# Explore Markov Chains with Examples


Have you ever wondered how Google ranks web pages? If you've done your research then you must know that it uses the PageRank Algorithm which is based on the idea of Markov chains. In 1998, Lawrence Page, Sergey Brin, Rajeev Motwani and Terry Winograd published "The PageRank Citation Ranking: Bringing Order to the Web", an article in which they introduced the now famous PageRank algorithm at the origin of Google. A little bit more than two decades later, Google has became a giant and, even if the algorithm has evolved a lot, the PageRank is still a "symbol" of the Google ranking algorithm (even if few people can really say the weight it still occupies in the algorithm).

From a theoretical point of view, it is interesting to notice that one common interpretation of the PageRank algorithm relies on the simple but fundamental mathematical notion of Markov chains. We will see in this unit that Markov chains are powerful tools for stochastic modelling that can be useful to any data scientist. More especially, we will answer basic questions such as: what are Markov chains, what good properties do they have and what can be done with them?

In this Unit on Introduction To Markov Chains will help you understand the basic idea behind Markov chains and how they can be modelled as a solution to real-world problems.

What is A Markov Chain?
What is The Markov Property?
What is A Transition Matrix?
What is A Markov Chain?

Andrey Markov first introduced Markov chains in the year 1906. He explained Markov chains as:
A stochastic process containing random variables, transitioning from one state to another depending on certain assumptions and definite probabilistic rules.
These random variables transition from one to state to the other, based on an important mathematical property called Markov Property.
This brings us to the question:

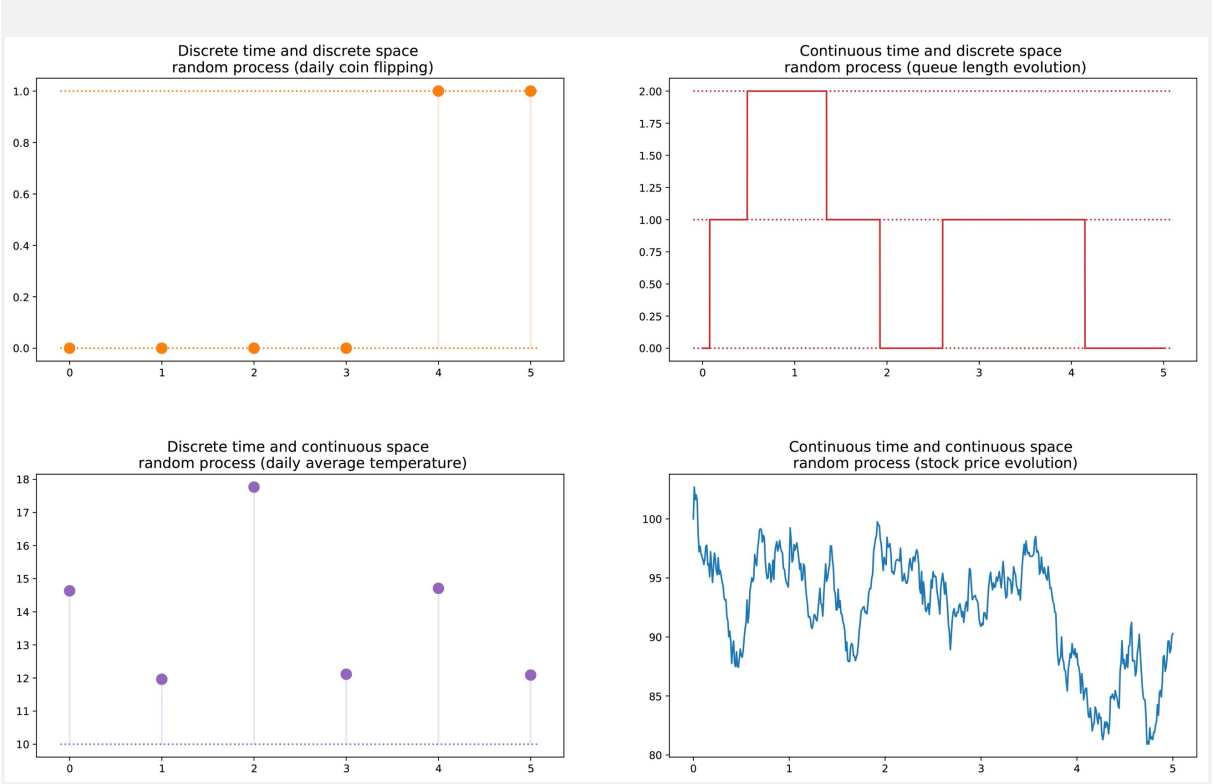
**What are Markov chains?**

**Random variables and random processes**

Before introducing Markov chains, let's start with a quick reminder of some basic but important notions of probability theory.

First, in non-mathematical terms, a random variable X is a variable whose value is defined as the outcome of a random phenomenon. This outcome can be a number (or "number-like", including vectors) or not. For example we can define a random variable as the outcome of rolling a dice (number) as well as the output of flipping a coin (not a number, unless you assign, for example, 0 to head and 1 to tail). Notice also that the space of possible outcomes of a random variable can be discrete or continuous: for example, a normal random variable is continuous whereas a poisson random variable is discrete.

We can then define a random process (also called stochastic process) or probabilistic process as a collection of random variables indexed by a set T that often represent different instants of time (we will assume that in the following). The two most common cases are: either T is the set of natural numbers (discrete time random process) or T is the set of real numbers (continuous time random process). For example, flipping a coin every day defines a discrete time random process whereas the price of a stock market option varying continuously defines a continuous time random process. The random variables at different instant of time can be independent to each other (coin flipping example) or dependent in some way (stock price example) as well as they can have continuous or discrete state space (space of possible outcomes at each instant of time).

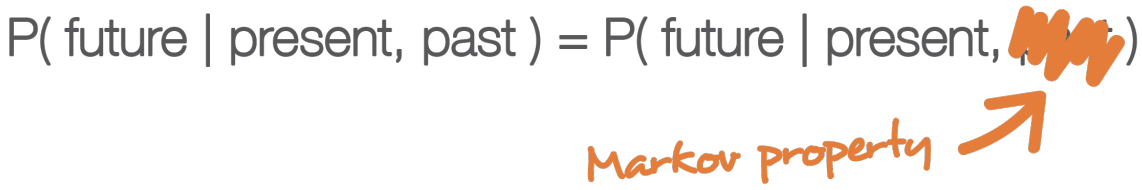



Different kind of random processes (discrete/continuous in space/time).

## Markov Property and Markov Chains

There exists some well known families of random processes: gaussian processes, poisson processes, autoregressive models, moving-average models, Markov chains and others. These particular cases have, each, specific properties that allow us to better study and understand them.

One property that makes the study of a random process much easier is the "Markov property". In a very informal way, the Markov property says, for a random process, that if we know the value taken by the process at a given time, we won't get any additional information about the future behaviour of the process by gathering more knowledge about the past. Stated in slightly more mathematical terms, for any given time, the conditional distribution of future states of the process given present and past states depends only on the present state and not at all on the past states (memoryless property). A random process with the Markov property is called Markov process.




The Markov property expresses the fact that at a given time step and knowing the current state, we won't get any additional information about the future by gathering information about the past.

Based on the previous definition, we can now define "homogenous discrete time Markov chains" (that will be denoted "Markov chains" for simplicity in the following). A Markov chain is a Markov process with discrete time and discrete state space. So, a Markov chain is a discrete sequence of states, each drawn from a discrete state space (finite or not), and that follows the Markov property.

Mathematically, we can denote a Markov chain by

$$X = (X_n)_{n \in \mathbb{N}} = (X_0, X_1, X_2, ...)$$

where at each instant of time the process takes its values in a discrete set E such that

$$X_n \in E \qquad \forall n \in \mathbb{N}$$

Then, the Markov property implies that we have

$$\mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n, X_{n-1} = s_{n-1}, X_{n-2} = s_{n-2}, ...) = \mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n)$$

Notice once again that this last formula expresses the fact that for a given historic (where I am now and where I was before), the probability distribution for the next state (where I go next) only depends on the current state and not on the past states.

Note. We have decided to describe only basic homogenous discrete time Markov chains in this introductory post. However, there also exists inhomogenous (time dependent) and/or time continuous Markov chains. We won't discuss these variants of the model in the following. Notice also that the definition of the Markov property given above is extremely simplified: the true mathematical definition involves the notion of filtration that is far beyond the scope of this modest introduction.

**Characterising the random dynamic of a Markov chain**

We have introduced in the previous subsection a general framework matched by any Markov chain. Let's now see what we do need in order to define a specific "instance" of such a random process.

Notice first that the full characterisation of a discrete time random process that doesn't verify the Markov property can be painful: the probability distribution at a given time can depend on one or multiple instants of time in the past and/or the future. All these possible time dependences make any proper description of the process potentially difficult.

However, thanks to the Markov property, the dynamic of a Markov chain is pretty easy to define. Indeed, we only need to specify two things: an initial probability distribution (that is a probability distribution for the instant of time n=0) denoted

$$\mathbb{P}(X_0 = s) = q_0(s) \qquad \forall s \in E$$

and a transition probability kernel (that gives the probabilities that a state, at time n+1, succeeds to another, at time n, for any pair of states) denoted

$$\mathbb{P}(X_{n+1} = s_{n+1} | X_n = s_n) = p(s_n, s_{n+1}) \qquad \forall (s_{n+1}, s_n) \in E \times E$$

With the previous two objects known, the full (probabilistic) dynamic of the process is well defined. Indeed, the probability of any realisation of the process can then be computed in a recurrent way.

Assume for example that we want to know the probability for the first 3 states of the process to be (s0, s1, s2). So, we want to compute the probability

$$\mathbb{P}(X_0 = s_0, X_1 = s_1, X_2 = s_2)$$

Here, we use the law of total probability stating that the probability of having (s0, s1, s2) is equal to the probability of having first s0, multiplied by the probability of having s1 given we had s0 before, multiplied by the probability of having finally s2 given that we had, in order, s0 and s1 before. Mathematically, it can be written

$$\mathbb{P}(X_0 = s_0, X_1 = s_1, X_2 = s_2) = \mathbb{P}(X_0 = s_0)\mathbb{P}(X_1 = s_1 | X_0 = s_0)\mathbb{P}(X_2 = s_2 | X_0 = s_0, X_1 = s_1)$$

Then appears the simplification given by the Markov assumption. Indeed, for long chains we would obtain for the last states heavily conditional probabilities. However, in a Markov case we can simplify this expression using that

$$\mathbb{P}(X_2 = s_2 | X_0 = s_0, X_1 = s_1) = \mathbb{P}(X_2 = s_2 | X_1 = s_1)$$

such that we have

$$\begin{aligned}\mathbb{P}(X_0 = s_0, X_1 = s_1, X_2 = s_2) &= \mathbb{P}(X_0 = s_0)\mathbb{P}(X_1 = s_1 | X_0 = s_0)\mathbb{P}(X_2 = s_2 | X_1 = s_1) \\ &= q_0(s_0)p(s_0, s_1)p(s_1, s_2)\end{aligned}$$

As they fully characterise the probabilistic dynamic of the process, many other more complex events can then be computed only based on both the initial probability distribution q0 and the transition probability kernel p. One last basic relation that deserves to be given is the expression of the probability distribution at time n+1 expressed relatively to the probability distribution at time n

$$q_{n+1}(s_{n+1}) \overset{\text{def}}{=} \mathbb{P}(X_{n+1} = s_{n+1}) = \sum_{s \in E} \mathbb{P}(X_n = s)\mathbb{P}(X_{n+1} = s_{n+1} | X_n = s) = \sum_{s \in E} q_n(s)p(s, s_{n+1})$$

Finite state space Markov chains
Matrix and graph representation
We assume here that we have a finite number N of possible states in E:

$$E = \{e_1, e_2, ..., e_N\}$$

Then, the initial probability distribution can be described by a row vectorq0 of size N and the transition probabilities can be described by a matrix p of size N by N such that

$$(q_0)_i = q_0(e_i) = \mathbb{P}(X_0 = e_i)$$

$$p_{i,j} = p(e_i, e_j) = \mathbb{P}(X_{n+1} = e_j | X_n = e_i) \qquad \text{(independent of n)}$$

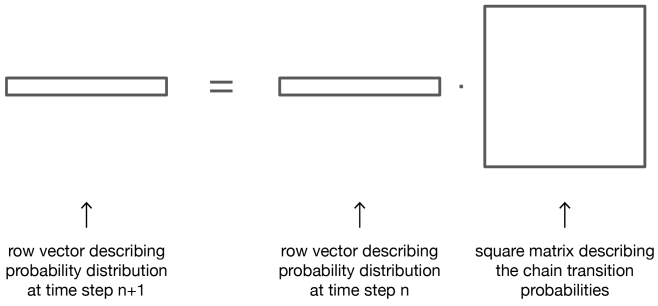
The advantage of such notation is that if we note denote the probability distribution at step n by a raw vector qn such that its components are given by

$$(q_n)_i = q_n(e_i) = \mathbb{P}(X_n = e_i)$$

then the simple matrix relations thereafter hold

$q_{n+1} = q_n p$ $\quad q_{n+2} = q_{n+1} p = (q_n p) p = q_n p^2$ $\quad ...$ $\quad q_{n+m} = q_n p^m$

(the proof won't be detailed here but can be recovered very easily).




When right multiplying a row vector representing probability distribution at a given time step by the transition probability matrix, we obtain the probability distribution at the next time step. So, we see here that evolving the probability distribution from a given step to the following one is as easy as right multiplying the row probability vector of the initial step by the matrix p. This also implies that we have

$$p_{i,j} = \mathbb{P}(X_{n+1} = e_j | X_n = e_i) \equiv \text{ probability of going from } e_i \text{ to } e_j \text{ in 1 step}$$

$$(p^2)_{i,j} = \mathbb{P}(X_{n+2} = e_j | X_n = e_i) \equiv \text{ probability of going from } e_i \text{ to } e_j \text{ in 2 steps}$$

...

$$(p^m)_{i,j} = \mathbb{P}(X_{n+m} = e_j | X_n = e_i) \equiv \text{ probability of going from } e_i \text{ to } e_j \text{ in m steps}$$

The random dynamic of a finite state space Markov chain can easily be represented as a valuated oriented graph such that each node in the graph is a state and, for all pairs of states (ei, ej), there exists an edge going from ei to ej if p(ei,ej)>0. The value of the edge is then this same probability p(ei,ej).

**Markov Chains properties**

In this section, we will only give some basic Markov chains properties or characterisations. The idea is not to go deeply into mathematical details but more to give an overview of what are the points of interest that need to be studied when using Markov chains. As we have seen that in the finite state space case we can picture a Markov chain as a graph, notice that we will use graphical representation to illustrate some of the properties bellow. However, one should keep in mind that these properties are not necessarily limited to the finite state space case.

Reducibility, periodicity, transience and recurrence

Let's start, in this subsection, with some classical ways to characterise a state or an entire Markov chain.

First, we say that a Markov chain is irreducible if it is possible to reach any state from any other state (not necessarily in a single time step). If the state space is finite and the chain can be represented by a graph, then we can say that the graph of an irreducible Markov chain is strongly connected (graph theory).

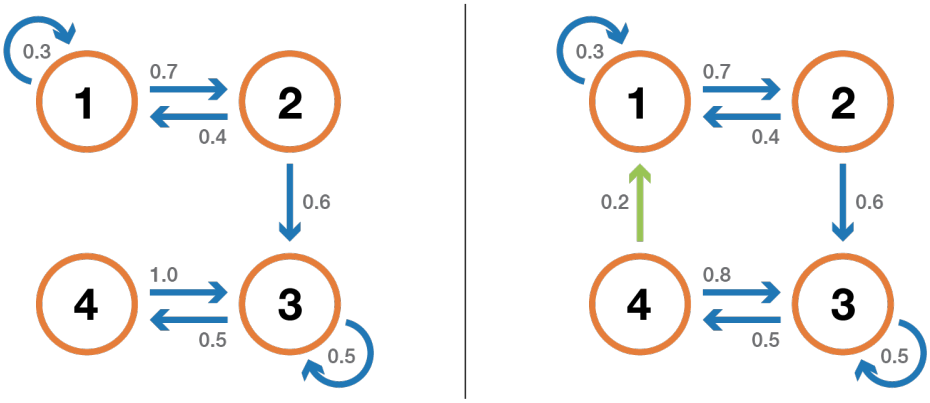



Illustration of the irreducibility property. The chain on the left is not irreducible: from 3 or 4 we can't reach 1 or 2. The chain on the right (one edge has been added) is irreducible: each state can be reached from any other state.

A state has period k if, when leaving it, any return to that state requires a multiple of k time steps (k is the greatest common divisor of all the possible return path length). If k = 1, then the state is said to be aperiodic and a whole Markov chain is aperiodic if all its states are aperiodic. For an irreducible Markov chain, we can also mention the fact that if one state is aperiodic then all states are aperiodic.

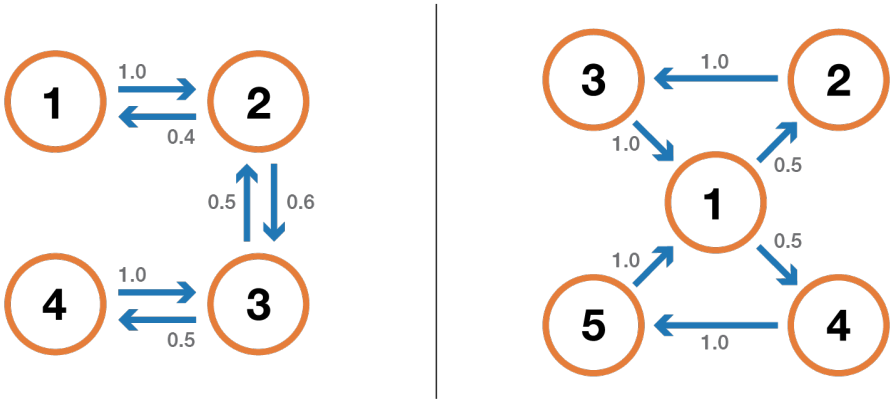



Illustration of the periodicity property. The chain on the left is 2-periodic: when leaving any state, it always takes a multiple of 2 steps to come back to it. The chain on the right is 3-periodic.

A state is transient if, when we leave this state, there is a non-zero probability that we will never return to it. Conversely, a state is recurrent if we know that we will return to that state, in the future, with probability 1 after leaving it (if it is not transient).

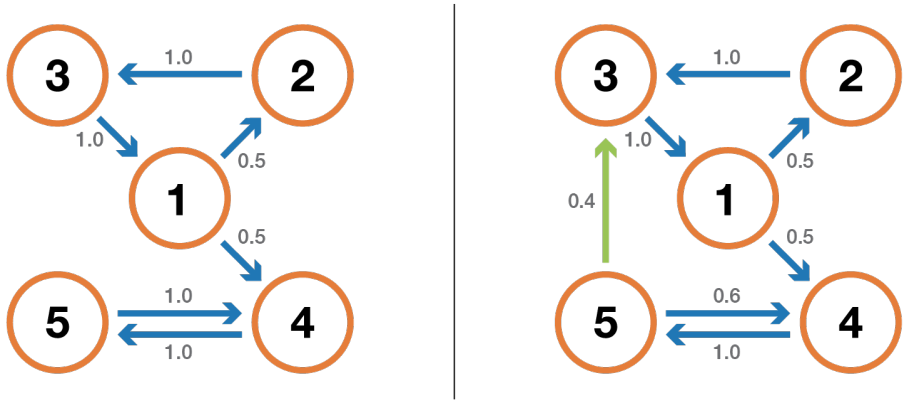



Illustration of the recurrence/transience property. The chain of the left is such that: 1, 2 and 3 are transient (when leaving these points we can't be absolutely sure that we will come back to them) and 3-periodic whereas 4 and 5 are recurrent (when leaving these points we are absolutely sure that we will come back to them at some time) and 2-periodic. The chain on the right has one more edge that makes the full chain recurrent and aperiodic.

For a recurrent state, we can compute the mean recurrence time that is the expected return time when leaving the state. Notice that even if the probability of return is equal to 1, it doesn't mean that the expected return time is finite. So, among the recurrent states, we can make a difference between positive recurrent state (finite expected return time) and null recurrent state (infinite expected return time).

Stationary distribution, limiting behaviour and ergodicity

We discuss, in this subsection, properties that characterise some aspects of the (random) dynamic described by a Markov chain.

A probability distribution $\pi$ over the state space E is said to be a stationary distribution if it verifies

$$\pi(e') = \sum_{e \in E} \pi(e) p(e, e') \qquad \forall e' \in E$$

As we have

$$\pi(e') = \text{probability of being in } e' \text{ at the current step}$$

$$\sum_{e \in E} \pi(e) p(e, e') = \text{probability of being in } e' \text{ at the next step}$$

Then a stationary distribution verifies

$$\text{probability of being in } e' \text{ at the current step} = \text{probability of being in } e' \text{ at the next step}$$

By definition, a stationary probability distribution is then such that it doesn't evolve through the time. So if the initial distribution q is a stationary distribution then it will stay the same for all future time steps. If the state space is finite, p can be represented by a matrix and π by a raw vector and we then have

$$\pi = \pi p = \pi p^2 = \dots$$

Once more, it expresses the fact that a stationary probability distribution doesn't evolve through the time (as we saw that right multiplying a probability distribution by p allows to compute the probability distribution at the next time step). Notice that an irreducible Markov chain has a stationary probability distribution if and only if all of its states are positive recurrent.

Another interesting property related to stationary probability distribution is the following. If the chain is recurrent positive (so that there exists a stationary distribution) and aperiodic then, no matter what the initial probabilities are, the probability distribution of the chain converges when time steps goes to infinity: the chain is said to have a limiting distributionthat is nothing else than the stationary distribution. In the general case it can be written

$$\lim_{n \to \infty} \mathbb{P}(X_n = e' | X_0 = e) = \lim_{n \to \infty} p^n(e, e') = \pi(e') \qquad \forall (e, e') \in E \times E$$

Let's emphasise once more the fact that there is no assumption on the initiale probability distribution: the probability distribution of the chain converges to the stationary distribution (equilibrium distribution of the chain) regardless of the initial setting.

Finally, ergodicity is another interesting property related to the behaviour of a Markov chain. If a Markov chain is irreducible then we also say that this chain is "ergodic" as it verifies the following ergodic theorem. Assume that we have an application f(.) that goes from the state space E to the real line (it can be, for example, the cost to be in each state). We can define the mean value that takes this application along a given trajectory (temporal mean). For the n-th first terms it is denoted by

$$\frac{1}{n}(f(X_0) + f(X_1) + \dots + f(X_{n-1})) = \frac{1}{n} \sum_{i=0}^{n-1} f(X_i)$$

We can also compute the mean value of application f over the set E weighted by the stationary distribution (spatial mean) that is denoted by

$$\sum_{e \in E} \pi(e) f(e)$$

Then ergodic theorem tells us that the temporal mean when trajectory become infinitely long is equal to the spatial mean (weighted by stationary distribution). The ergodic property can be written

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) = \sum_{e \in E} \pi(e) f(e)$$

Stated in another way, it says that, at the limit, the early behaviour of the trajectory becomes negligible and only the long run stationary behaviour really matter when computing the temporal mean.

---

**A classical example: the PageRank algorithm**

It's now time to come back to PageRank! Before going any further, let's mention the fact that the interpretation that we are going to give for the PageRank is not the only one possible and that authors of the original paper had not necessarily in mind Markov chains when designing the method. However, the following interpretation has the big advantage to be very well understandable.

The random web surfer

The problem PageRank tries to solve is the following: how can we rank pages of a given a set (we can assume that this set has already been filtered, for example on some query) by using the existing links between them?

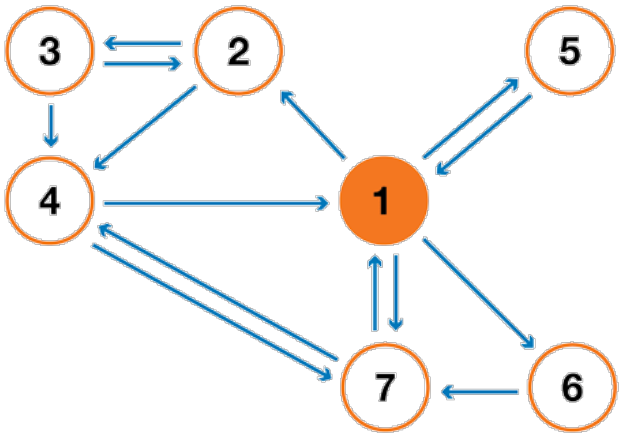
To solve this problem and be able to rank the pages, PageRank proceed roughly as follows. We consider that a random web surfer is on one of the pages at initial time. Then, this surfer starts to navigate randomly by clicking, for each page, on one of the links that lead to another page of the considered set (assume that links to pages out of this set are disallowed). For a given page, all the allowed links have then equal chance to be clicked.

We have here a the setting of a Markov chain: pages are the different possible states, transition probabilities are defined by the links from page to page (weighted such that on each page all the linked pages have equal chances to be chosen) and the memoryless properties is clearly verified by the behaviour of the surfer. If we assume also that the defined chain is recurrent positive and aperiodic (some minor tricks are used to ensure we meet this setting), then after a long time the "current page" probability distribution converges to the stationary distribution. So, no matter the starting page, after a long time each page has a probability (almost fixed) to be the current page if we pick a random time step.

The hypothesis behind PageRank is that the most probable pages in the stationary distribution must also be the most important (we visit these pages often because they receive links from pages that are also visited a lot in the process). The stationary probability distribution defines then for each state the value of the PageRank.

**A toy example**

In order to make all this much clearer, let's consider a toy example. Assume that we have a tiny website with 7 pages labelled from 1 to 7 and with links between the pages as represented in the following graph.




Trajectory : 1

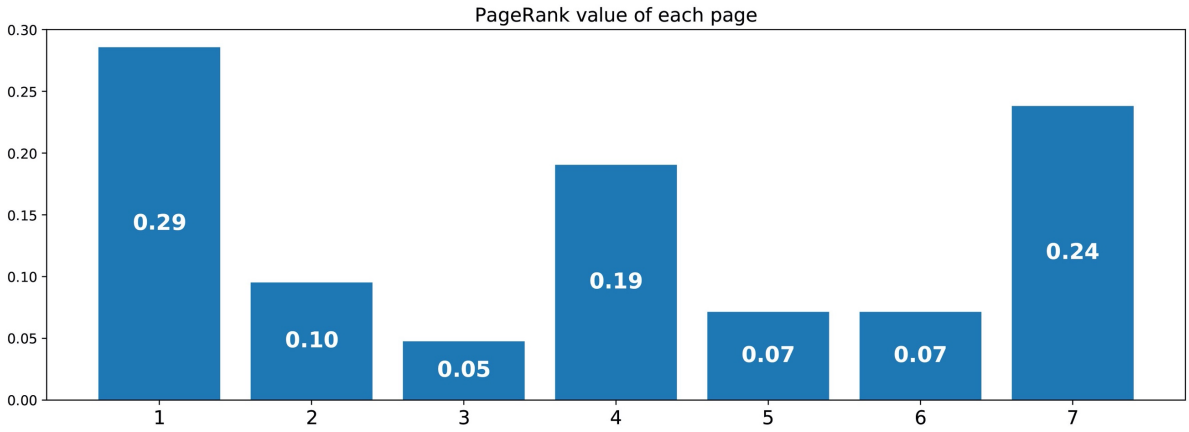
For clarity the probabilities of each transition have not been displayed in the previous representation. However, as the "navigation" is supposed to be purely random (we also talk about "random walk"), the values can be easily recovered using the simple following rule: for a node with K outlinks (a page with K links to other pages), the probability of each outlink is equal to 1/K. So, the probability transition matrix is given by

$$p = \begin{pmatrix} . & 0.25 & . & . & 0.25 & 0.25 & 0.25 \\ . & . & 0.5 & 0.5 & . & . & . \\ . & 0.5 & . & 0.5 & . & . & . \\ 0.5 & . & . & . & . & . & 0.5 \\ 1.0 & . & . & . & . & . & . \\ . & . & . & . & . & . & 1.0 \\ 0.5 & . & . & 0.5 & . & . & . \end{pmatrix}$$

where 0.0 values have been replaced by '.' for readability. Before any further computation, we can notice that this Markov chain is irreducible as well as aperiodic and, so, after a long run the system converges to a stationary distribution. As we already saw, we can compute this stationary distribution by solving the following left eigenvector problem

$$\pi = \pi p \qquad \text{with} \qquad \pi = \begin{pmatrix} \pi_1 & \pi_2 & \pi_3 & \pi_4 & \pi_5 & \pi_6 & \pi_7 \end{pmatrix}$$

Doing so we obtain the following values of PageRank (values of the stationary distribution) for each page



PageRank values computed on our toy example that contains 7 pages.
The PageRank ranking of this tiny website is then $1 > 7 > 4 > 2 > 5 = 6 > 3$.

---

**The main takeaways of this unit are the following:**
random processes are collections of random variables, often indexed over time (indices often represent discrete or continuous time)
for a random process, the Markov property says that, given the present, the probability of the future is independent of the past (this property is also called "memoryless property")
discrete time Markov chain are random processes with discrete time indices and that verify the Markov property
the Markov property of Markov chains makes the study of these processes much more tractable and allows to derive some interesting explicit results (mean recurrence time, stationary distribution…)
one possible interpretation of the PageRank (not the only one) consists in imagining a web surfer that randomly navigates from page to page and in taking the induced stationary distribution over pages as a factor of ranking (roughly, the most visited pages in steady-state must be the one linked by other very visited pages and then must be the most relevant)
To conclude, let's emphasise once more how powerful Markov chains are for problems modelling when dealing with random dynamics. Due to their good properties, they are used in various fields such as queueing theory(optimising the performance of telecommunications networks, where messages must often compete for limited resources and are queued when all ressources are already allocated), statistics (the well known "Markov Chain Monte Carlo" random variables generation technique is based on Markov chains), biology (modelling of biological populations evolution), computer science (hidden Markov models are important tools in information theory and speech recognition) and others.
Obviously, the huge possibilities offered by Markov chains in terms of modelling as well as in terms of computation go far behind what have been presented in this modest introduction and, so, we encourage the interested reader to read more about these tools that entirely have there place in the (data) scientist toolbox.

## What is the Markov Property?

Discrete Time Markov Property states that the calculated probability of a random process transitioning to the next possible state is only dependent on the current state and time and it is independent of the series of states that preceded it.
The fact that the next possible action/ state of a random process does not depend on the sequence of prior states, renders Markov chains as a memory-less process that solely depends on the current state/action of a variable.
Let's derive this mathematically:

Let the random process be, {Xm, m=0,1,2,⋯}.
This process is a Markov chain only if,

$$P(X_{m+1} = j|X_m = i, X_{m-1} = i_{m-1}, \cdots, X_0 = i_0) = P(X_{m+1} = j|X_m = i)$$

for all m, j, i, i0, i1, ⋯ im−1
For a finite number of states, S={0, 1, 2, ⋯, r}, this is called a finite Markov chain.
P(Xm+1 = j|Xm = i) here represents the transition probabilities to transition from one state to the other. Here, we're assuming that the transition probabilities are independent of time.
Which means that P(Xm+1 = j|Xm = i) does not depend on the value of 'm'. Therefore, we can summarise,

$$P_{ij} = P(X_{m+1} = j|X_m = i)$$

So this equation represents the Markov chain.
Now let's understand what exactly Markov chains are with an example.

## What Is A Transition Probability Matrix?

In the above section we discussed the working of a Markov Model with a simple example, now let's understand the mathematical terminologies in a Markov Process.
In a Markov Process, we use a matrix to represent the transition probabilities from one state to another. This matrix is called the Transition or probability matrix. It is usually denoted by P.

$$P = \begin{bmatrix} p_{11} & p_{12} & . & . & . & p_{1r} \\ p_{21} & p_{22} & . & . & . & p_{2r} \\ . & . & & . & & . \\ . & . & & . & & . \\ p_{21} & p_{22} & . & . & . & p_{2r} \end{bmatrix}$$

Note, pij≥0, and 'i' for all values is,

$$\sum_{k=1}^{r} p_{ik} = \sum_{k=1}^{r} P(X_{m+1} = k|X_m = \text{i})$$

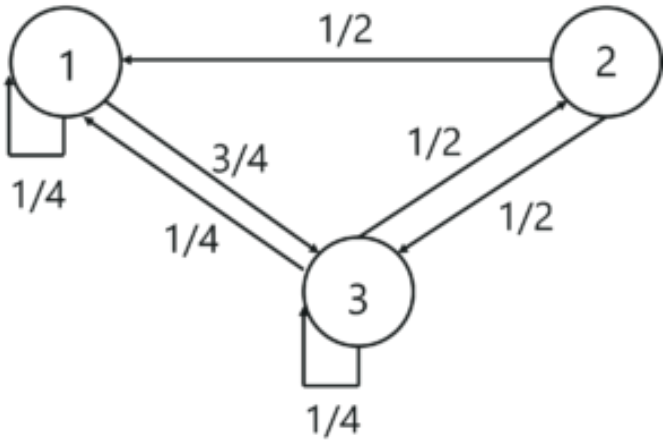
Let me explain this. Assuming that our current state is 'i', the next or upcoming state has to be one of the potential states. Therefore, while taking the summation of all values of k, we must get one.

## What Is A State Transition Diagram?

A Markov model is represented by a State Transition Diagram. The diagram shows the transitions among the different states in a Markov Chain. Let's understand the transition matrix and the state transition matrix with an example.

**Transition Matrix Example**
Consider a Markov chain with three states 1, 2, and 3 and the following probabilities:

$$P = \begin{bmatrix} 1/4 & 0 & 3/4 \\ 1/2 & 0 & 1/2 \\ 1/4 & 1/2 & 1/4 \end{bmatrix}$$




The above diagram represents the state transition diagram for the Markov chain. Here, 1,2 and 3 are the three possible states, and the arrows pointing from one state to the other states represents the transition probabilities pij. When, pij=0, it means that there is no transition between state ‘i’ and state ‘j’.
Now that we know the math and the logic behind Markov chains, let’s run a simple demo and understand where Markov chains can be used.