# Web Information Retrieval

## Prof. Luca Becchetti: (very) rough notes on Markov Chains and Random Walks

While reviewing a first draft of this notes, I came across the following tutorial on Markov chains @Towards Data Science, which I found extremely accurate and well done. I encourage you to take a look, it contains further points and nice examples that are worth reflecting upon.

### Basics

Consider the process described by the following picture:

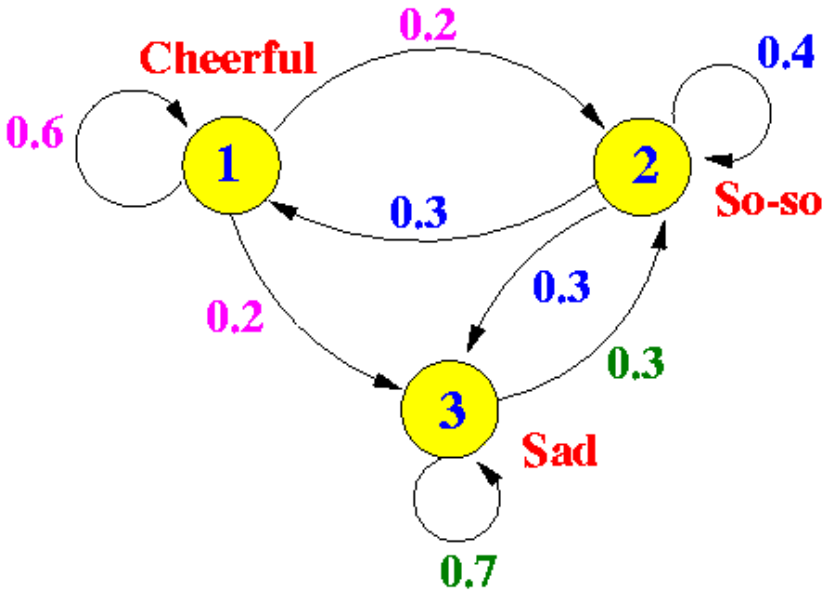



*Image source:* http://www.mathcs.emory.edu/~cheung/

Above we have a (directed) network $G = (V, E)$. Nodes represent *states*, directed links represent possible *transitions* between them. Each link $(i, j)$ (note that order is important!) is labelled with $P_{ij}$, the probability that, if in state $i$ at any round $t$, we "move" to state $j$ in round $t + 1$. For example, $P_{23} = 0.3$ in the picture above.

**Process.** The process starts in some initial state at round $0$ and it performs a transition to a different (possibly the same) state in each round, following an outgoing link of the current state with the probability associated to that link. Such a process is called a *Markov Chain* (often abbreviated as MC in the remainder of these notes).

**Remark.** To be precise, the one given above is an example of a *discrete* (transitions occur in discrete time steps), *homogeneous* (transition probabilities do not change over time), *finite-state* (we have a finite number of states, $3$ in the example above) MC. As you may suspect, a number of variants are possible (and studied). In the remainder, we restrict to discrete, homogeneous, finite-state MCs, even though I just write MC for brevity.

**Notation.** We use $X(t)$ to denote the state of the Markov chain after $t$ steps. We use $P$ to denote the $n \times n$ matrix, whose entry $(i, j)$ is $P_{ij}$. $P$ is called the *transition matrix* of the MC.

**Markov property.** As the picture above also suggests, a Markov Chain (hence, MC) exhibits the following, crucial property:

$$\mathbf{P}(X(t) = j | X(t-1) = a_{t-1}, \ldots, X(0) = a_0) = \mathbf{P}(X(t) = j | X(t-1) = a_{t-1}),$$

where $a_r$ denotes the state in which the MC found itself at the end of round $r$, $r = 1, \ldots, t-1$. Stated informally, what is going to happen next only depends on what is happening now. Note that

$$\mathbf{P}(X(t) = j | X(t-1) = i) = P_{ij}$$

**Properties of $\mathbf{P}$.** We assume henceforth that each state $i$ has at least one outgoing link and that the matrix is stochastic, i.e., $\sum_{j:i \rightsquigarrow j} P_{ij} = 1$, where $i \rightsquigarrow j$ means that a link from $i$ to $j$ exists. Note that, for Pagerank, we enforce this property by adding $n$ links from each *dead end* $j$, each pointing to a different state (including $j$ itself) and with associated transition probability $1/n$).

## MC evolution

After $t$ steps, the MC is in each state with some probability (possibly, 0). This corresponds to a probability distribution $\mathbf{p}(t)$. Here, $\mathbf{p}(t)$ is an $n$-dimensional vector, such that its $j$-th entry is $\mathbf{p}_j(t) = \mathbf{P}(X(t) = j)$, i.e., the unconditional probability that the MC ends up in state $j$ in round $t$. Since the MC has to be in some state at the end of each round (recall that we have no dead ends), we necessarily have $\sum_{j=1}^{n} \mathbf{p}_j(t) = 1$, for every $t$. There is an obvious relationship between $\mathbf{p}(t)$ and $\mathbf{p}(t-1)$. Namely, $\mathbf{p}_j(t)$ depends on $\mathbf{p}(t)$ as follows:

$$\mathbf{p}_j(t) = \sum_{i \rightsquigarrow j} \mathbf{p}_i(t-1) P_{ij}.$$

You should convince yourself that this relationship can be expressed in compact form as:

$$\mathbf{p}(t)^T = \mathbf{p}(t-1)^T P \qquad (1)$$

Iterating over different rounds we finally obtain:

$$\mathbf{p}(t)^T = \mathbf{p}(0)^T P^t.$$

**Rows or columns?** If you transpose both sides of Eq. (1) you immediately obtain:

$$\mathbf{p}(t) = (P^T)^t \mathbf{p}(0).$$

Now, $\mathbf{p}(t)$ and $\mathbf{p}(0)$ are column vectors and $P^T$ has columns (not rows) summing to 1, but the two forms are completely equivalent for all that matters. People often consider the first form when studying Markov chains, while a column representation is mostly preferred in the related literature on Spectral Graph Theory, a topic well beyond the scope of this lecture.

## Ergodic Markov chains

A number of questions are of interest, the most important being:

**1.** What happens to $\mathbf{p}(t)$ as $t \rightarrow \infty$?

**2.** Does $\lim_{t \rightarrow \infty} \mathbf{p}(t)$ exist? Is it unique?

The answer to the questions above depends on the structure of the MC, i.e., the topology of the underlying, directed graph. In the remainder, we focus on the class of ergodic MCs, for which the answer to the above questions is always yes. We emphasize that the MC defining Pagerank is ergodic.

**Definition.** We say the state $j$ is *accessible* from state $i$ if a directed path from $i$ to $j$ exists in the MC.

Note that this implies that, for all $i, j \in V$: $P_{ij}^k$ for some $k \geq 0$.

**Definition.** We say the $i$ and $j$ *communicate* if and only if $j$ is accessible from $i$ and viceversa.

Note that communicating relation is, mathematically speaking, an *equivalence relation*, namely, it satisfies the *reflexive*, *symmetric* and *transitive* properties. You are warmly invited to review the notion of equivalence relation, if needed.

**Definition.** A MC is *irreducible* if and only if it has a single communicating class. Otherwise, the MC is said *reducible*.

**Remark.** In practice, this means that the (directed graph) representing the MC is *strongly connected*, i.e., for every $i$ and $j$ there are directed paths connecting $i$ to $j$ and viceversa.

When a MC is reducible, $\lim_{t\to\infty} \mathbf{p}(t)$, provided it exists, in general depends on $\mathbf{p}(0)$.

**Definition.** A state $j$ of a MC is *periodic* if an integer $\Delta > 1$ exists, such that $\mathbf{P}(X(t+s) = j | X(t) = j) = 0$ , unless $s$ is divisible by $\Delta$. A MC is periodic if it contains at least one periodic state, otherwise it is said *aperiodic*.

**Fact.** In an irreducible Markov Chain, the presence of a single aperiodic state implies that all states are aperiodic, i.e., the chain is aperiodic.

Irreducible and aperiodic MCs identify the class of *ergodic* MCs. Ergodic MCs exhibit nice properties. First of all, if $P$ is the transition matrix of an ergodic MC, the following holds:

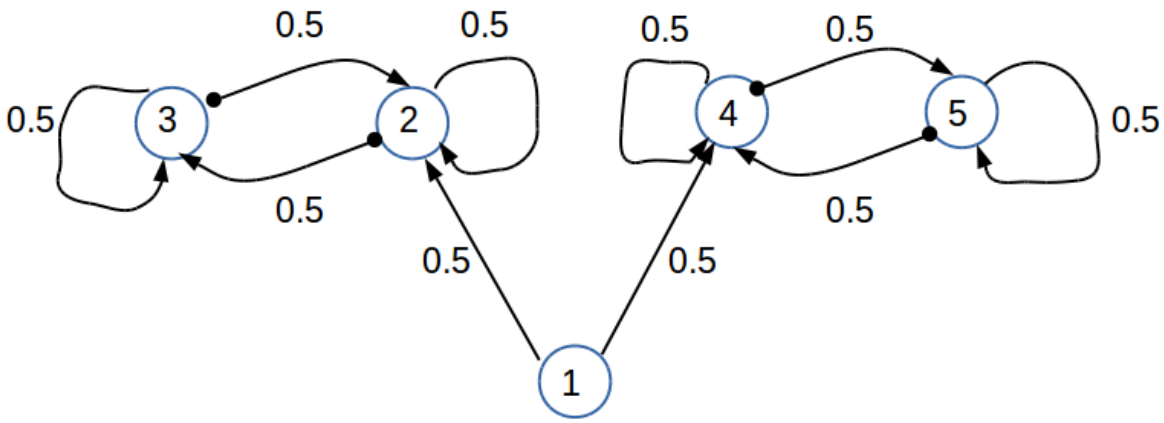
**Property 1.** If $P$ is the transition matrix of an ergodic MC, an integer $k \geq 1$ exists, such that $P_{ij}^t > 0$, for every pair $i, j$ of states, whenever $t \geq k$.

The behaviour of $\lim_{t\to\infty} \mathbf{p}(t)$ is interesting in ergodic MCs. Before we discuss this, we need the notion of stationary distribution.

**Definition.** A *stationary* (or *equilibrium*) distribution of a MC with transition matrix $P$ is a probability distribution $\pi$, such that $\pi^T = \pi^T P$.

Note that, by definition, if a MC reaches a stationary distribution, it maintains that distribution forever.

**Remark.** In general, one can have multiple, stationary distributions. Consider the following example:




The transition matrix of this Markov chain is:

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

It is easy to see that this MC has multiple, possible stationary distributions. One is $(\pi^{(1)})^T = (0, 1/2, 1/2, 0, 0)$, another one is $(\pi^{(2)})^T = (0, 0, 0, 1/2, 1/2)$. To see this, just observe that $(\pi^{(1)})^T P = \pi^{(1)}$ and $(\pi^{(2)})^T P = \pi^{(2)}$. Moreover, which stationary distribution is reached depends on the initial distribution $\mathbf{p}(0)$. For example, if $\mathbf{p}(0) = (0, 1, 0, 0, 0)$ the MC will converge to $\pi^{(1)}$ with probability 1 while, if $\mathbf{p}(0) = (1, 0, 0, 0, 0)$, the stationary distribution will be one between $\pi^{(1)}$ and $\pi^{(2)}$ with equal chances, i.e., the stationary distribution will be $\frac{1}{2}\pi^{(1)} + \frac{1}{2}\pi^{(2)}$ in this case (you can verify yourselves that this is yet another stationary distribution).

Ergodic MCs have interesting properties with respect to stationary distributions. In particular, we have the following, fundamental theorem:

**Theorem.** Any finite, ergodic MC has the following properties:

- The chain has a *unique* stationary distribution $\pi^T = (\pi_1, \ldots, \pi_n)$.
- For every pair $i, j$, the limit $\lim_{t\to\infty} P^t_{ij}$ exists and is *independent* of $i$.
- For every $j$: $\lim_{t\to\infty} P^t_{ij} = \pi_j$.

**Remarks.** The theorem above has a number of important implications. The first consequence (actually, just a restatement of the third claim of the theorem) is that

$$\lim_{t\to\infty} \mathbf{p}(0)P^t = \pi^T,$$

*independently* of $\mathbf{p}(0)$. This implies that, picking any arbitrary initial distribution (for example, $\mathbf{p}(0) = \mathbf{e}^i$, with $\mathbf{e}^i$ the $i$-th canonical vector), and applying the *power method*, i.e.,

$$\mathbf{p}(t)^T = \mathbf{p}(t-1)^T P, \ \ for\ t > 0,$$

we have that $\mathbf{p}(t)$ eventually converges to the stationary distribution, whenever $P$ is the transition matrix of an ergodic MC.

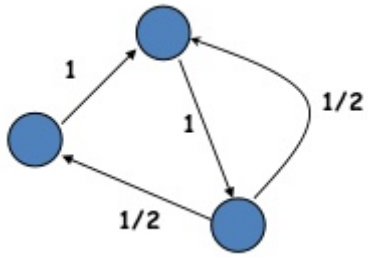
A second important consequence (an intuitive one, but one that would need a proof) is that

$$\lim_{t\to\infty} \frac{\sum_{k=0}^{t} \mathbf{p}_i(k)}{t} = \pi_i, for\ every\ i$$

If one considers that the numerator is the expected number of times that $i$ is visited over $t$ consecutive rounds, the above equality simply states that the expected fraction of time spent by a MC in a state over a (very) long time span is very close to the stationary probability for that state.

## Random walks and Markov chains

Given a (directed or undirected) graph $G = (V, E)$, a walk on $G$ is a (not necessarily simple) path on $G$, i.e., a sequence of vertices (possibly with repetitions), such that $v_2$ may occur right after $v_1$ in the sequence if and only if the (possibly directed) edge $(v_1, v_2)$ exists. In a random walk, the next edge that is traversed is chosen randomly: if at vertex $u$, in the next step the walker crosses one of the outgoing edges of $u$ with the same probability $1/d_u$, where $d_u$ denotes the out-degree of $u$ (which of course is just the degree if $G$ is undirected). This is the simplest version of random walks, more general notions are possible (e.g., in weighted graphs). Consider a random walk on a graph $G = (V, E)$, such as the following:




The random walk may be described by the following transition matrix $P$ (or one obtained permuting the rows of $P$, depending on how we number vertices of the graph):

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

A random walk is always a MC. If the graph is strongly connected, the resulting transition matrix is always stochastic (no dead ends) and thus it corresponds to a MC . In this case, the entries on row $i$ correspond to outgoing links from $i$ and they can take on two values ($d_i$ is the out-degree of $i$):

$$P_{ij} = \begin{cases} \frac{1}{d_i}, if\ (i,j) \in E, \\ 0, if\ (i,j) \notin E \end{cases}$$

In fact, we can also regard a MC as a random walk, with the caveat that, in this case, jumps from one node of the graph to its neighbours do not occur with the same probability.

**Important.** Note that an undirected, connected graph is also strongly connected, hence a random walk is always defined on any undirected, connected graph.

**Question:** what happens if an undirected graph consists of two or more connected components? Is a random walk defined? If yes, is it ergodic?

## Pagerank's random walk

Assume we have a Web graph $G = (V, E)$ with $n$ vertices. In theory, we would like to score vertices in $V$ using scores that reflects their centrality in $G$. The basic intuition is using a random walk on $G$ to achieve this goal. Unfortunately, in most cases, we cannot directly use $G$ to perform a random walk. The reason is that $G$ may not be strongly connected, since it can have dead ends, periodic nodes and so on. In general, a random walk in $G$ does not correspond to an ergodic MC (it may not even be defined, as is the case if we have dead ends). To obtain an ergodic MC that strongly depends on $G$'s underlying topology, we proceed as follows.

**Step 1: removing dead ends.** Assume $\hat{P}$ is $G$'s transition matrix. If $G$ contains a dead end $i$, the corresponding row in $\hat{P}$ will be a row of $0$'s. Such rows are removed simply by replacing each of them with the vector $\frac{1}{n}\mathbf{1}^T$, which corresponds to adding $n$ links from $i$ to each distinct vertex in $G$ (including $i$ itself).

Henceforth, we consider the matrix $P = \hat{P} + \frac{1}{n}\mathbf{a}\mathbf{1}^T$, where $\mathbf{a}$ is a column vector, such that $\mathbf{a}_i = 1$ if $i$ is a dead end, $\mathbf{a}_i = 0$ otherwise. Note that $P$ is now a stochastic matrix and that the corresponding MC might still be reducible and/or periodic.

**Step 2: teleportation.**

The random walk described by $P$ is modified as follows. When at a generic vertex $i$ of $G$ (i.e., its possibly modified version in which dead ends were removed):

- With probability $\alpha$:
  - Follow one of $i$'s outgoing links uniformly at random.
- With probability $1-\alpha$:
  - Jump to any vertex (including $i$) uniformly at random (hence, with probability $1/n$).

This corresponds to a MC/random walk described by the following equation for the generic vertex $i$:

$$\mathbf{p}_i(t) = \alpha \sum_{j:(j,i)\in E} \mathbf{p}_j(t-1)P_{ji} + \frac{1-\alpha}{n} = \alpha \sum_{j:(j,i)\in E} \frac{\mathbf{p}_j(t-1)}{d_j} + \frac{1-\alpha}{n}$$

Written in matrix form for all vertices, this equation becomes:

$$\mathbf{p}(t)^T = \alpha \mathbf{p}(t-1)^T P + \frac{1-\alpha}{n}\mathbf{1}^T = \mathbf{p}(t-1)^T \left(\alpha P + \frac{1-\alpha}{n}\mathbf{1}\mathbf{1}^T\right),$$

where the second equality follows since $\mathbf{p}(t-1)^T\mathbf{1} = 1$ (recall that $\mathbf{p}(t-1)^T$ is a probability distribution). You should convince yourself that steps 1 and 2 enforce ergodicity of the corresponding MC. To this end, note that i) each vertex can be reached from each other vertex (possibly, with tiny probability) and ii) teleportation removes any form of periodicity (in principle, we can reach any other vertex in one step, with probability at least $1/n$).

## References

This is a very succint excerpt of material that can be found in many textbooks about linear algebra and/or probability and algorithms. For example, you find a thorough introduction in Chapter 7 of the following book:

Michael Mitzenmacher and Eli Upfal. Probability and Computing, 2nd edition. Cambridge University Press, 2017.