# Markov Chains and Mixing Times

Martina Häberling
supervised by: Prof. Dr. Jean Bertoin

Universität Zürich

Institute of Mathematics

Dezember 2024

## Abstract

In this master thesis we see different techniques to determine a bound of the mixing time of Markov chains, following Chapter 4 to 8 of 'Markov Chains and Mixing Times' by David A. Levin and Yuval Peres and Chapter 3 of 'Reversible Markov Chains and Random Walks' by David A. Levin and James Allen Fill. We will study the distance of the distribution of the Markov chain to the stationary distribution. The mixing time is the number of steps for the distribution to reach stationarity. This is interesting for real-life applications, such as in shuffling cards and random walks on groups. During our search for the bounds of mixing times, we will look at different mathematical concepts and techniques, such as different shufflings, bottleneck ratio, transpositions and coupling. Furthermore, we will dive deeper into reversible Markov chains and graphs.

## Acknowledgements

I would like to thank my supervisor Prof. Dr. Jean Bertoin for this fascinating topic and his guidance through the process of my master thesis.
I am very thankful to everyone who supported me throughout my studies, especially my parents. Lastly, I would like to thank my former math teacher Katja Badal for encouraging me to study mathematics.

# Contents

**0 Introduction** **1**

**1 Preliminaries** **3**
1.1 Repetition . . . . . 3
1.2 Coupon collectors . . . . . 4

**2 Definition of Mixing Times** **5**
2.1 Total variation distance . . . . . 5
2.2 Coupling . . . . . 7
2.3 Standardizing Distance . . . . . 7
2.4 Mixing Times . . . . . 9

**3 Top to random shuffle** **11**
3.1 $\tau_{Top}$ . . . . . 11
3.2 Stopping Times . . . . . 11
3.3 Stationary Times and Strong Stationary Times . . . . . 11
3.4 Separation Distance . . . . . 12
3.5 Finding a bound . . . . . 12

**4 Bottleneck-Ratio** **14**
4.1 Binary Tree . . . . . 16

**5 Random Transposition** **19**
5.1 The Symmetric Group . . . . . 19
5.2 Upper bound . . . . . 21

**6 Reversible Markov Chains** **24**
6.1 Reversible Markov chains and Time-Reversal . . . . . 24
6.2 Reversed Distribution . . . . . 24
6.3 Example . . . . . 25

**7 Graphs** **30**
7.1 Graphs and Random Walks . . . . . 30
7.2 Fluid model and electrical networks . . . . . 32
7.3 Grand coupling and random colourings . . . . . 35
7.4 Example . . . . . 35

**8 Conclusion** **40**

# 0 Introduction

Markov chains were first studied in 1906 by Andrey Markov. They are stochastic processes, which describe a sequence of possible events, for which the probability only depends on the previous event. Around a century later Markov chains are used in many different fields, such as computer science, physics, statistics and many more. Classically, Markov chains were studied as fixed chains, but also the convergence of Markov chains became an interesting topic to explore. An important result was the theorem **Convergence to equilibrium**:

**Theorem 0.1.** *If $X$ is an irreducible, aperiodic Markov chain on the state space $\mathcal{X}$ and there is a stationary law $\pi$, then for any $x \in \mathcal{X}, y \in \mathcal{X}$:*

$$\lim_{n\to\infty} \mathbb{P}_x(X_n = y) = \pi(y).$$

Meaning that the Markov chain converges to the stationary distribution.
With that knowledge, the next question arose: What is the rate of convergence to stationarity of the distribution at time $t$, when $t$ goes to infinity? One started to describe the distance to the stationary distribution and then the mixing time was defined as the number of steps required to reach a certain distance to the stationary distribution. Now a question for the future is, how the mixing time will change, if the state space grows.

For example, this concept became useful to mathematicians (like David Aldous and Persi Diaconis), who started studying card shuffling and random walks on groups. Card shuffling can be looked at as a Markov chain of all possible configurations on the deck of cards and so the question would become: How many shuffles are needed until the distribution of the deck is approximately uniform? The result can be used in real life. For a card deck of 52 cards we would need to shuffle 8 times for a riffle shuffle or 205 times for a top-to-random shuffle.

Reversible finite Markov chains, which we also discuss in this thesis, can be identified as random walks on weighted graphs and be seen as resistor networks. This holds close resemblance to the classical electrical networks. Using this, one can find the mixing time of the reversed Markov chain and compare it to the mixing time of the original Markov chain.

This is nicely summarized in 'Markov Chains and Mixing Times' by David A. Levin and Yuval Peres [LP17] and 'Reversible Markov Chains and Random Walks' by David A. Levin and James Allen Fill [Fil14]. This thesis is a collection of results of those two books. More precisely of Chapter 4 to 8 of 'Markov Chains and Mixing Times' and Chapter 3 of 'Reversible Markov Chains and Random Walks'. I have created my own version of these chapters. None of the results in this paper are claimed to be original.
More books on this subject are the ones of Häggström [Hä02], Jerrum [Jer03], Lovàsz [Lov93] and Montenegro and Tetali [MT$^+$06].

After repeating and stating some definition useful for the thesis, we will look at the

definition of the mixing time and how we can build it, using the distance from the distribution of the Markov chain to the stationary distribution in Chapter 2. Additionally we will introduce Coupling, a useful tool for arguments in the following chapters. We continue in Chapter 3 with the top-to-random shuffle and bound for the first time the mixing time to this shuffle. It follows an introduction of Bottlenecks in Chapter 4. We apply this to calculate the lower bound of a lazy walk on a binary tree. In Chapter 5 we talk about random transportations and find an upper bound for the mixing time of the random transportation shuffle. Next, in Chapter 6, we see how the mixing time of a chain can be compared to the mixing time of the reversed chain. Lastly, in Chapter 7, we dive into the topic of graphs. We will see different kind of graphs and end in a last example of bounding the mixing time, this time using the metropolis chain on proper colourings.

# 1 Preliminaries

Before we start, we look at a few definitions and concepts which we will use throughout the text. All these definitions and statements are from my own notes of Probability 1 and 2 or from 'Markov chains and Mixing Times' [LP17].

## 1.1 Repetition

As our basis, we assume that Probability 1 and 2 are known. But we will still be repeating some useful definitions and results.

We are working on $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathcal{F}$ a $\sigma$-algebra and $\mathbb{P}$ a probability measure.
A **filtration** $(\mathcal{F}_n)_{n\geq 0}$ is an increasing sequence of sub-$\sigma$-algebras in $\mathcal{F}$.
A stochastic process (a sequence $(X_n)_{n\geq 0}$ of random variables) is **adapted** to a filtration $(\mathcal{F}_n)_{n\geq 0}$ if and only if $\forall n \geq 0$, $X_n$ is $\mathcal{F}_n$-measurable.
The **natural filtration** generated by a stochastic process $(X_n)_{n\geq 0}$ is $\mathcal{F}_n = \sigma(X_0, X_1, ..., X_n)$. It is the smallest filtration making $(X_n)$ an adapted process.

A **random sample** from a finite set $\mathcal{X}$ is a random uniform selection from $\mathcal{X}$.

The **simple random walk** on a graph $G$ is defined by choosing each step one of the neighbours of the current position uniformly at random. Giving us the transition matrix

$$P(x,y) = \begin{cases} \frac{1}{deg(x)} & \text{if } y \sim x, \\ 0 & \text{otherwise.} \end{cases}$$

We can define the increment distribution:
Take a probability distribution $\mu$ on a group $G$, we can define the left random walk on $G$ with **increment distribution** $\mu$ to be: Take a Markov chain with state space $G$ and let it move by multiplying the current state on the left by an element of $G$ selected according to $\mu$. We can see that the transition matrix $P$ then has entries $P(g, hg) = \mu(h) \ \forall\, g, h \in G$.

A **Markov chain** is a process that moves on the set $\mathcal{X}$ as follows: At $x \in \mathcal{X}$ the next position is chosen with a fixed probability distribution $P(x, \cdot)$, meaning that the process only depends on the position it is on, not where it was before.
More formally, $(X_t)$ is a Markov chain with state space $\mathcal{X}$ and transition matrix $P$, if $\forall\ x, y \in \mathcal{X};\ t \geq 1;\ H_{t-1} = \bigcap_{s=0}^{t-1}\{X_s = x_s\}$ satisfying $\mathbb{P}\{H_{t-1} \cap \{X_t = x\}\} > 0$, we get the Markov property:

$$\mathbb{P}\{X_{t+1} = y | H_{t-1} \cap \{X_t = x\}\} = \mathbb{P}\{X_{t+1} = y | X_t = x\} = P(x,y)$$

We can see, that the $x$-th row of the transition matrix $P$ is the distribution $P(x, \cdot)$, and therefore $P$ is stochastic, meaning all entries are non-negative and $\sum_{y\in\mathcal{X}} P(x,y) = 1$ $\forall\ x \in \mathcal{X}$.

The **stationary distribution** of the Markov chain is the probability distribution $\pi$ satisfying:

$$\pi(y) = \sum_{x \in \mathcal{X}} \pi(x) P(x, y) \quad \forall\ y \in \mathcal{X}$$

One of the last observation we need from the Probability theory is the Theorem **Convergence to equilibrium** for irreducible positive recurrent Markov chains.

**Theorem 1.1.** *Let $X$ be an irreducible, aperiodic Markov chain on $\mathcal{X}$. Suppose there exists a stationary distribution $\pi$. Then for any $x, y \in \mathcal{X}$: $\lim_{n\to\infty} \mathbb{P}_x(X_n = y) = \pi(y)$. More precisely, $\sum_{y \in \mathcal{X}} |\mathbb{P}_x\{X_n = y\} - \pi(y)| \xrightarrow{n \to \infty} 0$.*

A natural continuation is to ask: How long will we need to reach this equilibrium?

A last important result of Probability is the Markov inequality:

**Proposition 1.2.** *If $X$ is integrable and $a > 0$, then $\mathbb{P}\{|X| \geq a\} \leq \frac{\mathbb{E}(|X|)}{a}$.*

## 1.2 Coupon collectors

One concept we need in this thesis is the coupon collector, where $n$ different coupons exist, and the collector wants to collect them all. We assume that all coupons are equally likely to be acquired.
We can ask ourselves now: How many coupons does a collector need, to get every type?

We can see that this describes a Markov chain, as we can define the number of different coupons within the first t coupons $X_t$. We can see clearly $X_0 = 0$. When the collector has $k$ different types of coupons then $n - k$ coupons are missing, so of $n$ choices for a coupon, only $n - k$ will expand the collection. We get

$$\mathbb{P}\{X_{t+1} = k + 1 | X_t = k\} = \frac{n - k}{n},$$

$$\mathbb{P}\{X_{t+1} = k | X_t = k\} = \frac{k}{n}.$$

We can see that the chain is non-decreasing. When the collector has a complete collection, meaning the chain arrives at state $n$, then the chain will be absorbed, we can't have more types of coupons.
Our question changes to: How many steps do we need to reach the absorbing state?

**Proposition 1.3.** [1] *Let each new coupon be chosen uniformly and independently from the set of n possible types, and let $\tau$ be the (random) number of coupons collected, when the set first contains every type. For any $c > 0$,*

$$\mathbb{P}\{\tau > \lceil nlog(n) + cn \rceil\} \leq e^{-c}.$$

For more detailed and further information on coupon collectors, please refer to Feller [Fel91].

[1] Proposition 2.4 in [LP17].

# 2 Definition of Mixing Times

As stated above, we want to find the time that we need to reach stationary distribution, the mixing time. For this we need some definitions like the total variation distance and the standardizing distance.
Additionally we want to introduce the coalescence time, which we use for couplings. Lastly, we will find a relation between those times.

## 2.1 Total variation distance

Let us start with the first definition.
The **total variation distance** is defined as $||\mu - \nu||_{TV} = \max_{A \subseteq \mathcal{X}} |\mu(A) - \nu(A)|$, where $\mu$ and $\nu$ are two probability distributions in $\mathcal{X}$.

In words this means, that we can now calculate the difference of two distributions. For our use we will mostly use a probability distribution, defined through the transition matrix, and the stationary distribution. We can see this in the following example of the coin-tossing frog.

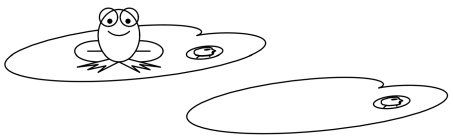

***Figure 1:*** *Frog sitting on a leaf with a coin. [LP17]*

**Example 2.1.** Imagine there is a frog sitting on a leaf, called $a$, next to another leaf, called $b$. The frog is holding a coin, which he is flipping every morning. If it lands on head, with probability $p$, the frog is jumping over to the other leaf and if it lands on tails, with probability $q$, the frog stays on the current leaf.
Therefore the transition matrix is $P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ and the stationary distribution is

$$\pi = (\frac{q}{p+q}, \frac{p}{p+q}).$$

Let us assume that the frog starts on the leaf $a$, which we can define as the starting state $\mu_0 = (1, 0)$. All following states can then be described as $\mu_t = \mu_{t-1}P$.
To calculate our total variation distance $||\mu_t - \pi||_{TV}$, we have to check, which states apply.
The four possible states $A \in \mathcal{X}$ are from $a \to a, a \to b, b \to a, b \to b$. The last two we do not need to consider as we have assumed that we start on the leaf $a$. We can see that the first two states are connected:

$$|P^t(a,a) - \pi(a)| = |(1 - P^t(a,b)) - (1 - \pi(b))| = |\pi(b) - P^t(a,b)| = |P^t(a,b) - \pi(b)|$$

The total variation distance becomes: $||\mu_t - \pi||_{TV} = |P^t(a,a) - \pi(a)|$.
Looking closer at $\mu_t(a) - \pi(a)$ we can see that it decreases exponentially fast as t increases:

$$\begin{aligned}
\mu_{t+1}(a) - \pi(a) &= (\mu_t(a)(1-p) + \mu_t(b)q) - \frac{q}{p+q} \\
&= (\mu_t(a)(1-p) + (1-\mu_t(a))q) - \frac{q}{p+q} \\
&= \mu_t(a)(1-p-q) + q - \frac{q}{p+q} \\
&= (1-p-q)(\mu_t(a) - \pi(a)) \\
&= (1-p-q)^{t+1}\mu_0
\end{aligned}$$

Therefore also the total variation distance decreases exponentially.

We can define the total variation distance differently, which can be useful in application, but is less intuitive.

**Proposition 2.2.** [2] *Let $\mu$ and $\nu$ be two probability distributions on $\mathcal{X}$. Then*

$$||\mu - \nu||_{TV} = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)|.$$

It follows by the same argument as in Proposition 2.2, that:

$$||\mu - \nu||_{TV} = \sum_{x\in\mathcal{X}, \mu(x)\geq\nu(x)} (\mu(x) - \nu(x)).$$

Using Proposition 2.2 we can show that the triangle inequality holds for the total variation distance:

$$\begin{aligned}
||\mu - \nu||_{TV} &= \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)| = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \eta(x) + \eta(x) - \nu(x)| \\
&\leq \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \eta(x)| + \frac{1}{2}\sum_{x\in\mathcal{X}} |\eta(x) - \nu(x)| = ||\mu - \eta||_{TV} + ||\eta - \nu||_{TV}
\end{aligned}$$

Lastly, we can write the total variation distance as follows:

**Proposition 2.3.** [3] *For $\mu$ and $\nu$ two probability distributions:*

$$||\mu - \nu||_{TV} = \frac{1}{2}\sup\{\sum_{x\in\mathcal{X}} f(x)\mu(x) - \sum_{x\in\mathcal{X}} f(x)\nu(x) : \max_{xin\mathcal{X}} |f(x)| \leq 1\}$$

[2] Proposition 4.2 in [LP17].
[3] Proposition 4.5 in [LP17].

## 2.2 Coupling

Another definition useful to know is coupling. This is practical to connect different Markov chains as we will see in Chapter 7. Here we will only show the connection to the total variation distance. However, there are many different applications of this technique.

A **coupling** of two probability distributions $\mu$ and $\nu$ is a pair of random variables $(X, Y)$ defined on a single probability space, so that the marginal distribution of $X$ is $\mu$ and of $Y$ is $\nu$. The marginal distribution is the probability distribution of the variables contained in the subset. Therefore a coupling $(X, Y)$ satisfies $\mathbb{P}\{X = x\} = \mu(x)$ and $\mathbb{P}\{Y = y\} = \nu(y)$.

**Example 2.4.** Let us take two fair-coins, meaning we get heads or tails with a probability of $\frac{1}{2}$. We can couple them, by defining $(X, Y)$ as a pair of independent coins such that we have $\mathbb{P}\{X = x, Y = y\} = \frac{1}{4} \ \ \forall x, y \in \{0, 1\}$.
Or we can couple them, by defining $X$ as the fair-coin toss and $Y = X$. We get the probabilities $\mathbb{P}\{X = Y = 0\} = \frac{1}{2}, \mathbb{P}\{X = Y = 1\} = \frac{1}{2}, \mathbb{P}\{X \neq Y\} = 0$.
Both couplings are valid, but have very different probabilities.

We can now connect the coupling with the total variation distance, through this Proposition.

**Proposition 2.5.** [4] *Let $\mu$ and $\nu$ be two probability distributions on $\mathcal{X}$. Then*

$$||\mu - \nu||_{TV} = \inf\{\mathbb{P}\{X \neq Y\} : (X, Y) \text{ is a coupling of } \mu \text{ and } \nu\}.$$

*The coupling $(X, Y)$ which realizes the infimum, is called **optimal**.*

Lastly, we will define the coupling of Markov chains. We do this to understand the different uses of coupling, but will not be further investigating this case.

A **coupling of Markov chains** with transition matrix $P$ is a process $(X_t, Y_t)_{t=0}^{\infty}$ with the property that both $(X_t)$ and $(Y_t)$ are Markov chains with the same transition matrix $P$. $(X_t)$ and $(Y_t)$ might have different starting distributions.

## 2.3 Standardizing Distance

We can define a mathematical tool, called the standardizing distance, that connects the total variation distance and the mixing time.

The **standardizing distance** is defined as: $d(t) = \max_{x \in \mathcal{X}} ||P^t(x, \cdot) - \pi||_{TV}$.
We also define: $\overline{d}(t) = \max_{x \in \mathcal{X}, y \in \mathcal{X}} ||P^t(x, \cdot) - P^t(y, \cdot)||_{TV}$.

[4] Proposition 4.7 in [LP17].

There are some useful results, which we can use afterwards.

**Proposition 2.6.** *If $d(t)$ and $\overline{d}(t)$ are as defined above, then $d(t) \leq \overline{d}(t) \leq 2d(t)$.*

*Proof.* Using the triangle inequality, we can see

$$\overline{d}(t) = \max_{x\in\mathcal{X}, y\in\mathcal{X}} ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} \leq \max_{x\in\mathcal{X}} ||P^t(x,\cdot) - \pi||_{TV} + \max_{y\in\mathcal{X}} ||\pi - P^t(y,\cdot)||_{TV} = 2d(t).$$

For the first inequality, we use the fact that $\pi$ is the stationary distribution, meaning $\pi(A) = \sum_{y\in\mathcal{X}} \pi(y)P^t(y,A)$. Then we get

$$\begin{aligned} |P^t(x,A) - \pi(A)| &= |P^t(x,A) - \sum_{y\in\mathcal{X}} \pi(y)P^t(y,A)| \\ &= |\sum_{y\in\mathcal{X}} \pi(y)(P^t(x,A) - P^t(y,A))| \\ &\leq \sum_{y\in\mathcal{X}} \pi(y)||P^t(x,\cdot) - P^t(y,\cdot)||_{TV}, \end{aligned}$$

$$d(t) \leq \sum_{y\in\mathcal{X}} \pi(y)\overline{d}(t) = \overline{d}(t).$$

We used $\sum_{y\in\mathcal{X}} \pi(y) = 1$, as well as taking the maximum over $A$ and $x$ and $y$. □

**Lemma 2.7.** *The function $\overline{d}$ is submultiplicative, which means that $\overline{d}(s+t) \leq \overline{d}(s)\overline{d}(t)$.*

*Proof.* Let us take $x, y \in \mathcal{X}$ and $(X_s, Y_s)$ an optimal coupling of $P^s(x,\cdot)$ and $P^s(y,\cdot)$, therefore $||\mu - \nu||_{TV} = \mathbb{P}\{X \neq Y\}$.

$$\Rightarrow ||P^s(x,\cdot) - P^s(y,\cdot)||_{TV} = \mathbb{P}\{X_s \neq Y_s\}$$

Using the definition of the mean, $\mathbb{E}(X) = \sum_x x\mathbb{P}(X = x)$, we get

$$P^{s+t}(x,y) = \sum_z P^s(x,z)P^t(z,y) = \sum_z \mathbb{P}\{X_s = z\}P^t(z,y) = \mathbb{E}(P^t(X_s,y)).$$

We can take the maximum over $x$ and $A$ with $y \in A$, and get:

$$\begin{aligned} \overline{d}(s+t) &= \max_{x\in\mathcal{X}} \max_{A\subseteq\mathcal{X}} |P^{s+t}(x,A) - P^{s+t}(y,A)| \\ &= \max_{x\in\mathcal{X}} \max_{A\subseteq\mathcal{X}} |\mathbb{E}(P^t(X_s,A) - P^t(Y_s,A))| \\ &\leq \mathbb{E}(\max_{x\in\mathcal{X}} \max_{A\subseteq\mathcal{X}} |P^t(X_s,A) - P^t(Y_s,A)|\mathbb{1}_{\{X_s\neq Y_s\}}) \\ &= \mathbb{E}(\overline{d}(t)\mathbb{1}_{\{X_s\neq Y_s\}}) \\ &= \mathbb{P}\{X_s \neq Y_s\}\overline{d}(t) \\ &\leq \max_{x\in\mathcal{X}} \mathbb{P}\{X_s \neq Y_s\}\overline{d}(t) \\ &= \overline{d}(s)\overline{d}(t) \end{aligned}$$

□

## 2.4 Mixing Times

Now, we are ready to define the mixing time which describes a parameter that measures the time required by a Markov chain for the standardizing distance to be small.

The **mixing time** is defined by $t_{mix}(\epsilon) = \min\{t : d(t) \leq \epsilon\}$ and $t_{mix} = t_{mix}(\frac{1}{4})$.

We can lastly define, for a coupling $\{(X_t, Y_t)\}$, the **coalescence time** of a chain to be $\tau_{couple} = \min\{t : X_s = Y_s \ \ \forall s \geq t\}$.

For the coalescence time we have a useful connection to the total variation distance.

**Theorem 2.8.** *Let $\{(X_t, Y_t)\}$ be a coupling satisfying if $X_s = Y_s$, then $X_t = Y_t \ \forall t \geq s$. Starting with $X_0 = x$ and $Y_0 = y$. Then*

$$||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} \leq \mathbb{P}_{xy}\{\tau_{couple} > t\}.$$

*Proof.* We can see that $P^t(x, z) = \mathbb{P}_{x,y}\{X_t = z\}$ and $P^t(x, z) = \mathbb{P}_{x,y}\{X_t = z\}$.
Therefore we have a coupling by definition ($\mathbb{P}\{X = x\} = \mu(x)$ in our case: $\mu = P^t(x, \cdot)$).
By Proposition 2.5 we get

$$||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} = \inf\{\mathbb{P}_{x,y}\{X_t \neq Y_t\}; \text{ for coupling } (X_t, Y_t)\}.$$

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

By the property of the coupling we get

$$\mathbb{P}_{x,y}\{X_t \neq Y_t\} = \mathbb{P}_{x,y}\{\tau_{couple} > t\},$$

which means that $(X_t)$ and $(Y_t)$ have not yet met.
Putting this together, we get our desired result:

$$||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} \leq \mathbb{P}_{xy}\{\tau_{couple} > t\}$$

□

Now we can put the mixing time and the coalescence time in relation to each other.

**Corollary 2.9.** *Suppose that for each pair of states $x, y \in \mathcal{X}$ there is a coupling $(X_t, Y_t)$ with $X_0 = x$ and $Y_0 = y$. For each coupling, let $\tau_{couple}$ be the coalescence time. Then*

$$d(t) \leq \max_{x,y \in \mathcal{X}} \mathbb{P}_{x,y}\{\tau_{couple} > t\}$$

*and therefore*

$$t_{mix} \leq 4 \cdot \max_{x,y \in \mathcal{X}} \mathbb{E}_{x,y}(\tau_{couple}).$$

*Proof.* We can combine Proposition 2.6 and 2.8 into

$$d(t) \leq \overline{d}(t) = \max_{x,y \in \mathcal{X}} ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} \leq \max_{x,y \in \mathcal{X}} \mathbb{P}_{xy}\{\tau_{couple} > t\}.$$

Per definition, we have $t_{mix} = t_{mix}(\frac{1}{4}) = \min\{t : d(t) \leq \frac{1}{4}\}$. By Markov's inequality

$$d(t) \leq \max_{x,y} \mathbb{P}_{x,y}\{\tau_{couple} > t\} \leq \max_{x,y} \mathbb{E}_{x,y}(\tau_{couple}) \cdot \frac{1}{t}.$$

Therefore if $t_{mix} \leq 4 \cdot \max_{x,y} \mathbb{E}_{x,y}(\tau_{couple})$ we get

$$d(t_{mix}) \leq \max_{x,y} \mathbb{E}_{x,y}(\tau_{couple}) \cdot \frac{1}{4 \cdot \max_{x,y} \mathbb{E}_{x,y}(\tau_{couple})} = 1/4.$$

□

# 3 Top to random shuffle

This section is the first example on how to bound the mixing time.

Let us introduce a specific way to shuffle a deck of $n$ cards: Each step, we take the top card of the deck and insert it uniform at random back in the deck. This will eventually mix up the deck.
We can interpret the successive arrangements of the deck as a random walk on the group $S_n$ of $n!$ possible permutations of the cards, which eventually have uniform distribution[5].
We can now ask ourselves, how long we have to shuffle to get close to random?

To answer this, we need some theory and definitions.

## 3.1 $\tau_{Top}$

Let $\tau_{Top}$ be the time of one shuffle after the original bottom card has moved to the top of the deck.

This is quite a useful definition, which we will need later, but first we have to consider an interesting result.

**Proposition 3.1.** [6] *Let $(X_t)$ be the random walk on $S_n$ corresponding to the top to random shuffle on $n$ cards. Given at time $t$ that there are $k$ cards under the original bottom card, each of the $k!$ possible orderings of these cards are equally likely. Then the distribution of $X_{\tau_{Top}}$ is uniform over $S_n$, and the time $\tau_{Top}$ is independent of $X_{\tau_{Top}}$.*

## 3.2 Stopping Times

Suppose that $\{X_t\}$ is adapted to the filtration $\{F_t\}$. We call $\{X_t\}$ a **Markov chain with respect to** $\{F_t\}$ if $\mathbb{P}_x\{X_{t+1} = y|F_t\} = P(X_t, y)$, where $P$ is a transition matrix. A "normal" Markov chain satisfies this when $\{F_t\}$ is the natural filtration.
If the filtration is not specified, we will assume that it is the natural filtration.

Recall that a **stopping time** for a filtration $\{F_t\}$ is a random variable $\tau$ with values in $\mathbb{N}_0 = \{0, 1, 2, ...\}$ such that $\{\tau = t\} \in F_t$.

**Example 3.2.** Going back to the top-to-random shuffle and $\tau_{Top}$, let $A$ be the set of arrangements having the original bottom card on top. Then we can see that $\tau_{Top} = \tau_A + 1$ is a stopping time.

## 3.3 Stationary Times and Strong Stationary Times

Let $(X_t)$ be an irreducible Markov chain with stationary distribution $\pi$. We assume that $\{F_t\}$ is a filtration and $\{X_t\}$ is adapted to $\{F_t\}$. A **stationary time** $\tau$ for $\{X_t\}$

[5] See Proposition 2.12 in [LP17].
[6] Proposition 6.1 in [LP17].

is an $\{F_t\}$-stopping time, possibly depending on the starting position $x$, such that the distribution of $\{X_\tau\}$ is $\pi$:

$$\mathbb{P}_x\{X_\tau = y\} = \pi(y) \quad \forall y.$$

Let us take a Markov chain $(X_t)$ with respect to the filtration $\{F_t\}$, with stationary distribution $\pi$. A **strong stationary time** for $(X_t)$ and starting position $x$ is an $\{F_t\}$-stopping time $\tau$, such that for all times $t$ and all $y$ the following holds:

$$\mathbb{P}_x\{\tau = t, X_\tau = y\} = \mathbb{P}_x\{\tau = t\}\pi(y)$$

This can interpreted as $X_\tau$ having distribution $\pi$ and being independent of $\tau$.

**Example 3.3.** In the top-to-random shuffle, we can see that $\tau_{Top}$ is a strong stationary time. Using Proposition 3.1, we get:

$$\mathbb{P}_x\{\tau_{Top} = t, X_{\tau_{Top}} = y\} = \mathbb{P}_x\{\tau_{Top} = t\}\mathbb{P}_x\{X_{\tau_{Top}} = y\} = \mathbb{P}_x\{\tau_{Top} = t\}\pi(y)$$

A more detailed introduction to strong stationary times can be found in Aldous and Diaconis [AD86] and [AD87]. More examples of strong stationary times are found in Pak [Pak97].

## 3.4 Separation Distance

The **separation distance** is defined as $s_x(t) = \max_{y\in\mathcal{X}}[1 - \frac{P^t(x,y)}{\pi(y)}]$.
Further we define $s(t) = \max_{x\in\mathcal{X}} s_x(t)$.

## 3.5 Finding a bound

Now, to find a bound of the top-to-random-shuffle, we need a few results.

**Lemma 3.4.** [7] *If $\tau$ is a strong stationary time for starting state $x$, then*

$$s_x(t) \leq \mathbb{P}_x\{\tau > t\}.$$

**Lemma 3.5.** [8] *The separation distance $s_x(t)$ satisfies $||P^t(x,\cdot) - \pi||_{TV} \leq s_x(t)$, and therefore $d(t) \leq s(t)$.*

With those results, we can combine this knowledge, as well as Proposition 1.3 to answer our initial question.

For our top-to-random shuffle, we have seen that $\tau_{Top}$ is a strong stationary time. If we look at the original bottom card and that there are $k$ cards below it, the chance to

[7] Lemma 6.12 in [LP17].
[8] Lemma 6.16 in [LP17].

rise the original bottom card by one is $\frac{(k+1)}{n}$, until a card is placed under it. So we can compare the distribution of $\tau_{Top}$ with the coupon collector's time.
We get by Lemma 3.5, Lemma 3.4 and Proposition 1.3

$$d(nlogn + cn) \leq s_x(nlogn + cn) \leq \mathbb{P}_x\{\tau > nlogn + cn\} \leq e^{-c}.$$

By setting $\epsilon = e^{-c}$ we get

$$t_{mix}(\epsilon) \leq nlogn + log(\epsilon^{-1})n.$$

Therefore we have found our bound for the mixing time of the top-to-random shuffle.

# 4 Bottleneck-Ratio

We have seen that we can find an upper bound of the mixing time. But how do we know that this is the best bound? We can further calculate a lower bound and compare these bounds.

To do this we need the concept of Bottlenecks.
**Bottlenecks** of a Markov chain are a geometric feature used to control mixing time, by making parts of the states difficult to reach from a starting position, which limits the speed of convergence.

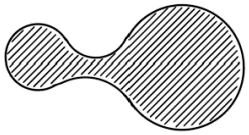

***Figure 2:*** *Graph with a Bottleneck. (Compare to [LP17])*

Let us first look at some definitions.
For this we assume that $P$ is an irreducible and aperiodic transition matrix for a Markov chain on state space $\mathcal{X}$ with stationary distribution $\pi$.

We define the **edge measure** $\mathcal{Q}$ as $\mathcal{Q}(x,y) = \pi(x)P(x,y)$ and $\mathcal{Q}(A,B) = \sum_{x\in A, y\in B} \mathcal{Q}(x,y)$.
This $\mathcal{Q}(A,B)$ is the probability of moving from the set A to the set B in one step, when starting from $\pi$.

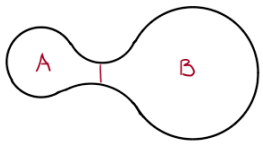


***Figure 3:*** *Graph with two sets.*

**Example 4.1.** We can calculate this edge measure for the simple random walk with vertices $\mathcal{X}$ and edge set $\mathcal{E}$.

$$\begin{aligned}\mathcal{Q}(x,y) &= \pi(x)P(x,y) \\ &= \begin{cases} \frac{deg(x)}{2|\mathcal{E}|} \cdot \frac{1}{deg(x)} = \frac{1}{2|\mathcal{E}|} & \text{if } \{x,y\} \text{ is an edge,} \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

The **bottleneck ratio** of the set $S$ is defined as $\Phi(S) = \frac{\mathcal{Q}(S,S^c)}{\pi(S)}$ and the bottleneck ratio of the whole chain is defined as $\Phi_* = \min_{S:\pi(S)\leq\frac{1}{2}} \Phi(S)$.

**Example 4.2.** For the simple random walk, we can see that $2|\mathcal{E}|\mathcal{Q}(S, S^c)$ is the size of the boundary $\partial S$ of $S$. This can be described as all the edges having one vertex in $S$ and one in $S^c$.
Then we can calculate the bottleneck ratio.

$$\Phi(S) = \frac{\mathcal{Q}(S, S^c)}{\pi(s)} = \frac{|\partial S|}{2|\mathcal{E}|\pi(s)} = \frac{|\partial S|}{2|\mathcal{E}|} \cdot \frac{2|\mathcal{E}|}{\sum\limits_{x\in S} deg(x)} = \frac{|\partial S|}{\sum\limits_{x\in S} deg(x)}$$

**Example 4.3.** If the simple random walk is lazy, the edge measure and bottleneck ratio are:

$$\mathcal{Q}(x, y) = \pi(x)P(x, y) = \frac{deg(x)}{2|\mathcal{E}|} \cdot \frac{1}{2deg(x)} \cdot \mathbb{1}_{\{\{x,y\}\ \in\ \mathcal{E}\}} = \frac{1}{4|\mathcal{E}|} \cdot \mathbb{1}_{\{\{x,y\}\ \in\ \mathcal{E}\}}$$

$$\Phi(S) = \frac{\mathcal{Q}(S, S^c)}{\pi(s)} = \frac{|\partial S|}{4|\mathcal{E}|\pi(s)} = \frac{|\partial S|}{4|\mathcal{E}|} \cdot \frac{2|\mathcal{E}|}{\sum\limits_{x\in S} deg(x)} = \frac{|\partial S|}{2\sum\limits_{x\in S} deg(x)}$$

If the graph of a simple random walk is regular with $deg = d$, then the bottleneck ratio becomes:

$$\Phi(S) = \frac{|\partial S|}{\sum\limits_{x\in S} deg(x)} = \frac{|\partial S|}{|S| \cdot d}$$

We can see that in this case the bottleneck ratio is proportional to the ratio of the size of the boundary and the volume of $S$.

We will now state a useful Theorem to connect the bottleneck ratio to the mixing time.

**Theorem 4.4.** *If $\Phi_*$ is the bottleneck ratio as defined above, then $t_{mix} = t_{mix}(\frac{1}{4}) \geq \frac{1}{4\Phi_*}$.*

*Proof.* Let us consider a stationary chain $\{X_t\}$, such that $X_0$ (and $X_t\ \forall t$) has distribution $\pi$.

$$\begin{aligned}
\Rightarrow \mathbb{P}_\pi\{X_0 \in A, X_t \in A^c\} &\leq \sum_{r=1}^{t} \mathbb{P}_\pi\{X_{r-1} \in A, X_r \in A^c\} && \text{by the triangle inequality} \\
&= t \cdot \mathbb{P}_\pi\{X_0 \in A, X_1 \in A^c\} && \text{all have the same distribution} \\
&= t \cdot \mathcal{Q}(A, A^c) && \text{per definition.}
\end{aligned}$$

Dividing by $\pi(A)$ shows that:

$$\mathbb{P}_\pi\{X_t \in A^c | X_0 \in A\} = \frac{\mathbb{P}_\pi\{X_0 \in A, X_t \in A^c\}}{\pi(A)} \leq \frac{t \cdot \mathcal{Q}(A, A^c)}{\pi(A)} = t \cdot \Phi(A).$$

So there exists an $x$ such that:

$$1 - P^t(x, A) \leq t \cdot \Phi(A) \Rightarrow P^t(x, A) \geq 1 - t \cdot \Phi(A).$$

Therefore

$$\begin{aligned} d(t) &= \max_{x \in \mathcal{X}, A \subset \mathcal{X}} |P^t(x, A) - \pi(A)| \\ &\geq P^t(x, A) - \pi(A) \\ &\geq 1 - t \cdot \Phi(A) - \pi(A). \end{aligned}$$

If $\pi(A) \leq \frac{1}{2}$ and $t < \frac{1}{4\Phi(A)}$, then

$$d(t) > 1 - \frac{1}{4\Phi(A)}\Phi(A) - \frac{1}{2} = \frac{1}{4}.$$

We can therefore see that $t_{mix} \geq \frac{1}{4\Phi(A)}$.
When we maximize over $A$ with $\pi(A) \leq \frac{1}{2}$, we get

$$t_{mix} \geq \frac{1}{4 \min\limits_{A:\pi(A) \leq \frac{1}{2}} \Phi(A)} = \frac{1}{4\Phi_*}.$$

And this completes our proof. □

Similar results can be found in Fill [Fil91], Mihail [Mih89] and Chen, Lovàsz and Pack [CLP99].

## 4.1 Binary Tree

Let us look at an example on how we can use the bottleneck ratio to find the lower bound of the mixing time, using the binary tree.
First a few definitions:

- A **tree** is a connected graph with no cycles.
- A **rooted** tree has a distinguished vertex, called the **root**.
- The **depth** of a vertex $v$ is its graph distance to the root.
- A **level** of the tree consists of all vertices at the same depth.
- The **children** of $v$ are the neighbours of $v$ with depth larger than $v$.
- A **leaf** is a vertex with degree one.

- A **rooted finite b-ary tree of depth k**, denoted by $T_{b,k}$, is a tree with a distinguished vertex $\rho$, the root, which satisfies that:
  - $\rho$ has degree b
  - every vertex at distance $j$ from the root, where $1 \leq j \leq k-1$, has degree $b+1$
  - the vertices at distance $k$ from $\rho$ are leaves.

  This tree has $n = \frac{(b^{k+1}-1)}{(b-1)}$ vertices and $n-1$ edges.

For our example we will now use a lazy walk on the rooted binary tree of depth $k$ ($b = 2$), with number of vertices $= n = 2^{k+1} - 1$, number of edges $= n - 1$.
Our goal is to show $t_{mix} \geq (n-2)\frac{1}{2}$.

For this we take $v_0$ to be our root and its neighbours right and left $v_r, v_l$ respectively, as seen below. Let us define a descendant $w$ of $v$ if the shortest path from $w$ to $v_0$ passes through $v$. We take $S$ to be the right-hand side of the tree, that is $v_r$ and all of it's descendants. And lastly, $|v|$ is the length of the shortest path from $v$ to $v_0$.

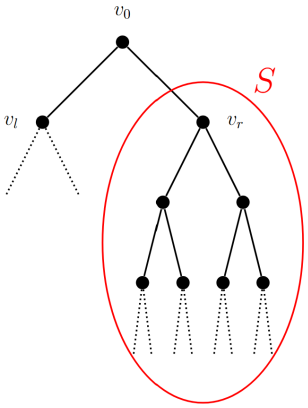


***Figure 4:*** *Binary Tree as described above.*

Because we are working on a simple random walk, we get for any vertex $y \in V$ we have $\pi(y) = \frac{deg(y)}{2|\mathcal{E}|}$. More specifically we get the stationary distribution

$$\pi(v) = \begin{cases} \frac{2}{2n-2} & \text{for } v = v_0, \\ \frac{3}{2n-2} & \text{for } 0 < |v| < k, \\ \frac{1}{2n-2} & \text{for } |v| = k. \end{cases}$$

Summing over $S$, we have $\frac{n-1}{2}$ vertices in $S$, from those we have $\frac{\frac{n-1}{2}-1}{2} = \frac{n-3}{4}$ vertices that are not leaves and $\frac{\frac{n-1}{2}+1}{2} = \frac{n+1}{4}$ vertices that are leaves. By using that for a binary tree of depth k, we have $2^k$ leafs and $2^{k+1} - 1$ vertices.

We get

$$\begin{aligned}\pi(S) &= \frac{n-3}{4}\frac{3}{2n-2} + \frac{n+1}{4}\frac{1}{2n-2} \\ &= \frac{1}{2n-2}(\frac{1}{4}(3n-9+n+1)) \\ &= \frac{n-2}{2n-2} \\ &\leq \frac{1}{2}.\end{aligned}$$

Because we only have one edge from $S$ to $S^c$ the edge measure of the lazy random walk simplifies to:

$$\begin{aligned}\mathcal{Q}(S, S^c) = \mathcal{Q}(v_r, v_0) &= \pi(v_r)P(v_r, v_0) \\ &= (\frac{3}{2n-2})\frac{1}{6} = \frac{1}{4(n-1)}\end{aligned}$$

$$\Rightarrow \Phi(S) = \frac{\mathcal{Q}(S, S^c)}{\pi(S)} = \frac{1}{4(n-1)}\frac{2(n-1)}{n-2} = \frac{1}{2(n-2)}.$$

By Theorem 4.4 we get the mixing time. We also use the fact that $\Phi_* \leq \Phi(S)$. Then:

$$t_{mix} \geq \frac{1}{4\Phi_*} \geq \frac{1}{4\Phi(S)} = \frac{2(n-2)}{4} = \frac{n-2}{2}.$$

So we were able to find the lower bound of the mixing time using the bottleneck ratio. In Section 5.3.4 of [LP17], we can find that $t_{mix} \leq 16n$. Therefore the mixing time must lay between $16n$ and $\frac{n-2}{2}$, which is a wide range.

# 5 Random Transposition

In this chapter, we will look at an example of random transposition, which describes a way for shuffling cards and define the upper bound of the mixing time.
But first we need to look at the theory of the symmetric group, which describes a general model for shuffling.

## 5.1 The Symmetric Group

A **permutation** of $\{1, 2, ..., n\}$ is a bijection from $\{1, 2, ..., n\}$ to itself. The set of all permutations form a group, called the **symmetric group** $S_n$, under the composition operation. The identity element of the group is the identity function $id(x) = x$. Every permutation has a well-defined inverse function, which is the inverse element in the group.

We can look at this as a mechanism of shuffling cards, by applying permutations $\sigma$ of $S_n$ with a probability of $\mu(\sigma)$ on a deck of cards. This is the same as running the random walk on the group with increment distribution $\mu$. Therefore the resulting chain is irreducible if the support of $\mu$ generates all of $S_n$[9]. If $\mu(id) > 0$, the chain will be aperiodic, as all states have period one. Lastly every shuffle chain has uniform stationary distribution.

**Example 5.1.** We can see that the following permutation can be seen as a shuffling of cards.

| i | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\sigma(i)$ | 3 | 1 | 2 | 4 |

This describes a shuffling where the first card lands on the third place, the second card lands on the first place, the third card lands on the second place and the fourth card stays at its place.

For permutations we can use the **cycle notation**. It is written as $(a\ b\ c)$ and describes the permutation $\sigma$ for which $b = \sigma(a)$, $c = \sigma(b)$, $a = \sigma(c)$. They are read right to left, when written consecutively.

**Example 5.2.**

$$(1\ 2)(2\ 3)(3\ 4)(2\ 3)(1\ 2) = (1\ 4)$$

We can break this down for each number:

[9] See Proposition 2.13 in [LP17].

$$\sigma(1) : 1 \to 2 \to 3 \to 4 \to 4 \to 4$$
$$\sigma(2) : 2 \to 1 \to 1 \to 1 \to 1 \to 2$$
$$\sigma(3) : 3 \to 3 \to 2 \to 2 \to 3 \to 3$$
$$\sigma(4) : 4 \to 4 \to 4 \to 3 \to 2 \to 1$$

Which is the same as just once exchanging 1 and 4.

Every permutation can be written as a product of disjoint cycles.
A cycle of length $n$ is called an **$n$-cycle**. A 2-cycle is called a **transposition**. Fixed points are 1-cycles, which are not written specifically.

Using those cycles we can now generate an exactly uniform random permutation. For this we can use the following algorithm:
Set $\sigma_0$ to be the identity permutation. Inductively for $k = 1, 2, .., n-1$ construct $\sigma_k$ from $\sigma_{k-1}$, by swapping the cards $k$ and $J_k$, with $J_k$ being an integer chosen uniformly in $\{k, ..., n\}$ and independently of $\{J_1, ..., J_{k-1}\}$. Which leaves us with

$$\sigma_k(i) = \begin{cases} \sigma_{k-1}(i) & \text{if } i \neq J_k, i \neq k, \\ \sigma_{k-1}(J_k) & \text{if } i = k, \\ \sigma_{k-1}(k) & \text{if } i = J_k. \end{cases} \tag{1}$$

Which is $\sigma_k = \sigma_{k-1} \circ (k\ J_k)$.

For any permutation $\sigma$ we can we can define the sign of the product to be

$$\mathcal{M} = \prod_{1 \leq i < j \leq n} (\sigma(j) - \sigma(i)).$$

For the identity, we get $\mathcal{M}(id) > 0$, because every term is positive. For all $\sigma \in S_n$ and transposition $(a\ b)$, we can see that $\mathcal{M}(\sigma \circ (a\ b)) = -\mathcal{M}(\sigma)$. Without loss of generality we can assume that $a < b$, then for all $c$ such that $a < c < b$ two factors change sign ($a \leftrightarrow c$ and $a \leftrightarrow b$) as well as $a \leftrightarrow b$. So if there are $n$ numbers between $a$ and $b$, we will have an odd number of sign changes, which leads to a sign change for the permutation.

With this we can define the parity of a permutation:
A permutation $\sigma$ is **even** if $\mathcal{M}(\sigma) > 0$, which means that every way of writing it as a product of transpositions contains an even number of factors.
A permutation $\sigma$ is **odd** if it is not even, which means that every way of writing it as a product of transpositions contains an odd number of factors.
All even permutations form a subgroup of $S_n$, known as the **alternating group** $A_n$.

We can also define the parity of $n$-cycles, using that it can be written as a product of $n-1$ transpositions: $(a_1 ... a_n) = (a_1\ a_2) ... (a_{n-1}\ a_n)$
An $n$-cycle is **even (odd)** if $n$ is odd (even).

The sign of any permutation is determined by its disjoint cycle decomposition.

**Example 5.3.** Let us take $T$ to be the set of all transpositions, with $T \subseteq S_n$, and let $\mu$ be the uniform probability distribution on $T$. Using the algorithm above, we can see that $< T >= S_n$, because we have only used the identity and transpositions in the algorithm to build $S_n$. It follows that our chain is irreducible.

Since every element of the support of $\mu$ is odd (since they are 2-cycles), if the walk started with identity, we would end on an odd (even) position after an odd (even) number of steps. Hence the walk is periodic.

Periodicity occurs in random walks on groups, when the entire support of the increment distribution falls into a single coset of some subgroup. This means, that the walk will not be able to leave that coset.
It follows that we need to assume $\mu(id) > 0$ to ensure aperiodicity.

## 5.2 Upper bound

We want to find an upper bound for the mixing time of the following chain:
Consider the random transposition shuffle in (1) started from the identity permutation. To avoid periodicity we choose at time $t$ two cards $L_t, R_t$ independently and uniformly at random. If they are the same card, do nothing, otherwise exchange them. To this the distribution $\mu$ is

$$\mu(\sigma) = \begin{cases} 1/n & \text{if } \sigma = id, \\ 2/n^2 & \text{if } \sigma = (i\ j), \\ 0 & \text{otherwise.} \end{cases}$$

This chain is aperiodic and is irreducible as in Example 5.3.

We can also look at this shuffling in another way: At each time $t$ we choose a card $X_t$ and a position $Y_t$ independently. We then transpose the chosen card to the chosen position. If the card is already at this position, we do nothing.
We want to couple two decks to find the upper bound. To couple the decks we use the same choices of $(X_t)$ and $(Y_t)$ to shuffle both decks. Let there be two trajectories $(\sigma)$ and $(\sigma')$ and let $a_t$ be the number of cards that have the same position in $(\sigma)$ and $(\sigma')$.

First, if the card $X_t$ is in the same position in $(\sigma)$ and $(\sigma')$, then $a_{t+1} = a_t$.
If $X_t$ is in a different position in $(\sigma)$ and $(\sigma')$ and the card at position $Y_t$ is the same in $(\sigma)$ and $(\sigma')$, we loose one alignment but also create a new one. Therefore $a_{t+1} = a_t$.
Lastly, if $X_t$ is in a different position in $(\sigma)$ and $(\sigma')$ and the card at position $Y_t$ is different in $(\sigma)$ and $(\sigma')$, at least one new alignment is created. It is possible to create up to 3 new alignments, which we will consider in example 5.4

**Example 5.4.** Here a few options for possible new alignments.
One alignment:

$$\begin{array}{cccc} 2 & 3 & 1 & 4 \\ 4 & 1 & 3 & 2 \end{array} \xrightarrow[Y_t=1]{X_t=1} \begin{array}{cccc} 1 & 3 & 2 & 4 \\ 1 & 4 & 3 & 2 \end{array}$$

Two alignments:

$$\begin{array}{cccc} 2 & 3 & 1 & 4 \\ 3 & 1 & 4 & 2 \end{array} \xrightarrow[Y_t=1]{X_t=1} \begin{array}{cccc} 1 & 3 & 2 & 4 \\ 1 & 3 & 4 & 2 \end{array}$$

Three alignments:

$$\begin{array}{ccc} 2 & 3 & 1 \\ 3 & 1 & 2 \end{array} \xrightarrow[Y_t=1]{X_t=1} \begin{array}{ccc} 1 & 3 & 2 \\ 1 & 3 & 2 \end{array}$$

To be able to conclude what the upper bound for the mixing time is, we need the following result.

**Proposition 5.5.** *Let $\tau$ be the time required for the two decks to coincide. Then, no matter the initial configurations of the decks,*

$$\mathbb{E}(\tau) < \frac{\pi^2}{6} n^2.$$

*Proof.* Let us break $\tau$ into transpositions between the first time that $a_t \geq i-1$ and the first time that $a_t \geq i$, denoted by $\tau_i$:

$$\tau = \tau_1 + ... + \tau_n$$

$\tau_i$ is allowed to be 0, as $a_0$ can be bigger than 0 and $a_t$ can increase by more than 1.

Let us look at two cases:

Case 1: When for at least one $t$ it holds that $a_t = i$:
We have $n-i$ unaligned cards and the probability that the number of aligned cards increases is $\frac{(n-i)^2}{n^2}$, because we would have to choose one of the unaligned cards and position. In this case $\tau_{i+1}$ is a geometric random variable with probability $\frac{(n-i)^2}{n^2}$. It follows that the mean is $\mathbb{E}(\tau_{i+1}|a_t = i) = \frac{n^2}{(n-i)^2}$.

Case 2: When for no $t$ it holds that $a_t = i$:
Then we have $\tau_{i+1} = 0$.

We can combine these cases to get our desired result.

$$\begin{aligned}
\mathbb{E}(\tau) &< \sum_{i=0}^{n-1} \frac{n^2}{(n-i)^2} && \text{since we start from 0 and not from 1,} \\
&< n^2 \sum_{l=1}^{\infty} \frac{1}{l^2} && \text{for } (n-i) \in (0, n-1) \text{ we get } l \in (1, \infty), \\
&= n^2 \frac{\pi^2}{6} && \text{by the Basel Problem.}
\end{aligned}$$

□

By using Corollary 2.9, we can conclude:

$$t_{mix} \leq 4 \max \mathbb{E}_{xy}(\tau) < \frac{2}{3}\pi^2 n^2.$$

This gives us the desired upper bound for the transposition shuffle. There is an even better bound for this mixing time, that we will not discuss here, but is found in Chapter 8.2.3 of [LP17].

# 6 Reversible Markov Chains

In this section we are going to look at reversible Markov chains and at the properties of the mixing time in regard to the reversed Markov chain.

## 6.1 Reversible Markov chains and Time-Reversal

A Markov chain is **reversible** if the chain satisfies the **detailed balance equation**:

$$\pi(x)P(x,y) = \pi(y)P(y,x) \quad \forall x, y \in \mathcal{X},$$

with $\pi$ being the unique stationary distribution of $P$.

Now, for a general chain we can define the **time-reversed chain** to have the transition matrix $\widehat{P}$ where $\pi(x)P(x,y) = \pi(y)\widehat{P}(y,x)$ so that the chain is reversible if and only if $\widehat{P} = P$.
The stationary $\widehat{P}$ chain equates to the $P$ chain run backwards in time.
The stationary $\widehat{P}$ chain equates to the $P$ chain run backwards in time.

We can see that some results about reversible chains are just results about general chains using both $P$ and $\widehat{P}$.
As an example we can look at the cyclic tour property with $T_i = \min\{t \geq 0 : X_t = i\}$.

**Lemma 6.1.** *For the states $x_0, x_1, ..., x_m$ of a reversible chain*

$$\mathbb{E}_{x_0}(T_{x_1}) + \mathbb{E}_{x_1}(T_{x_2}) + ... + \mathbb{E}_{x_m}(T_{x_0}) = \mathbb{E}_{x_0}(T_{x_m}) + \mathbb{E}_{x_m}(T_{x_{m-1}}) + ... + \mathbb{E}_{x_1}(T_{x_0}).$$

*Proof.* For a general chain we have

$$\mathbb{E}_{x_0}(T_{x_1}) + \mathbb{E}_{x_1}(T_{x_2}) + ... + \mathbb{E}_{x_m}(T_{x_0}) = \widehat{\mathbb{E}_{x_0}}(T_{x_1}) + \widehat{\mathbb{E}_{x_1}}(T_{x_2}) + ... + \widehat{\mathbb{E}_{x_m}}(T_{x_0}), \qquad (2)$$

with $\widehat{\mathbb{E}}$ being the expectation of the time-reversed chain $\widehat{\mathbb{P}}$.
So, for the reversible chain we immediately get the desired result.

The equation (2) is quite intuitive, like imagining how a movie is played backwards, frame by frame. For a more detailed explanation see the notes in [Fil14]. □

## 6.2 Reversed Distribution

For a distribution $\mu$ on a group $G$, the **reversed distribution** $\widehat{\mu}$ is defined by

$$\widehat{\mu}(g) = \mu(g^{-1}) \quad \forall g \in G.$$

Let $P$ be the transition matrix of the random walk with increment distribution $\mu$. Then the random walk with increment distribution $\widehat{\mu}$ is exactly the time reversal

$$\widehat{P}(x,y) = \frac{\pi(y)P(y,x)}{\pi(x)},$$

with $\pi$ being the stationary distribution.
This is our result, when $\mu = \widehat{\mu}$, the random walk on $G$ with increment distribution $\mu$ is reversible and $P = \widehat{P}$.

We can state a nice Corollary about mixing times of random walks using the following Lemma.

**Lemma 6.2.** [10] *Let $P$ be the transition matrix of a random walk on a group $G$ with increment distribution $\mu$ and let $\widehat{P}$ be that of the walk on $G$ with increment distribution $\widehat{\mu}$. Let $\pi$ be the uniform distribution on $G$. Then for any $t \geq 0$:*

$$||P^t(id, \cdot) - \pi||_{TV} = ||\widehat{P}^t(id, \cdot) - \pi||_{TV}.$$

**Corollary 6.3.** *If $t_{mix}$ is the mixing time of a random walk on a group and $\widehat{t}_{mix}$ is the mixing time of the reversed walk, then*

$$t_{mix} = \widehat{t}_{mix}.$$

*Proof.* We have the definitions

$$t_{mix} = \min\{t : d(t) \leq \frac{1}{4}\} = \min\{t : \max_{x \in \mathcal{X}} ||P^t(x, \cdot) - \pi||_{TV} \leq \frac{1}{4}\}$$

$$\widehat{t}_{mix} = \min\{t : \max_{x \in \mathcal{X}} ||\widehat{P}^t(x, \cdot) - \pi||_{TV} \leq \frac{1}{4}\}.$$

By Lemma 6.2 we can directly conclude that:

$$t_{mix} = \widehat{t}_{mix}.$$

□

## 6.3 Example

If we consider a process other than a random walk, we can have very different mixing times for the chain and its time reversal.

As an example of such a chain, we can look at the winning streak chain with window n. First let us look at the normal chain:
Let us imagine a creature with bounded memory tossing a fair coin (probability 1/2 to land on each side) repeatedly and trying to track the length of the last run of heads. The creature can only remember up to n tosses. This implies that the length of the last run of heads is at most $n$.

***Figure 5:*** *Creature with bounded memory[yet24].*

We can imagine the tosses, as an n bit word $b = (b_1, ...b_n)$, where

[10] Lemma 4.13 in [LP17].

1 represents head and 0 represents tail. For the length of the block of 1's starting at the right-most bit of b we can write

$$r(b) = \begin{cases} 0 & \text{if } b_n = 0, \\ \max\{m : b_{n-m+1} = ... = b_n = 1\} & \text{otherwise.} \end{cases}$$

Take arbitrary initial bits $B_{-n+1}, ..., B_0$ and independent fair bits $B_1, B_2, ...$ and let $X_t = r(B_{t-n+1}, B_{t-n+2}, ..., B_t)$.
We have created a Markov chain with state space $\{0, ...n\}$ and non-zero transitions given by

$$P(i, 0) = \frac{1}{2} \text{ for } 0 \leq i \leq n, \ P(n, n) = \frac{1}{2}, \ P(i, i+1) = \frac{1}{2} \text{ for } 0 \leq i \leq n.$$

We start at a state $a$, meaning that we fix initial bits with $r(B_{-n+1}, ..., B_0) = a$.

**Example 6.4.** Let us look at a short example with $n = 7$, and $X_t = 3$, $X_{t+1} = 4$, $X_{t+2} = 0$ and $X_{t+3} = 1$.[a]

| | | |
|---|---|---|
| Time $t$ | : | $\boxed{0\,0\,1\,0\,1\,1\,1}\,1\,0\,1$ |
| Time $t+1$ | : | $0\,\boxed{0\,1\,0\,1\,1\,1\,1}\,0\,1$ |
| Time $t+2$ | : | $0\,0\,\boxed{1\,0\,1\,1\,1\,1\,0}\,1$ |
| Time $t+3$ | : | $0\,0\,1\,\boxed{0\,1\,1\,1\,1\,0\,1}$ |

We can see that the value of $X_t$ can only increase by 1, stay at $n = 7$ or fall back onto 0, when $t$ increases. This can be displayed in the graph

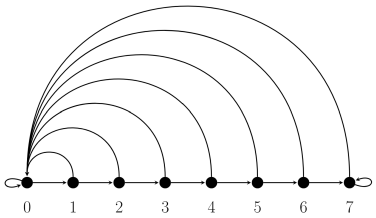


***Figure 6:*** *Number of heads in a row $X_t$. (Compare to [LP17])*

[a] Similar example in [LP17].

We will quickly confirm that

$$\pi(i) = \begin{cases} \frac{1}{2^{i+1}} & \text{if } i = 0, 1, .., n-1, \\ \frac{1}{2^n} & \text{if } i = n, \end{cases}$$

is stationary for $P$.

$$\pi P = \pi : (\pi(0), ..., \pi(n)) \begin{pmatrix} \text{P(0,0)} & ... & \text{P(0,n)} \\ ... & ... & ... \\ \text{P(n,0)} & ... & \text{P(n,n)} \end{pmatrix} = (\pi(0), ..., \pi(n))$$

$$\Rightarrow \pi(0)P(0,0) + ... + \pi(n)P(n,0) = \frac{1}{2} \cdot \frac{1}{2} + ... + \frac{1}{2^n} \cdot \frac{1}{2} = \frac{1}{2} = \pi(0)$$

$$\Rightarrow \pi(0)P(0,i) + ... + \pi(i-1)P(i-1,i) + ... + \pi(n)P(n,i) = \frac{1}{2^i} \cdot \frac{1}{2} = \frac{1}{2^{i+1}} = \pi(i),$$

because only one $P(i-1,i)$ is not 0.

$$\Rightarrow \pi(0)P(0,n) + ... + \pi(n-1)P(n-1,n) + ... + \pi(n)P(n,n) = \frac{1}{2^n} \cdot \frac{1}{2} + \frac{1}{2^n} \cdot \frac{1}{2} = \frac{1}{2^n} = \pi(n)$$

Therefore, $\pi(i)$ is stationary.

Now, we can fix a pair of initial values $a, b \in \{0, ..., n\}$ and create two chains $(X_t^x)$, $(Y_t^y)$ and couple them by coupling the two bit streams $(B_t^x)_{t=-n+1}^{\infty}$, $(B_t^y)_{t=-n+1}^{\infty}$, so we have

$$X_t = r(B_{t-n+1}^x, B_{t-n+2}^x, ..., B_t^x),\ Y_t = r(B_{t-n+1}^y, B_{t-n+2}^y, ..., B_t^y).$$

Let $x, y$ be bit strings of length $a, b$: $x = (B_{-n+1}^x, ..., B_0^x), y = (B_{-n+1}^y, ..., B_0^y)$. For $t \geq 1$ we will set $B_t^x = B_t^y = B_t$, with $(B_t)_{t=0}^{\infty}$ is an independent and identically distributed (iid) fair bit sequence. The first time an added bit is 0, both chains fall into state 0 and stay coupled from then on.

We can now calculate the bound on $t_{mix}$.

$$\Rightarrow \mathbb{P}\{\tau_{couple} > t\} \leq P(0,1) \cdot ... \cdot P(t-1,t) = 2^{-t}$$

By Corollary 2.9

$$d(t) \leq 2^{-t}.$$

Using this with the definition of $t_{mix}$ and setting $\epsilon = 2^{-t}$, we get

$$t_{mix}(\epsilon) \leq \lceil log_2 \left(\frac{1}{\epsilon}\right) \rceil.$$

This only depends on $\epsilon$, so we can compute:

$$t_{mix} = t_{mix}\left(\frac{1}{4}\right) \leq \lceil log_2(4) \rceil = 2$$

Now, let us look at the time reversed chain.
By definition $\widehat{P}$ has nonzero entries:

$$\widehat{P}(x,y) = \frac{\pi(y)P(y,x)}{\pi(x)}.$$

So we can calculate the transition matrix .

$$\widehat{P}(0,i) = \frac{\pi(i)P(i,0)}{\pi(0)} = \pi(i)\frac{1/2}{1/2} = \pi(i) \text{ for } 0 \leq i \leq n$$

$$\widehat{P}(n,n) = \frac{\pi(n)P(n,n)}{\pi(n)} = 1/2$$

$$\widehat{P}(n,n-1) = \frac{\pi(n-1)P(n-1,n)}{\pi(n)} = 1/2$$

$$\widehat{P}(i,i-1) = \frac{\pi(i-1)P(i-1,i)}{\pi(i)} = \frac{1}{2^i}2^{(i+1)}\frac{1}{2} = 1 \text{ for } 1 \leq i \leq n$$

We can still have the same interpretation of a coin flip as in the normal chain. However this time we have a window of width n moving left along a string of independent random bits. Then the sequence of the lengths $(\widehat{X}_t)$ of the rightmost block of 1's in the window is a version of the reversed chain.

**Example 6.5.** Take the same example as above, but this time we have $\widehat{X}_t = 1$, $\widehat{X}_{t+1} = 0$, $\widehat{X}_{t+2} = 4$ and $\widehat{X}_{t+3} = 3$.[a]

| | |
|---|---|
| Time $t$ : | 0 0 1 [0 1 1 1 1 0 1] |
| Time $t+1$ : | 0 0 [1 0 1 1 1 1 0] 1 |
| Time $t+2$ : | 0 [0 1 0 1 1 1 1] 0 1 |
| Time $t+3$ : | [0 0 1 0 1 1 1] 1 0 1 |

This time $X_t$ will decrease by one until we reach 0, then we will jump to a random number between 0 and $n = 7$. Except at $X_t = 7$, then we have a probability of $\frac{1}{2}$ for it to stay at 7. Described in a graph:

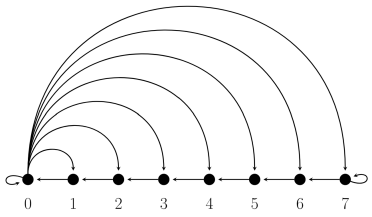


***Figure 7:*** *Number of heads in a row $X_t$ for the reversed chain. (Compare to [LP17])*

[a] Similar example in [LP17].

We can see that if $(B_t)_{t=-\infty}^{n-1}$ are iid bits, then $\widehat{X}_t = r(B_{-t}, ..., B_{-t+n-1})$. In particular $\widehat{X}_n = r(B_{-n}, ..., B_{-1})$, which is uniform over all $2^n$ bitwords of length n, independent of initial bits.

We would like to show now that the distribution of $\widehat{X}_t$ is exactly stationary, regardless of initial distribution, after $n$ steps. Meaning that the mixing time will be $n$.

If we consider our chain starts at step $n$. On the first move, it moves to $n-1$ with probability $\frac{1}{2}$. After n-1 moves, it moves to state 1.

$$\begin{aligned} \Rightarrow \widehat{P}^{n-1}(n,1) &= \widehat{P}(n,n-1) \cdot \widehat{P}(n-1,n-2) \cdot ... \cdot \widehat{P}(2,1) \\ &= \frac{1}{2} \cdot 1 \cdot .... \cdot 1 = \frac{1}{2} \end{aligned}$$

By definition

$$\begin{aligned} \widehat{d}(n-1) &= \max_{x \in \mathcal{X}} ||\widehat{P}^{n-1}(x,\cdot) - \pi||_{TV} \\ &= \max_{x \in \mathcal{X}} \max_{A \subseteq \mathcal{X}} |\widehat{P}^{n-1}(x,A) - \pi(A)| \\ &\geq |\widehat{P}^{n-1}(n,1) - \pi(1)| = |\frac{1}{2} - \frac{1}{2^2}| = \frac{1}{4}. \end{aligned}$$

We have to take one more step away from 1 for it to be stationary.
By definition

$$\widehat{t}_{mix}(\epsilon) = n \quad \forall \epsilon < \frac{1}{4}.$$

We were able to see in this example, that there are Markov chains that have significantly different mixing times to their reversed chains. For the Markov chain $t_{mix}(\epsilon) = 2$ and for the reversed chain $\widehat{t}_{mix}(\epsilon) = n$.
In contrast we showed in Chapter 6.2, that for the random walk, we receive the same mixing time.

Please refer to the books of Kelly [Kel11] and Keilson [Kei79] for further reference.

# 7 Graphs

In this last section, we will dive into the theory of graphs. On a graph we can use the random walk as the Markov chain. We will shortly see the connection to electrical networks. And end with the last example for bounding mixing time on the Metropolis chain.

## 7.1 Graphs and Random Walks

A **graph** is a construct, consisting of a finite vertex set $\mathcal{V} = \{v, x, y, ...\}$ and an edge set $\mathcal{E} = \{\epsilon_1, \epsilon_2, ...\}$. It is connected and undirected, has no multiple edges and no self-loops.

In a **weighted graph**, each edge $(v, w)$ also has a weight $0 < \omega_{v,x} = \omega_{x,v} < \infty$. In this graph we allow for self-loops.

On the weighted graph, we can define a Markov chain, with the convention, that an absent edge has weight 0.
A discrete **random walk** on a weighted graph is a Markov chain with transition matrix

$$P(v, x) = \omega_{v,x}/\omega_v \quad x \neq v,$$

with $\omega_v = \sum_x \omega_{v,x}$.

We call $\omega = \sum_v \omega_v$ the total edge-weight, with each edge counted twice (in each direction).
The chain will automatically be reversible with the stationary distribution

$$\pi_v = \frac{\omega_v}{\omega}.$$

$$\Rightarrow \pi_v p_{vx} = \frac{w_v}{w}\frac{w_{vx}}{w_v} = \frac{w_{xv}}{w}\frac{w_x}{w_x} = \pi_x p_{xv}$$

We can see that the chain is irreducible, because we have defined the graph to be connected.
For the chain to be aperiodic, it is necessary that the graph is not bipartite. A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets U and V, meaning, every edge connects a vertex in U to one in V.

An **unweighted graph** is a graph with each weight equal to 1.
The transition matrix becomes

$$P(v, x) = \begin{cases} \frac{1}{d_v} & \text{if } (v, x) \text{ is an edge,} \\ 0 & \text{if not,} \end{cases},$$

with $d_v$ being the degree of vertex $v$.

And the stationary distribution becomes

$$\pi_v = \frac{d_v}{2|\mathcal{E}|},$$

with $|\mathcal{E}|$ being the number of edges of the graph.

For an **unweighted regular graph** (each vertex has the same number of edges) we have the stationary distribution

$$\pi_v = \frac{1}{n}.$$

For those different graphs we can see that the mean hitting times take the following form.

**Lemma 7.1.** [11] *For a random walk on an n-vertex graph, with* $T_i^+ = \min\{t \geq 1 : X_t = i\}$, *we get*

$$\mathbb{E}_v(T_v^+) = \begin{cases} \frac{\omega}{\omega_v} & \text{if weighted,} \\ \frac{2|\mathcal{E}|}{d_v} & \text{if unweighted,} \\ n & \text{if unweighted regular.} \end{cases}$$

**Lemma 7.2.** [12] *Given an edge* $(v, x)$, *and* $T = \min\{t \geq 1 : X_t = v, X_{t-1} = x\}$, *we have*

$$\mathbb{E}_{\mathcal{V}}(T) = \begin{cases} \frac{\omega}{\omega_{v,x}} & \text{if weighted,} \\ 2|\mathcal{E}| & \text{if unweighted.} \end{cases}$$

**Example 7.3.** Let us look at two glued graphs and find an upper bound for its mixing time.

Take two complete graphs with $n$ vertices, and glue them together at a vertex $v^*$. We denote the new graph as $\mathcal{G}$.

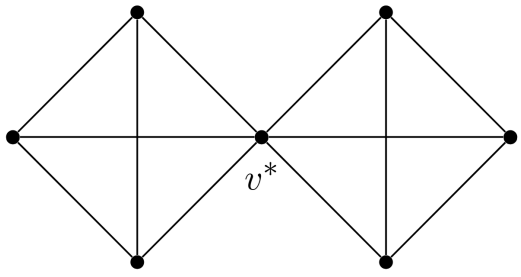


***Figure 8:*** *Two complete graphs with 4 vertices, glued at vertex* $v^*$.

We can see that the vertex $v^*$ has a degree of $2n - 2$, and every other vertex has a degree of $n - 1$.

[11] Lemma 3.5 in [Fil14].
[12] Lemma 3.7 in [Fil14].

We want the graph $\mathcal{G}$ to be regular, so we have to add one loop at vertex $v^*$ and $n$ loops at any other vertex.

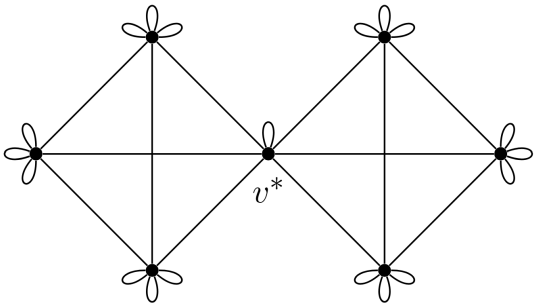


***Figure 9:*** *Two glued graphs at vertex $v^*$ made regular.*

The uniform distribution of the random walk on the graph $\mathcal{G}$ is stationary, because it is regular with degree $2n-1$.
It is clear to see that if we are at $v^*$ each vertex is equally likely to be hit. So we take $\tau_{v^*}$ to be the hitting time of $v^*$, then $\tau = \tau_{v^*} + 1$ is a strong stationary time.
If we are not at $v^*$, the probability of hitting $v^*$ next is $\frac{1}{2n-1}$. Then $\tau_{v^*}$ is geometric with parameter $\frac{1}{2n-1}$.
We can calculate

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

Therefore, we can see for $t = 4n$

$$\mathbb{P}_x\{\tau > t\} \leq e^{-\frac{t-1}{2n-1}} \leq e^{-2}.$$

We can conclude that $t_{mix} \leq 4n$.

## 7.2 Fluid model and electrical networks

We will take a short detour to look at graphs in different contexts, with fluids and electricity.

We will first look at the **fluid model**, which describes a way of defining a continuous time chain on a weighted graph.

Let us consider having a finite number of buckets that hold quantities of fluids. Some of those buckets are connected by tubes.
If bucket $i$, with $q_i$ quantities of fluid, and bucket $j$, with $q_j$ quantities of fluid, are connected, the flow rate is proportional to the pressure difference, more precisely $\omega_{ij}(q_i - q_j)$ from bucket $i$ to bucket $j$, with a parameter $\omega_{ij} = \omega_{ji}$. We will ignore the fluid in the tubes for our concept. We can see that the quantities of fluid will evolve with time $t$

according to

$$\frac{dq_i(t)}{dt} = \sum_{i \neq j} \omega_{ij}(q_i - q_j),$$

called the **forward equation** of $q_i(t)$ for the continuous time chain with transition rates $p_{ij} = \omega_{ij}$, if $i \neq j$. This is the probability of being at position $i$ at time $t$.

If we look at the time when $t$ goes to infinity, we are able to see intuitively, that the fluid will distribute itself uniformly over all buckets, meaning that the stationary distribution of this model is always uniform. Using a bigger parameter $\omega_{ij}$ the fluid will settle faster.
All this makes sense in the fluid model, however for the usual discrete Markov chain and its continuation this does not hold.

Before we can compare the fluid model to the electrical network, we need some definitions.

A **flow** $\mathbf{f} = (f_{ij})$ on a graph satisfies:

$$f_{ij} = \begin{cases} -f_{ji} & \text{if } (i,j) \text{ is an edge,} \\ 0 & \text{otherwise.} \end{cases}$$

We can define the **net flow** away from $i$ as:

$$f_{(i)} = \sum_{j \neq i} f_{ij}.$$

and by symmetry we get $\sum_i f_{(i)} = 0$.

On disjoint nonempty subsets $A, B$ of vertices, we can define a **unit flow from B to A** as a flow satisfying

$$\sum_{i \in B} f_{(i)} = 1; \;\; f_{(j)} = 0 \;\; \forall j \notin A \cup B.$$

This leads to

$$\sum_{i \in A} f_{(i)} = -1.$$

Using the random walk on a weighted graph, we can define a special flow. Given $v_0 \notin A$ and $T_A = \min\{t \geq 0 : X_t \in A\}$, the unit flow from $v_0$ to $A$ $\mathbf{f}^{v_0 \to A}$ is defined by

$$f_{ij} = \mathbb{E}_{v_0} \sum_{t=1}^{T_A} (\mathbb{1}_{(X_{t-1}=i, X_t=j)} - \mathbb{1}_{(X_{t-1}=j, X_t=i)}).$$

This describes the difference between the mean number of transitions $i \to j$ and $j \to i$ for a chain that started at $v_0$ and ended, when it first hit A. This works as well in continuous time.

This model can be used in networks:

A **network** is a finite undirected connected graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$, endowed additionally with non-negative numbers $\{c(e)\}$ called **conductances**, that are associated to the edges of $G$. It is obvious that $c(x, y) = c(y, x)$.
The **resistance** of the edge $e$ is defined by $r(e) = \frac{1}{c(e)}$.

An **electrical network** is a weighted graph together with a function $g$, called voltage, and a flow $I$, called current. This is constructed the following way:
Take a weighted graph, if a wire is connecting two vertices $x$ and $y$ it will have conductance $\omega_{xy}$ and resistance $\frac{1}{\omega_{xy}}$.
We fix a vertex $v_0$ and a subset $A$ of vertices without $v_0$. We apply voltage 1 at $v_0$ and voltage 0 at the vertices of $A$. We get the voltage $g(v)$ for those vertices

$$g(v_0) = 1, \quad g(v) = 0 \ \forall v \in A. \tag{3}$$

To determine the **current**, we can use Ohm's law:

$$\text{Current} = \frac{\text{Potential difference}}{\text{Resistance}}$$

In our case for one wire

$$\Rightarrow I_{xy} = (g(x) - g(y))\omega_{xy}. \tag{4}$$

We can see that $I = (I_{xy})$ is a flow.
By Kirchoff's law, we get

$$I_{(v)} = 0, \ v \notin \{v_0\} \cup A. \tag{5}$$

(3), (4), (5) imply for $v \notin \{v_0\} \cup A$

$$g(v) = \sum_x p_{vx} g(x),$$

with $p_{vx}$ the transition matrix for a weighted graph defined in Chapter 7.1. This specifies $g$ uniquely.

With this we can see that a weighted graph uniquely determines both a random walk (Chapter 7.1) and an electrical network. Therefore we can identify the voltage and the current in terms of the random walk. We can summarise this in a Proposition.

**Proposition 7.4.** [13] *Consider a weighted graph as an electrical network, where a wire linking $x$ and $y$ has conductance $\omega_{xy}$. Suppose that the voltage function $g$ satisfies $g(v_0) = 1$, $g(v) = 0 \quad \forall v \in A$. Then the voltage at any vertex $v$ is given in terms of the associated random walk by*

$$g(v) = \mathbb{P}_v(T_{v_0} < T_A) \in [0, 1]$$

*and the current $I_{xy}$ along each wire $(x, y)$ is $\frac{f_{xy}}{r}$, with $\mathbf{f} = f^{v_0 \to A}$ and*

$$r = \frac{1}{\omega_{v_0} \mathbb{P}_{v_0}(T_A < T^+_{v_0})} \in (0, \infty).$$

Since we know that $\mathbf{f}$ is a unit flow from $v_0$ to $A$ and $g(v_0) = 1$, it follows that $I_{(v_0)} = g(v_0)/r$.

[13] Proposition 3.10 in [LP17]

## 7.3 Grand coupling and random colourings

Before we can go over to bound the mixing time, we need two concepts: grand coupling and random colouring.

A **grand coupling** is a collection of random variables $\{X_t^x : x \in \mathcal{X}, t = 0, 1, 2, ...\}$ such that for each $x \in \mathcal{X}$, the sequence $\{X_t^x\}$ is a Markov chain with transition matrix $P$ started from $x$.

A **random mapping representation** of $P$ on the state space $\mathcal{X}$ is a function $f : \mathcal{X} \times \Lambda \to \mathcal{X}$, along with a $\Lambda$-valued random variable $Z$, satisfying $\mathbb{P}\{f(x, z) = y\} = P(x, y)$.
We can see that $(X_n) = f(X_{n-1}, Z_n)$ is a Markov chain with transition matrix $P$ and initial distribution $\mu$, if $Z_1, Z_2, ...$ is a sequence of independent random variables, with the same distribution as $Z$.

Using this, we can build a grand coupling:
Take a function $f : \mathcal{X} \times \Lambda \to \mathcal{X}$ and a $\Lambda$-valued random variable $Z$ such that $\mathbb{P}\{f(x, z) = y\} = P(x, y)$. It exists, by Proposition 1.5 in [LP17]. Then take $Z_1, Z_2, ...$ a sequence of independent and identically distributed random variables, with the same distribution as $Z$. Define $X_0^x = x,\ X_t^x = f(X_{t-1}^x, Z_t)$ for $t \geq 1$.
So, $(X_t^x)_{t=0}^{\infty}$ is a Markov chain starting at $x$, with transition matrix $P$, and we have created a grand coupling.

Given a graph, we can label all the vertices using colours and by that adding limitations on this graph. More formally, we can define the following.
Fix a set $\{1, 2, ..., q\}$ of number of colours. A **proper q-colouring** of a graph $G = (\mathcal{V}, \mathcal{E})$ is an assignment of colours to each vertex in $\mathcal{V}$, but without neighbour vertices having the same colour.

More generally, we can define a **configuration** as a function from $\mathcal{V}$ to $S$, with $\mathcal{V}$ being the vertex set of a graph and $S$ is a finite set. This can be interpreted as a labeling of vertices with elements in $S$. We can see, that colouring is an example of a configuration.

## 7.4 Example

The **Metropolis chain** for a probability $\pi$ and a symmetric transition matrix $\Psi$ is defined as

$$P(x, y) = \begin{cases} \Psi(x, y)[1 \wedge \frac{\pi(y)}{\pi(x)}] & \text{if } y \neq x, \\ 1 - \sum_{z: z \neq x} \Psi(x, z)[1 \wedge \frac{\pi(z)}{\pi(x)}] & \text{if } y = x. \end{cases}$$

Then $\pi$ is a stationary distribution of the Metropolis chain.
The transition matrix of the Metropolis chain is constructed by selecting a vertex $v$ uniformly from $\mathcal{V}$ and then selecting a colour $k$ uniformly from $\{1, 2, ..., q\}$. If no neighbour of $v$ has colour $k$, then the original colour of $v$ is changed to $k$, otherwise nothing changes.

Now, we have all information necessary to bound the mixing time of the Metropolis

chain.

**Theorem 7.5.** *Let $G$ be a graph with $n$ vertices and maximal degree $\Delta$. For the Metropolis chain on proper colourings of $G$, if $q > 3\Delta$ and $c_{met}(\Delta, q) = 1 - (3\Delta/q)$, then*

$$t_{mix}(\epsilon) \leq c_{met}(\Delta, q)^{-1} n \left( logn + log\left(\frac{1}{\epsilon}\right)\right).$$

*Proof.* Similar to the Metropolis chain, the grand coupling creates a pair $(v, c)$ consisting of a vertex and a colour, uniformly from $\mathcal{V} \times \{1, ..., q\}$ and independent from the past. As described before for each vertex $x \in \mathcal{X}$, $X_t^x$ is updated to the colour $k$ if no neighbour has that colour, otherwise it stays the same.
We will use the same vertex and same colour for all Markov chains $(X_t^x)$.
So for two colourings $x, y \in \mathcal{X}$ we can define the number of vertices, where $x$ and $y$ are different:

$$\rho(x, y) = \sum_{v \in \mathcal{V}} \mathbb{1}_{\{x(v) \neq y(v)\}}$$

Now we will first look at the case when $\rho(x, y) = 1$.
In this case, $x$ and $y$ agree everywhere except at one vertex, we call it $v_0$. We can define a set $\mathcal{N}$ of all colours of the neighbours of $v_0$ in $x$. In the case below: $\mathcal{N} = \{blue, red, yellow\}$. This is the same set as the set of all colours of the neighbours of $v_0$ in $y$.

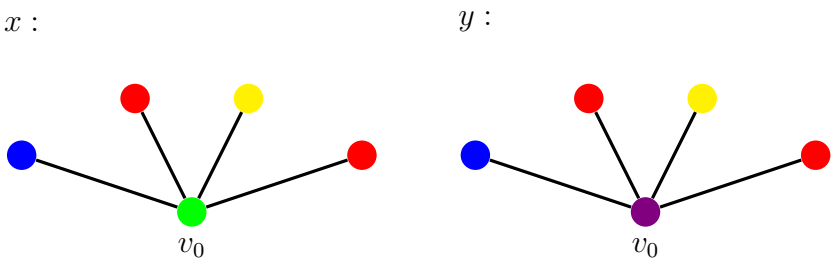


***Figure 10:*** *Two colourings with different colours for vertex $v_0$.*

After updating a colour we have two possibilities, either $v_0$ is selected, with a colour not in $\mathcal{N}$ or a neighbour is selected.

Let us first assume that $v_0$ is selected and that the new colour is not in $\mathcal{N}$.

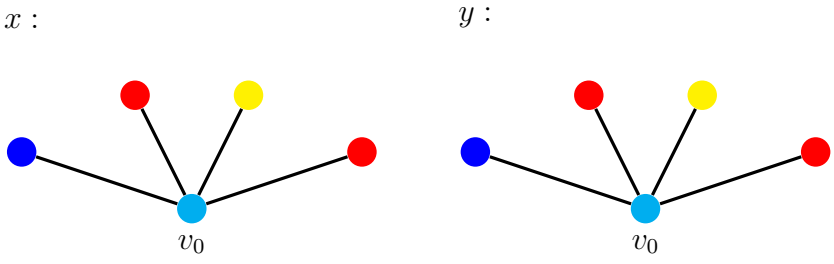


***Figure 11:*** *Two colourings with now the same colours for vertex $v_0$.*

We can calculate

$$\mathbb{P}\{\rho(X_1^x, X_1^y) = 0\} = \left(\frac{1}{n}\right)\left(\frac{q - |\mathcal{N}|}{q}\right) \geq \frac{q - \Delta}{nq}.$$

The first equation is given by the probability of choosing this particular vertex times the number of colours we can use over the number of colours in general. The inequality is given because the number of neighbours (= number of colours of neighbours) is at most the maximal degree: $|\mathcal{N}| \leq \Delta$.

Let us now assume that a neighbour $\omega$ of $v_0$ is selected. If the new colour is $x(v_0)$ or $y(v_0)$, then $\rho$ increases by at most 1.

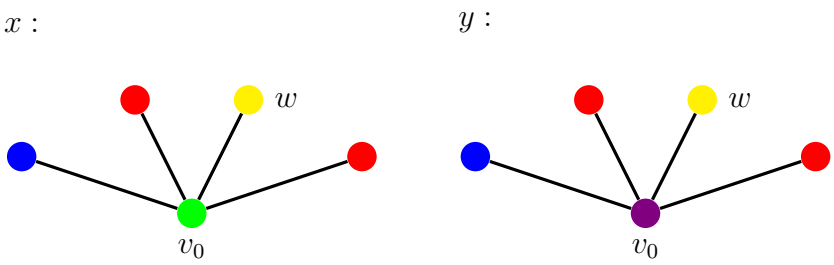


***Figure 12:*** *Two different colourings with chosen vertex $w$.*

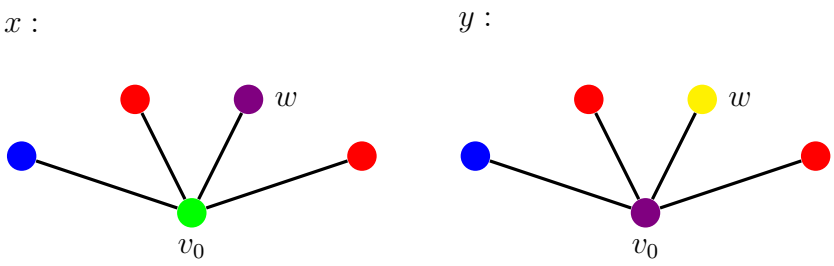


***Figure 13:*** *Two different colourings with chosen vertex $w$ and chosen color violet only accepted by $x$.*

If the new colour is neither $x(v_0)$ nor $y(v_0)$, then the new colour will be accepted in $x$ and in $y$.

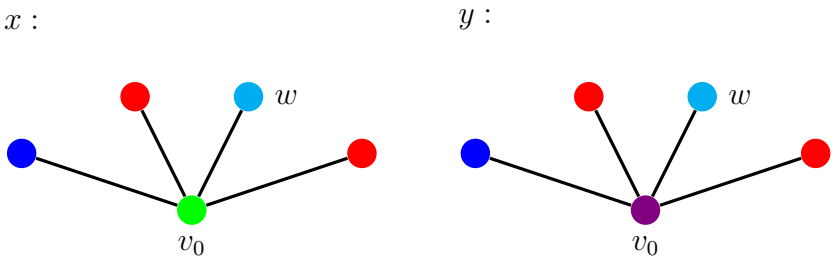


***Figure 14:*** *Two different colourings with chosen vertex $w$ and chosen color blue accepted both colorings.*

Therefore we will get a new disagreement if and only if a neighbour of $v_0$ is selected and the colour is either $x(v_0)$ or $y(v_0)$. Therefore we get:

$$\mathbb{P}\{\rho(X_1^x, X_1^y) = 2\} \leq \left(\frac{\Delta}{n}\right)\left(\frac{2}{q}\right).$$

This is given by the maximum of possible neighbours over the possibility for one neighbour times the two possible colours $x(v_0), y(v_0)$ over all possible colours.

Given those two bounds we can calculate:

$$\begin{aligned}\mathbb{E}(\rho(X_1^x, X_1^y) - 1) &= \mathbb{P}\{\rho(X_1^x, X_1^y) - 1 = -1\} \cdot (-1) + \mathbb{P}\{\rho(X_1^x, X_1^y) - 1 = 1\} \cdot (1) \\ &\leq -\frac{q-\Delta}{nq} + \frac{2\Delta}{nq} = \frac{3\Delta - q}{nq}\end{aligned}$$

$$\Rightarrow \mathbb{E}(\rho(X_1^x, X_1^y)) \leq 1 - (-\frac{3\Delta - q}{nq}) = 1 - \frac{1 - \frac{3\Delta}{q}}{n} = 1 - \frac{c_{met(\Delta,q)}}{n},$$

by definition of $c_{met}$.
Assuming $q > 3\Delta$, it holds that:

$$\mathbb{E}(\rho(X_1^x, X_1^y)) \leq 1 - \frac{c_{met(\Delta,q)}}{n} < 1,$$

given the case that $\rho(x, y) = 1$.

Now, we will look at the case $\rho(x, y) = r$ for $r \in \mathbb{N}_0$.
We have $r$ colourings $x_0 = x, x_1, ..., x_k, ..., x_r = y$ with the property that $\rho(x_k, x_{k+1}) = 1$. Using the case and discussion from above, we have $\mathbb{E}(\rho(X_1^{x_k}, X_1^{x_{k-1}})) \leq 1 - \frac{c_{met(\Delta,q)}}{n}$, because $\rho$ is a metric. We can calculate

$$\begin{aligned}\mathbb{E}(\rho(X_1^x, X_1^y)) &\leq \sum_{k=1}^{r} \mathbb{E}(\rho(X_1^{x_k}, X_1^{x_{k-1}})) \leq \sum_{k=1}^{r} \left(1 - \frac{c_{met(\Delta,q)}}{n}\right) \\ &= r\left(1 - \frac{c_{met(\Delta,q)}}{n}\right) = \rho(x, y)\left(1 - \frac{c_{met(\Delta,q)}}{n}\right).\end{aligned}$$

If we condition over the event that $X_{t-1}^x = x_{t-1}$ and $X_{t-1}^y = y_{t-1}$, then the random vector $(X_t^x, X_t^y)$ has the same distribution as $(X_1^{x_{t-1}}, X_1^{y_{t-1}})$.

$$\mathbb{E}(\rho(X_t^x, X_t^y) | X_{t-1}^x = x_{t-1}, X_{t-1}^y = y_{t-1}) = \mathbb{E}(\rho(X_1^{x-1}, X_1^{y-1})) \leq \rho(x_{t-1}, y_{t-1})\left(1 - \frac{c_{met(\Delta,q)}}{n}\right)$$

Taking the expectation over $(X_{t-1}^x, X_{t-1}^y)$

$$\begin{aligned}\mathbb{E}(\rho(X_t^x, X_t^y) &= \mathbb{E}_{(X_{t-1}^x, X_{t-1}^y)}(\rho(X_t^x, X_t^y) | X_{t-1}^x = x_{t-1}, X_{t-1}^y = y_{t-1}) \\ &\leq \mathbb{E}_{(X_{t-1}^x, X_{t-1}^y)}(\rho(x_{t-1}, y_{t-1}))\left(1 - \frac{c_{met(\Delta,q)}}{n}\right) \\ &\leq \rho(X_{t-1}^x, X_{t-1}^y)\left(1 - \frac{c_{met(\Delta,q)}}{n}\right).\end{aligned}$$

By iteration

$$\mathbb{E}(\rho(X_t^x, X_t^y)) \leq \rho(x,y)\left(1 - \frac{c_{met(\Delta,q)}}{n}\right)^t.$$

Now we can use Markov's inequality to get

$$\begin{aligned}
\mathbb{P}\{X_t^x \neq X_t^y\} &= \mathbb{P}\{\rho(X_t^x, X_t^y) \geq 1\} \\
&\leq \frac{1}{1}\mathbb{E}(\rho(X_t^x, X_t^y)) \\
&\leq \rho(x,y)\left(1 - \frac{c_{met(\Delta,q)}}{n}\right)^t \\
&= \rho(x,y)\left(\left(1 + \frac{-c_{met(\Delta,q)}}{n}\right)^n\right)^{t/n} \\
&\leq \rho(x,y)e^{-t\cdot\frac{c_{met(\Delta,q)}}{n}} \\
&\leq ne^{-t\cdot\frac{c_{met(\Delta,q)}}{n}}.
\end{aligned}$$

For the second to last inequality we have used the fact that $(1 + \frac{x}{n})^n$ converges to $e^x$ as $n$ goes to infinity. In the last inequality we have used that $\rho$ is at most $n$, if all vertices are different.

Since this holds for all colourings $x, y \in \mathcal{X}$, we can apply Corollary 2.9 and $\mathbb{P}_{x,y}\{X_t \neq Y_t\}$ $= \mathbb{P}_{x,y}\{\tau_{couple} > t\}$ from Theorem 2.8.

$$\Rightarrow d(t) \leq ne^{-t\cdot\frac{c_{met(\Delta,q)}}{n}}.$$

If we take $t_{mix}(\epsilon) \leq c_{met}(\Delta,q)^{-1}n(logn + log(\frac{1}{\epsilon}))$, then:

$$d(t_{mix}(\epsilon)) \leq ne^{-\cdot\frac{c_{met(\Delta,q)}}{n}\cdot c_{met}(\Delta,q)^{-1}n(logn+log(\frac{1}{\epsilon}))} = ne^{-logn}e^{-log(1/\epsilon)} = \epsilon$$

This concludes that $c_{met}(\Delta,q)^{-1}n(logn+log(\frac{1}{\epsilon}))$ is indeed the upper bound of the mixing time.

□

# 8 Conclusion

We were able to show that we can bound the mixing time for various different chains. Starting with the top to random shuffle, giving us $t_{mix}(\epsilon) \leq nlog(n) + log(\epsilon^{-1})n$. Next, we were able to see that if we have the bottleneck ratio, we can calculate the mixing time like: $t_{mix} \geq \frac{1}{4\Phi_*}$. Specifically for a lazy walk on a binary tree we get $t_{mix} \geq \frac{n-2}{2}$. For the random transportation shuffle, we were able to see the upper bound of $t_{mix} \leq \frac{2}{3}\pi^2 n^2$. When looking at the reversible Markov chains, we have received two different results. On the one hand, we saw that for a random walk on a group the mixing time of the random walk is the same as the mixing time for the reversed walk $t_{mix} = \widehat{t}_{mix}$. On the other hand, we saw an example of the winning streak chain, where the mixing time of the normal chain and the reversed chain are different. Lastly, we looked at two different models that we can connect to a graph: The fluid model and electrical network. We were able to connect the electrical network to a weighted graph with a random walk. In the end we were able to see that for the metropolis chain on proper colourings of a graph the mixing time can be bounded by $t_{mix}(\epsilon) \leq c_{met}(\Delta, q)^{-1} n(logn + log(\frac{1}{\epsilon}))$.

I hope you have found this topic as fascinating and interesting as I have. I would recommend to everyone who has liked it, to continue reading 'Markov Chains and Mixing Times' [LP17] to see more examples and applications of mixing times.

# References

[AD86] David Aldous and Persi Diaconis. Shuffling cards and stopping times. *The American Mathematical Monthly*, 93(5):333–348, 1986.

[AD87] David Aldous and Persi Diaconis. Strong uniform times and finite random walks. *Advances in Applied Mathematics*, 8(1):69–97, 1987.

[CLP99] Fang Chen, László Lovász, and Igor Pak. Lifting markov chains to speed up mixing. In *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, pages 275–281, 1999.

[Fel91] William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 81. John Wiley & Sons, 1991.

[Fil91] James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible markov chains, with an application to the exclusion process. *The annals of applied probability*, pages 62–87, 1991.

[Fil14] David Aldous; James Allen Fill. *Reversible Markov Chains and Random Walks on Graphs*. UC Berkley statistics departement, 2014.

[Hä02] O. Häggström. *Finite Markov chains and algorithmic applications*. London Mathematical Society Student Texts, vol. 52, Cambridge University Press, Cambridge, 2002.

[Jer03] Mark Jerrum. *Counting, sampling and integrating: algorithms and complexity*. Springer Science & Business Media, 2003.

[Kei79] Julian Keilson. Rarity and exponentiality. In *Markov Chain Models—Rarity and Exponentiality*, pages 130–163. Springer, 1979.

[Kel11] Frank P Kelly. *Reversibility and stochastic networks*. Cambridge University Press, 2011.

[Lov93] László Lovász. Random walks on graphs. *Combinatorics, Paul erdos is eighty*, 2(1-46):4, 1993.

[LP17] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[Mih89] Milena Mihail. Conductance and convergence of markov chains-a combinatorial treatment of expanders. In *FOCS*, volume 89, pages 526–531, 1989.

[MT+06] Ravi Montenegro, Prasad Tetali, et al. Mathematical aspects of mixing times in markov chains. *Foundations and Trends® in Theoretical Computer Science*, 1(3):237–354, 2006.

[Pak97] Igor Pak. *Random walks on groups: strong uniform time approach*. Harvard University, 1997.

[yet24] `https://ch.pinterest.com/pin/98305204342170870/`, 19.08.24.