof. [5] Let

$$A = \{y \in \Omega : f(y) \ge m/2\}.$$

Under $P^t(x,\cdot)$, Chebyshev's inequality gives

$$\mathbb{P}_x(X_t \notin A) = \mathbb{P}_x(f(X_t) < m/2) \le \mathbb{P}_x(|f(X_t) - m| \ge m/2) \le \frac{4\sigma_t^2}{m^2}.$$

Hence

$$\mathbb{P}_x(X_t \in A) \ge 1 - \frac{4\sigma_t^2}{m^2}.$$

Under stationarity, again by Chebyshev's inequality,

$$\pi(A) = \pi(f \ge m/2) \le \pi(|f| \ge m/2) \le \frac{4\sigma_\pi^2}{m^2}.$$

Therefore

$$\|P^t(x,\cdot) - \pi\|_{\mathrm{TV}} \ge \mathbb{P}_x(X_t \in A) - \pi(A) \ge 1 - \frac{4\sigma_t^2}{m^2} - \frac{4\sigma_\pi^2}{m^2}.$$

□

This proposition [5] shows that if one can find a statistic whose expectation at time t is still large compared with its fluctuations, then the chain cannot yet be close to stationarity.

7.2. **Connection with Shuffling.** For card shuffling, natural distinguishing statistics include the position of a fixed card, the number of fixed points, and the number of rising sequences.

For the riffle shuffle, the number of rising sequences is especially important. A single riffle shuffle can produce only a limited amount of disorder, because it interleaves ordered packets. Repeated riffle shuffles increase the possible number of rising sequences, and the Bayer-Diaconis [1] analysis uses this statistic to obtain information about convergence to uniformity.

The inverse shuffle representation gives a simpler lower-bound intuition. After k inverse riffle shuffles, each card has a binary label of length k, so there are only 2^k possible labels. If 2^k is much smaller than n, then many cards must share labels. Cards with the same label keep their original relative order, so the deck still contains visible structure from its initial ordering.

8. Eigenvalues and Wilson's Method for Lower Bounds

Having now introduced general methods for obtaining lower bounds on mixing times using distinguishing statistics, it follows to now develop a more systematic approach based on the spectral properties of the Markov chain.

This method, often referred to as *Wilson's method* [5, 7], uses eigenvalues and eigenfunctions of the transition matrix to construct effective distinguishing statistics and obtain quantitative lower bounds on mixing time.

8.1. **Eigenvalues of Markov Chains.** Let P be the transition matrix of a Markov chain on a finite state space Ω. It is possible to view P as a linear operator acting on functions $f : \Omega \to \mathbb{R}$ via

$$(Pf)(x) = \sum_{y \in \Omega} P(x, y) f(y).$$

An eigenfunction of P is a non-zero function f satisfying

$$Pf = \lambda f$$

for some scalar λ, which is called the corresponding eigenvalue.

For a transition matrix, $\lambda = 1$ is always an eigenvalue, with constant eigenfunction. Moreover, every eigenvalue λ satisfies

$$|\lambda| \leq 1.$$

When the chain is reversible, the eigenvalues are real and can be ordered.

8.2. **Interpretation of Eigenvalues.** Eigenvalues describe how quickly different components of the chain decay toward stationarity.

If f is an eigenfunction with eigenvalue λ, then

$$\mathbb{E}_x[f(X_t)] = \lambda^t f(x).$$

Thus, the expectation of $f(X_t)$ decays at an exponential rate determined by λ. If λ is close to 1, then the decay is slow.

This observation suggests that eigenfunctions corresponding to eigenvalues close to 1 can be used as distinguishing statistics. If such a function

still has large expectation at time t, then the chain cannot yet be close to stationarity.

8.3. **Reversible Chains.** The spectral approach is particularly effective for reversible Markov chains. A chain with stationary distribution π is reversible if

$$\pi(x)P(x,y) = \pi(y)P(y,x)$$

for all $x, y \in \Omega$.

In this case, the operator P is self-adjoint with respect to the inner product

$$\langle f, g\rangle_\pi = \sum_{x\in\Omega} f(x)g(x)\pi(x).$$

This implies that:

- all eigenvalues are real
- there exists an orthonormal basis of eigenfunctions
- functions can be decomposed into eigencomponents

Orthonormal means that the eigenfunctions are orthogonal and normalised with respect to the inner product

$$\langle f, g\rangle_\pi = \sum_{x\in\Omega} f(x)g(x)\pi(x).$$

8.4. **Spectral Decomposition.** Let $\{f_1, f_2, \dots\}$ be an orthonormal basis of eigenfunctions with corresponding eigenvalues

$$1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots .$$

Any function f can be written as

$$f = \sum_i a_i f_i.$$

Applying P^t gives

$$P^t f = \sum_i a_i \lambda_i^t f_i.$$

Thus, components corresponding to eigenvalues with $|\lambda_i| < 1$ decay over time, while components corresponding to eigenvalues close to 1 persist.

8.5. **Using Eigenfunctions for Lower Bounds.** Suppose that f is an eigenfunction with eigenvalue $\lambda < 1$, normalised so that

$$\mathbb{E}_\pi[f] = 0.$$

Then

$$\mathbb{E}_x[f(X_t)] = \lambda^t f(x).$$

If $|\lambda|$ is close to 1, then λ^t remains large for moderate t. This means that $f(X_t)$ still has significant bias away from its stationary mean.

If, in addition, the variance of f under π is bounded, then this implies that the distribution at time t is far from stationarity.

This idea leads to a quantitative lower bound.

8.6. **Wilson's Method.** Wilson's method provides a systematic way to obtain lower bounds on mixing time using eigenfunctions of the transition matrix.[5] The key idea is to construct a function which evolves in a simple and controlled way under the Markov chain, and whose distribution remains far from its stationary distribution for a significant amount of time.

Let $f : \Omega \to \mathbb{R}$ be a function satisfying the following properties:

- $\mathbb{E}_\pi[f] = 0$,
- $Pf = \lambda f$ for some eigenvalue λ with $|\lambda| < 1$,
- $\mathrm{Var}_\pi(f) = \sigma^2$ is finite.

The condition $Pf = \lambda f$ means that f is an eigenfunction of the transition operator. Iterating this relation gives

$$P^t f = \lambda^t f,$$

and hence

$$\mathbb{E}_x[f(X_t)] = \lambda^t f(x).$$

Thus, the expectation of $f(X_t)$ decays exponentially in t. However, if λ is close to 1, then this decay is slow, and $\lambda^t f(x)$ remains large for a significant amount of time.

At the same time, the variance of $f(X_t)$ remains controlled. In particular, under mild conditions one can ensure that

$$\mathrm{Var}(f(X_t)) \leq \sigma^2$$

for all t, where $\sigma^2 = \mathrm{Var}_\pi(f)$.

This leads to comparing the behaviour of $f(X_t)$ under the distribution at time t with its behaviour under the stationary distribution.

Under π, the function f has mean 0 and variance σ^2. By Chebyshev's inequality, for any $a > 0$,

$$\pi(|f| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Thus, under stationarity, the values of f are typically small.

On the other hand, under $P^t(x, \cdot)$, the random variable $f(X_t)$ has mean $\lambda^t f(x)$. If $\lambda^t f(x)$ is large compared to σ, then $f(X_t)$ is typically far from 0.

To make this precise, consider the set

$$A = \{y \in \Omega : f(y) \geq \tfrac{1}{2}\lambda^t f(x)\}.$$

Under the distribution at time t, the value of $f(X_t)$ is concentrated near $\lambda^t f(x)$, so the probability that $X_t \in A$ is close to 1.

Under the stationary distribution however, the probability that $f(y)$ is this large is small, since f is centred at 0 with bounded variance.

This shows that the two distributions place most of their mass on different regions of the state space, and are therefore far apart in total variation distance.

This argument leads to the following method due to David Wilson.[7]

Theorem 8.1 (Wilson's method)**.** *Let (X_t) be an irreducible, aperiodic Markov chain on a finite state space Ω with transition matrix P. Let $\Phi : \Omega \to \mathbb{R}$ be an eigenfunction of P with real eigenvalue λ satisfying*

$$\frac{1}{2} < \lambda < 1.$$

Let $0 < \varepsilon < 1$, and suppose that there exists $R > 0$ such that

$$\mathbb{E}_x \left[\left| \Phi(X_1) - \Phi(x) \right|^2 \right] \leq R$$

for every $x \in \Omega$.

Then, for every starting state $x \in \Omega$,

$$t_{\text{mix}}(\varepsilon) \geq \frac{1}{2 \log(1/\lambda)} \left[\log \left(\frac{(1-\lambda)\Phi(x)^2}{2R} \right) + \log \left(\frac{1-\varepsilon}{\varepsilon} \right) \right].$$

The role of this theorem is to turn a slowly decaying eigenfunction into a lower bound on the mixing time. The condition $\lambda < 1$ ensures that the expectation of $\Phi(X_t)$ eventually decays, while the condition $\lambda > 1/2$ ensures that this decay is slow enough to be useful. The quantity R controls the one-step fluctuations of the eigenfunction.

In applications, one usually chooses Φ so that $|\Phi(x)|$ is large at a carefully chosen starting state x, while λ is close to 1. Then $\lambda^t \Phi(x)$ remains large for a long time, showing that the chain still retains information about its initial state.

The significance of this result is that it provides a direct way to obtain lower bounds on mixing time. If there exists an eigenvalue λ close to 1, then the quantity λ^t decays slowly, and the chain retains memory of its initial state for a long time.

Thus, Wilson's method refines the distinguishing statistics approach by selecting a statistic with particularly simple and well-understood evolution. This makes it a powerful tool for proving sharp lower bounds, and plays an important role in the analysis of cut-off phenomena.

8.7. **Application to Card Shuffling.** In card shuffling, the most useful statistics are those which measure how much of the original order remains in the deck.

For the riffle shuffle, the central statistic in the Bayer-Diaconis [1] analysis is the number of rising sequences. Equivalently, one may count descents in the inverse permutation. This statistic is important because the probability of obtaining a particular permutation after an a-shuffle depends only on its number of rising sequences.

Another common statistic is the number of inversions. An inversion is a pair of cards which appear in the opposite order from their original order. If the deck is perfectly ordered, the number of inversions is 0. Under the uniform distribution on S_n, the expected number of inversions is

$$\frac{1}{2}\binom{n}{2}.$$

Thus, the inversion count measures how far the deck is from its original order.

For riffle shuffles, inversion-type statistics are natural because they detect remaining order in the deck. Diaconis and Fulman [2] note that certain eigenfunctions for riffle shuffling can be described using the relative order of pairs of cards. In particular, functions which record whether card i appears

before card j are closely related to eigenfunctions, and the inversion statistic is obtained by summing this information over all pairs of cards.

If after k shuffles the expected number of inversions is still far from its uniform value, then the deck cannot yet be close to uniformly random. Bayer and Diaconis [1] use the exact formula for the probability of a permutation with a given number of rising sequences to show that the transition to stationarity occurs around

$$\frac{3}{2}\log_2 n.$$

9. The Cut-off Phenomenon

One of the features of the riffle shuffle is that convergence to stationarity does not occur linearly, or even at any gradual rate. Instead, the transition from a highly ordered deck to a well-mixed deck happens quite abruptly over a small window of shuffles. This behaviour is known as the *cut-off phenomenon*, and will help explain the 7 (or 8) shuffles result. The aim will not be to create a fully fleshed out proof, the full Bayer–Diaconis [1] asymptotic theorem and its related works are incredibly complex and would not fit in this paper, but the following should outline what is necessary to understand.

9.1. **Cut-off Phenomenon.** Let (X_t) be a sequence of Markov chains indexed by the size of the state space, and let

$$d_n(t) = \max_x \|P^t(x,\cdot) - \pi\|_{\mathrm{TV}}$$

be the worst-case total variation distance from stationarity.

Informally, a family of Markov chains has a cut-off if the convergence to stationarity happens suddenly. That is, the distance stays close to 1 for a long time, and then drops to close to 0 over a much shorter time window.

More formally, suppose there is a sequence of times t_n and a sequence $w_n = o(t_n)$ such that, for each fixed $c > 0$,

$$\lim_{n\to\infty} d_n(t_n - cw_n) = 1$$

and

$$\lim_{n\to\infty} d_n(t_n + cw_n) = 0.$$

Then the family is said to exhibit cut-off at time t_n, with cut-off window w_n.

The important point is that the window w_n is much smaller than the mixing time itself. Thus, the chain does not gradually become random. Instead, it remains far from random until shortly before t_n, and then rapidly becomes close to random.

9.2. **Rising Sequences.** For the riffle shuffle, the cut-off phenomenon is closely connected to the number of rising sequences.

Definition 9.1. *Let $\pi \in S_n$ be a permutation. A* rising sequence *of π is a maximal subsequence of consecutive card values appearing in increasing order.*[1]

For example, consider the arrangement

$$1, 5, 2, 3, 6, 7, 4.$$

The rising sequences are

$$1, 2, 3, 4$$

and

$$5, 6, 7.$$

Thus this permutation has 2 rising sequences.

Rising sequences measure how much of the original order remains after shuffling. A completely ordered deck has only one rising sequence. A highly mixed deck tends to have many rising sequences.

The importance of rising sequences comes from the fact that after riffle shuffling, the probability of a permutation depends only on its number of rising sequences.

9.3. **The Probability Formula for an a-shuffle.** The riffle shuffle explored so far in this paper is a 2-shuffle. More generally, an a-shuffle is performed by cutting the deck into a piles and interleaving those piles while preserving the relative order within each pile.

Repeated riffle shuffles fit into this framework because performing k ordinary riffle shuffles is equivalent to performing one 2^k-shuffle. This follows naturally from the inverse shuffle representation that after k inverse riffle shuffles, each card has a binary string of length k, and hence one of 2^k possible labels.

Theorem 9.2. *Let $Q_a(\pi)$ be the probability of obtaining the permutation π after an a-shuffle. If π has r rising sequences, then*

$$Q_a(\pi) = \frac{1}{a^n}\binom{a+n-r}{n}.$$

Proof Idea. The inverse a-shuffle assigns to each card an independent label from

$$\{0, 1, \ldots, a-1\}.$$

The deck is then sorted according to these labels, preserving the original relative order among cards with the same label.

To obtain a fixed permutation π, cards lying in the same rising sequence may be assigned the same label, but labels must increase when moving from one rising sequence to the next. Thus, producing π is equivalent to assigning labels to the r rising sequences in a weakly increasing way using labels from $\{0, 1, \ldots, a-1\}$.

The number of such assignments is

$$\binom{a+n-r}{n}.$$

Since there are a^n total possible label assignments, the probability is

$$Q_a(\pi) = \frac{1}{a^n}\binom{a+n-r}{n}.$$

□

This formula is important because it gives exact probabilities for riffle shuffling. Instead of needing to track all of the $n!$ permutations individually, it is only needed to know how many permutations have each possible number of rising sequences.

9.4. **Connecting Rising Sequences to Total Variation Distance.** Let $A(n, j)$ denote the Eulerian number(the number of permutations in S_n with exactly j descents, with descents being a sequence of decreasing order) A permutation has r rising sequences if and only if it has $r-1$ descents in the appropriate inverse permutation. Therefore, the number of permutations with r rising sequences is given by an Eulerian number.

Using the formula above, the total variation distance from uniform after an a-shuffle can be written as

$$\|Q_a - U\|_{\mathrm{TV}} = \frac{1}{2}\sum_{r=1}^{n} A(n, r-1)\left|\frac{1}{a^n}\binom{a+n-r}{n} - \frac{1}{n!}\right|.$$

The difficult part is then to estimate this sum asymptotically as n grows.

Bayer and Diaconis [1] carried out this asymptotic analysis and showed that the critical point occurs at

$$k = \frac{3}{2}\log_2 n.$$

More precisely, if

$$k = \frac{3}{2}\log_2 n + \theta,$$

then for large n,

$$\|Q^{*k} - U\|_{\mathrm{TV}} = 1 - 2\Phi\left(-\frac{2^{-\theta}}{4\sqrt{3}}\right) + O(n^{-1/4}),$$

where

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-u^2/2}\,du$$

is the standard normal distribution function.

This formula describes the cut-off's shape. When θ is negative, the term $2^{-\theta}$ is large, and the total variation distance is close to 1. When θ is positive, $2^{-\theta}$ is small, and the distance decreases rapidly toward 0.

Thus, the transition from unmixed to mixed occurs in a narrow window around

$$\frac{3}{2}\log_2 n.$$

9.5. **A 52-Card Deck.** For a standard deck of $n = 52$ cards,

$$\frac{3}{2}\log_2 52 \approx 8.6.$$

Results show [1] that the total variation distance behaves approximately as follows, for k shuffles:

k	1	2	3	4	5	6	7	8
$\|Q^{*k} - U\|_{\mathrm{TV}}$	1.000	1.000	1.000	1.000	0.924	0.614	0.334	0.167

This presents the cut-off phenomenon clearly. The distance remains very close to 1 for the first few shuffles, and then drops sharply between 5 and 8 shuffles, and will tend towards 0 as k increases.

A well-known rule of thumb is that about 7 riffle shuffles are sufficient to randomise a standard 52-card deck. From this table, and using $t_{\text{mix}} = t_{\text{mix}}(1/4)$, 8 shuffles would be suggested. The result of 7 riffles is more applicable in casual circumstances, where $t_{\text{mix}} = t_{\text{mix}}(1/3)$ would be acceptable.

9.6. **Practical Explanation.** The cut-off phenomenon can be understood practically using the inverse shuffle representation.

After k riffle shuffles, each card is assigned a binary string of length k. The total number of possible labels is 2^k. When 2^k is much smaller than n, many cards share the same label, and the deck retains significant structure.

However, once 2^k becomes comparable to or larger than n, it becomes unlikely that many cards share labels. At this point, the ordering of the cards becomes effectively random.

The transition from $2^k \ll n$ to $2^k \gg n$ occurs when $k \approx \log_2 n$, which explains the logarithmic scaling of the mixing time. The constant $\frac{3}{2}$ arises from a further analysis of the distribution of permutations after repeated shuffles.

10. Wash Shuffling and Spatial Mixing

So far, this paper has focused mainly on riffle shuffling. The riffle shuffle has a particularly clean mathematical structure, in that it can be modelled as a random walk on S_n, and the inverse shuffle representation allows repeated shuffles to be studied using random binary labels.

However, not all common shuffling methods have this structure. A different practical method is the *wash shuffle*, also called the *smoosh shuffle*. This is discussed in Chapter 14 of *The Mathematics of Shuffling Cards*.[2]

The wash shuffle is performed by spreading the cards face down on a table, moving them around by hand for some amount of time, and then gathering them back together into a deck. This is common in casino and card-room settings, especially when several decks are being mixed at once.

The main distinction from the riffle shuffle is that wash shuffling is not naturally a process of cutting and interleaving ordered packets. Instead, it is a form of *spatial mixing*. The cards move around in a two-dimensional region before being collected back into a one-dimensional order. Therefore, instead of tracking cuts, interleavings, or binary labels, one has to track the physical positions of cards on a table. Naturally, this is significantly more complex, so both a theoretical and computer model will be looked at.

10.1. **What is Being Modelled?** Start with n labelled cards spread out on a table of a set size. The cards are then repeatedly pushed around by the shuffler's hands. At the end, the cards are gathered back together to form a deck order.

Like in riffle and top-to-random shuffles, the aim is to see how long it takes for the deck to become sufficiently close to stationarity, but instead of a fixed number of shuffles, the answer will be a number of seconds. Consider

wash shuffling as continuous-time process, while riffling is more a discrete sequence.

A possible failure of this shuffle is that cards which began next to each other are more likely to remain near each other after shuffling. This motivates statistics such as the number of originally adjacent pairs that remain adjacent after the shuffle. If too many original neighbours remain together, then the deck is not behaving like a uniformly random permutation.

10.2. **A Spatial Model for Wash Shuffling.** First, it is needed to introduce a mathematical model for wash shuffling, in this case the *gather-and-spread* model.[2] The model is designed to capture the spatial nature of the shuffle while still being simple enough to analyse.

The table is represented by the unit square

$$D = [0, 1]^2.$$

The cards are represented by labelled points

$$Z_1(t), Z_2(t), \ldots, Z_n(t) \in D.$$

At time $t = 0$, these points may be placed in a row, representing the initial ordered deck. The shuffle then evolves by repeatedly applying two operations:

- gathering nearby cards together,
- spreading them in a random direction.

The shuffler's hand is modelled as a disc

$$U = \{z \in \mathbb{R}^2 : |z| \leq \delta\},$$

where $\delta > 0$ represents the radius of the hand.

Gathering. Choose a point

$$z_0 \in D.$$

The hand is placed with centre z_0. All cards under the hand are gathered to the point z_0, while cards outside the hand remain fixed.

This can be represented by a map

$$G_{z_0} : D \to D$$

defined by

$$G_{z_0}(z) = \begin{cases} z_0, & z \in z_0 + U, \\ z, & z \notin z_0 + U. \end{cases}$$

Thus, any card lying under the hand is moved to the hand's centre. This models the way a hand can collect a local group of cards into a small pile or cluster.

Spreading. After gathering, the cards under the hand are moved in a random direction. Choose an angle

$$\theta \in [0, 2\pi].$$

If the card is under the hand, it is moved a distance s in the direction θ. Ignoring boundary effects, this movement is

$$(x, y) \mapsto (x + s\cos\theta,\ y + s\sin\theta).$$

Cards outside the hand are unchanged.

Boundary effects are handled by keeping cards inside the table. If a card would move outside the square, it is instead kept on the boundary.

Breaking Ties. The gathering step can place several cards at exactly the same point. To avoid cards remaining permanently stuck together, the model introduces an additional randomisation.

Fix

$$0 < p < 1.$$

When a card is about to be spread, it tosses a Bernoulli(p) coin. If the coin lands heads, the card moves; otherwise it remains in place.

This breaks ties between cards that were gathered to the same point, because cards gathered together do not necessarily all move in the same way afterwards.

The Hopping Hand. The hand does not act in just one fixed location, but repeatedly chooses new random locations and directions.

This is modelled using a Poisson point process. At random times

$$0 < t_1 < t_2 < \cdots,$$

the hand chooses a location

$$w_i \in D$$

and a direction

$$\theta_i \in [0, 2\pi].$$

At each time t_i, the gather-and-spread operation is applied. This produces a continuous-time Markov process

$$(Z_1(t), \ldots, Z_n(t)).$$

Finally, to turn the spatial configuration back into a deck order, the card positions are projected onto the x-axis and collected from left to right. Ties are broken randomly. This gives a random permutation

$$\pi(t) \in S_n.$$

10.3. **Computer Modelling.** It is of interest to generate a computer model of this procedure. As mentioned, generating a mixing time bound for this is not as simple as other shuffles, so a computer model could be useful. The simplest approach is to represent each card by the position of its centre. Then the gather-and-spread model can be simulated by repeatedly choosing a random hand position and moving nearby card-centres.

In this paper, the programming I have created uses the aforementioned gather-and-spread model, modelling each card by a single point at the centre of the card. There are some drawbacks to this, such as losing some information in relation to cards interactions which each other (friction, collisions, etc). The resulting model should therefore be interpreted as a simplified spatial model rather than a complete physical simulation. Also, let it be noted that from real world testing, it was found one can perform 25 *gather-and-spread* iterations in 10 seconds. (As in, one can place their hand, move a pile, and remove their hand 25 times in 10 seconds.)

10.4. **Testing Mixing in Wash Shuffles.** Because the exact distribution of a real wash shuffle is hard to compute, there are empirical tests of mixing.

The difficulty is that the state space is enormous:

$$|S_{52}| = 52!.$$

Testing directly whether the distribution on all 52! permutations is uniform is not realistic with ordinary sample sizes. Instead, one tests particular statistics which should reveal obvious non-randomness.

A natural statistic is the number of originally adjacent pairs that remain adjacent after shuffling. For a permutation $\sigma \in S_n$, define

$$T_A(\sigma) = \sum_{i=1}^{n-1} \mathbf{1}_{\{\sigma_{i+1}=\sigma_i+1\}}.$$

This counts adjacent pairs from the original deck which still appear as adjacent pairs in the same order after shuffling.

For example, if

$$\sigma = 1, 3, 2, 4, 6, 5, 7, 8, 9,$$

then the pairs $7, 8$ and $8, 9$ remain adjacent, so

$$T_A(\sigma) = 2.$$

At the start, when the deck is ordered,

$$T_A(\sigma) = n - 1.$$

Under the uniform distribution on S_n, $T_A(\sigma)$ is approximately Poisson with mean 1. Thus, for a well-shuffled deck, one expects

$$T_A(\sigma)$$

to usually be close to 0, 1, or 2.

This is the statistic that will be used for this paper's computer analysis.

10.5. **Computer Analysis Results.** The results for running 100 trials of 60, 30 and 15 seconds of wash shuffling are as follows.

T_A	15 seconds (37 iterations)	30 seconds (75 iterations)	60 seconds (150 iterations)
0	0	26	38
1	0	41	34
2	3	20	20
3	8	7	5
4	12	3	3
5+	77	3	0

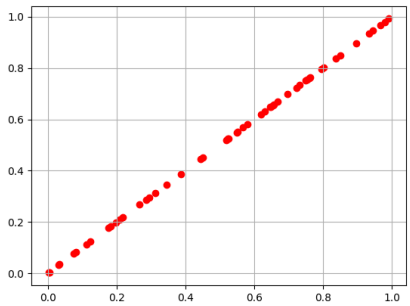
Where the expected values are

T_A	0	1	2	3	4	5+
Poisson(1), 100 trials	36.8	36.8	18.4	6.1	1.5	0.4

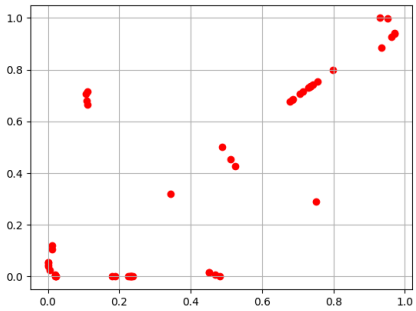
Clearly, 15 seconds is not sufficient, 60 seconds is very close to the Poisson(1) benchmark, and 30 seconds provides an intermediate result. Comparing the T_A totals for each supports this as well.

Shuffle / distribution	Total T_A over 100 trials
Expected under stationarity	100
15 seconds	683
30 seconds	129
60 seconds	101

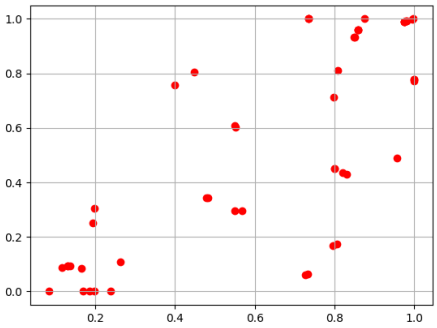
These numbers are consistent with those found in other computer analysis [2], which find approximately 30 seconds sufficient for a wash shuffle.



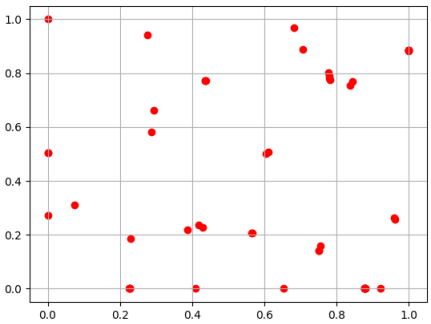
Starting position before wash shuffling.



Simulated wash shuffle after 15 seconds.



Simulated wash shuffle after 30 seconds.



Simulated wash shuffle after 60 seconds.

FIGURE 2. Simulated wash shuffle positions at different times.

10.6. **Theoretical Mixing Bounds.** First, a brief overview of *coupling*. A *coupling* of two probability distributions μ and ν on the same state space Ω is a pair of random variables (X, Y) such that

$$X \sim \mu \qquad \text{and} \qquad Y \sim \nu.$$

That is, X has distribution μ and Y has distribution ν, but they are constructed on the same probability space and may depend on each other.

For Markov chains, a coupling means running two copies of the chain together, usually using shared randomness, so that each copy has the correct transition probabilities. The aim is often to construct the two chains so that they eventually meet. Once the two chains have met, they are made to move together thereafter.

Coupling is useful because it gives a way to bound total variation distance. If (X_t, Y_t) is a coupling where X_t starts from a fixed initial state and Y_t starts from stationarity, then

$$\|P^t(x, \cdot) - \pi\|_{\mathrm{TV}} \leq \mathbb{P}(X_t \neq Y_t).$$

Thus, if the two coupled chains are likely to have met by time t, then the chain starting from x must be close to stationarity.

The theoretical analysis of the gather-and-spread model uses coupling. The idea is to run two versions of the spatial process together. One version starts from a fixed initial ordering. The other starts with the same card positions, but with a uniformly random labelling of the cards. If the two processes can be coupled so that all labels become matched, then the resulting projected permutation is uniform.

Let τ^* be the first time at which all cards in the original process have been paired with their corresponding cards in the uniformly labelled process. Then the coupling gives the total variation bound

$$\|P^t - U\|_{\mathrm{TV}} \leq \mathbb{P}(\tau^* \geq t).$$

This is similar in spirit to the strong stationary time inequality used earlier. In both cases, one controls total variation distance by bounding the probability that a certain random time has not yet occurred.

The full gather-and-spread model is difficult to analyse. Instead, it is simpler to describe a limiting approximation [2], in which the card positions evolve as a jump diffusion. The cards move like reflecting Brownian motions inside the square, with occasional gathering events.

This jump diffusion model gives an explicit logarithmic upper bound on the mixing time. If $t^*(\varepsilon)$ is the time needed for the induced permutation to be within ε of uniformity, then

$$t^*(\varepsilon) \leq \frac{1}{\eta} \log\left(\frac{n}{\varepsilon}\right),$$

where $\eta > 0$ is a constant depending on the parameters of the model.

The exact complicated expression for η is:

$$\eta = \frac{\delta^2}{2p\pi\sigma^2(1+2\delta)^2} K_0\left(\frac{\sqrt{2}+2\delta}{\sigma\delta}\sqrt{\pi p(1-p)}\right),$$

where K_0 is the modified Bessel function of the second kind.[6]

For this paper, the important point is the order of the bound:

$$t^*(\varepsilon) = O\left(\log\left(\frac{n}{\varepsilon}\right)\right).$$

This means that, in the idealised jump diffusion version of the gather-and-spread model, the time needed to mix grows only logarithmically with the number of cards.

This is interesting because the same broad logarithmic behaviour appears in riffle shuffling, but for a completely different reason.

11. Closing Remarks

This paper has studied card shuffling as a problem in the theory of finite Markov chains, with the central question being how long it takes to shuffle a deck of cards, most notably with the riffle shuffle.

The main result proven was the upper bound for riffles on a n card deck,

$$t_{\text{mix}} \leq 2\log_2(4n/3)$$

for sufficiently large n. This proof used strong stationary times and inversion of shuffling (specifically how that is viable under the symmetric group S_n), to provide the answer of 13 shuffles for a $n = 52$ card deck.

Further analysis in this paper covered Wilson's method for lower bounds, explored the cut-off phenomenon and the known 7 shuffles result, and finally explored spatial mixing and computer modelling for more complex shuffles.

While the content in this paper is specific to card shuffling, this mathematics has a wide range of use cases. Topics such as MCMC (Markov chain Monte Carlo), biology and genetics, cryptography and cybersecurity, random walks on graphs, and statistical physics (to name but a few) all have considerable crossover with the mathematics explored in this dissertation, and all provide other real world use cases for this paper's methods.

References

[1] D. Bayer and P. Diaconis. Trailing the dovetail shuffle to its lair. *The Annals of Applied Probability*, 2(2):294–313, 1992.

[2] P. Diaconis and J. Fulman. *The Mathematics of Shuffling Cards*. The American Mathematical Society, 2023.

[3] P. Flajolet, D. Gardy, and L. Thimonier. Birthday paradox, coupon collectors, caching algorithms and self-organizing search. *Discrete Applied Mathematics*, 39(3):207–229, 1992.

[4] E. W. Gilbert. Theory of shuffling. *Bell Laboratories Technical Memorandum*, 1955.

[5] D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times, second edition*. The American Mathematical Society, 2017.

[6] S. G. Samko, A. A. Kilbas, and O. I. Marichev. *Fractional Integrals and Derivatives: Theory and Applications*, chapter 1. 1993.

[7] D. B. Wilson. Mixing times of lozenge tiling and card shuffling markov chains. *The Annals of Applied Probability*, 14(1), Feb. 2004.