Networked Life: Useful mathematics

Costas Courcoubertis, Barnabé Monnot

April 28, 2016

1 Linear algebra

1.1 Some basic linear algebra

1.1.1 Vectors

We review some basic notions of linear algebra. Unless otherwise stated, all our variables will belong to $\mathbb{R}$, set of all **real numbers**. We let $\mathbb{R}_+ = [0, +\infty)$.

A **vector** x belongs to $\mathbb{R}^n$ if x can be written as

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \ x_1 \in \mathbb{R}, \cdots, x_n \in \mathbb{R}$$

We let all of our vectors be column vectors by default.

Addition between vectors and multiplication by a scalar are well-defined operations

$$x + y = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{pmatrix}, \quad \lambda x = \begin{pmatrix} \lambda x_1 \\ \lambda x_2 \\ \vdots \\ \lambda x_n \end{pmatrix}, \quad x, y \in \mathbb{R}^n, \lambda \in \mathbb{R}$$

If $x \in \mathbb{R}^n$ is a vector, then we let x^T denote the **transpose** of x

$$x^T = (x_1, \cdots, x_n)$$

Let x and y be two vectors in $\mathbb{R}^\times$, then their usual **scalar product** is written as

$$x \cdot y = x^T y = y^T x = \sum_{i=1}^{n} x_i y_i$$

1.1.2 Matrices

A is a **matrix** in $\mathbb{R}^{n \times p}$ if it can be written as

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ A_{n1} & \cdots & \cdots & A_{np} \end{pmatrix}, \ A_{ij} \in \mathbb{R}, \ \forall i \in \{1, \cdots, n\}, j \in \{1, \cdots, p\}$$

Definition 1.1 (Square matrix)**.** A matrix is called **square** if its number of rows is equal to its number of columns.

The **identity matrix** I is a square matrix such that all diagonals elements are equal to 1, and the others are 0.

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

The **transpose** of $A \in \mathbb{R}^{n \times p}$ is noted $A^T \in \mathbb{R}^{p \times n}$ and is such that $(A^T)_{ij} = A_{ji}$.

$$A^T = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n1} \\ \vdots & \ddots & \ddots & \vdots \\ A_{1p} & \cdots & \cdots & A_{pn} \end{pmatrix}$$

Definition 1.2 (Symmetric matrix)**.** A is **symmetric** if $A = A^T$ (it implies that A is square).

Matrix multiplication This is the matrix multiplication you are used to. If $A \in \mathbb{R}^{n \times p}$ and $B \in \mathbb{R}^{p \times m}$, then $AB \in \mathbb{R}^{n \times m}$ is defined as

$$(AB)_{ij} = \sum_{k=1}^{p} A_{ik} B_{kj}$$

In other terms, $(AB)_{ij}$ is the scalar product of A's i-th row with B's j-th column. Note that in general, $AB \neq BA$, and you can only multiply two matrices if the left operand's number of columns is equal to the right operand's number of rows.

Matrix multiplication is also well-defined for matrix/vector operations:

$$Ax = \begin{pmatrix} \sum_{k=1}^{p} A_{1k} x_k \\ \sum_{k=1}^{p} A_{2k} x_k \\ \vdots \\ \sum_{k=1}^{p} A_{nk} x_k \end{pmatrix}, \; A \in \mathbb{R}^{n \times p}, \; x \in \mathbb{R}^p$$

Definition 1.3 (Invertible matrix)**.** A square matrix $A \in \mathbb{R}^{n \times n}$ is **invertible** if there exists a matrix $A^{-1} \in \mathbb{R}^{n \times n}$ such that

$$AA^{-1} = A^{-1}A = I$$

Equivalently A is invertible if $\det(A) = 0$, where $\det(A)$ is the determinant of A.

1.1.3 Basis and dimensions

Let $\{x_i\}_{i=1}^k$ be a family of k vectors in $\mathbb{R}^n$. We say that the family is **linearly independent** if

$$\lambda_1 x_1 + \cdots + \lambda_k x_k = 0 \Rightarrow \lambda_1 = \cdots = \lambda_k = 0, \; \lambda_i \in \mathbb{R} \; \forall i$$

We call the **span** of $\{x_i\}_{i=1}^k$, or the space generated by $\{x_i\}_{i=1}^k$, the set

$$\text{span}(\{x_i\}_{i=1}^k) = \{y \in \mathbb{R}^n; \; \exists \lambda_1, \cdots, \lambda_k \text{ such that } y = \sum_{i=1}^{k} \lambda_i x_i\}$$

$\{x_i\}_{i=1}^n$ is a **basis** of $\mathbb{R}^n$ if it is a linearly independent family of vectors and $\text{span}(\{x_i\}_{i=1}^n) = \mathbb{R}^n$. Note that we need as many vectors in our family as the exponent in $\mathbb{R}^n$.

All basis of a space have the same number of elements. That number is called the **dimension** of a space (so the dimension of $\mathbb{R}^n$ is n).

Definition 1.4 (Kernel, image, rank)**.** The **kernel** of a matrix $A \in \mathbb{R}^{n\times p}$, noted $\ker(A)$, is the set of vectors $x \in \mathbb{R}^p$ such that $Ax = 0$ (i.e each component of Ax is equal to 0).

$$\ker(A) = \{x \in \mathbb{R}^p;\ Ax = 0\}$$

The **image** of $A \in \mathbb{R}^{n\times p}$, noted $\text{im}(A)$ is the set of vectors $y \in \mathbb{R}^n$ such that you can find $x \in \mathbb{R}^p$ and $Ax = y$.

$$\text{im}(A) = \{y \in \mathbb{R}^n;\ \exists x \in \mathbb{R}^p \text{ such that } Ax = y\}$$

The **rank** of a matrix is the dimension of its image. Alternatively, it is the largest number of linearly independent rows of A.

Theorem 1.1 (Rank theorem)**.** *For $A \in \mathbb{R}^{n\times p}$*

$$\dim(\ker(A)) + \dim(im(A)) = p$$

1.2 Spectral theory

1.2.1 Eigenvalues, eigenvectors

We say that $\lambda \in \mathbb{C}$ is an **eigenvalue** of $A \in \mathbb{R}^{n\times n}$ (A square matrix) if there exists a vector $v \in \mathbb{R}^n$ such that $Av = \lambda v$. v is then called the **eigenvector** associated to λ.

The easiest way to find eigenvalues is by finding the roots of A's **characteristic polynomial**, noted $p(\lambda)$

$$p(\lambda) = \det(\lambda I - A)$$

Solving $p(\lambda) = 0$ gives us the roots $\{\lambda_1, \cdots, \lambda_n\}$ of p.

Definition 1.5 (Algebraic multiplicity)**.** We call **algebraic multiplicity** (AM) of λ_i the multiplicity of λ_i in p, i.e if $p(\lambda) = (\lambda - \lambda_i)^k q(\lambda)$, where q is another polynomial such that $q(\lambda_i) \neq 0$, then the algebraic multiplicity of λ_i is k.

Definition 1.6 (Geometric multiplicity)**.** We call **geometric multiplicity** (GM) of λ_i the dimension of the kernel of $\lambda_i I - A$.

When for all λ_i roots of p the algebraic multiplicity equals the geometric multiplicity, then the matrix A is **diagonalizable**, i.e A can be written as

$$A = TDT^{-1}$$

where $D = \text{diag}(\lambda_1, \cdots, \lambda_n)$ and T is formed with the corresponding eigenvectors, $T = [v_1|\cdots|v_n]$, v_i being a basis element of $\ker(\lambda_i I - A)$.

The case of symmetric matrices We will often be concerned with symmetric matrices, when looking at symmetric externalities between agents for example. These matrices have a lot of nice properties. We start with some definitions:

Definition 1.7 (Positive definite matrices)**.** A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is **positive definite** if for all $x \in \mathbb{R}^n \backslash \{0\}$, $x^T Ax > 0$. Equivalently, A is positive definite when all eigenvalues of A are positive.

Definition 1.8 (Positive semi-definite matrices)**.** A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is **positive semi-definite** if for all $x \in \mathbb{R}^n$, $x^T Ax \geq 0$. Equivalently, A is positive semi-definite when all eigenvalues of A are nonnegative.

Proposition 1.1 (Properties of symmetric matrices)**.** *Let $A, B \in \mathbb{R}^{n\times n}$ symmetric matrices. Then*

- $(AB)^T = B^T A^T$
- *All eigenvalues of A are real*
- *A is diagonalizable*

1.2.2 Jordan normal form

What if the matrix is not diagonalizable, i.e if we don't have AM = GM for all eigenvalues? We can arrive at a similar matrix decomposition, but we need to change both T and D.

Let's show it in an example. Suppose A is a matrix with eigenvalues $-1, 1$ and 2, AM(1) = AM(2) = 2, AM(-1) = 3, GM(1) = GM(-1) = 2, GM(2) = 1. Then we can find a matrix T such that $A = TJT^{-1}$ where

$$J = \begin{pmatrix} 1 & & & & & & \\ & 1 & & & & & \\ & & -1 & 1 & & & \\ & & & -1 & & & \\ & & & & -1 & & \\ & & & & & 2 & 1 \\ & & & & & & 2 \end{pmatrix}$$

(all the non-assigned cells are 0). This is called the **Jordan normal form**.

The idea is to have appear on the diagonal the eigenvalues as many times as their algebraic multiplicity. Then, when the geometric multiplicity is strictly less than the AM, we have ones appear on the upper diagonal. How many of them? AM minus GM exactly.

We form the transformation matrix T with a notion of **generalized eigenvectors** on which we will not expand here.

1.2.3 Spectral radius

We usually care (a lot) about the largest eigenvalue (in absolute value): this is called the **spectral radius** of a matrix, noted $\rho(A)$.

We will see later for example that this eigenvalue, and the gap between the largest and second largest eigenvalue, when obtained from a particular matrix representation of a graph gives us a measure of connectedness of this graph.

It is also an important measure to study the stability of **matrix powers**. Define A^k recursively, as $A^k = A^{k-1}A$, for $k \geq 1$, with $A^0 = I$. Then we have the following result:

Proposition 1.2.

$$\rho(A) < 1 \Leftrightarrow \lim_{k\to+\infty} A^k = 0$$

And the corollary:

Corollary 1.3.

$$\rho(A) < 1 \Rightarrow \sum_{k=0}^{\infty} A^k = (I - A)^{-1}$$

Proof. It is a simple proof that looks like the one of the sum of a geometric series. Try to do it! □

The proposition can be proved using the powers of a matrix put in a Jordan normal form. Note that for a matrix A decomposed as $A = TDT^{-1}$, we always have $A^k = TD^kT^{-1}$, which is super convenient when the matrix in the middle is easy to handle (such as diagonal or in Jordan form).

The Perron-Frobenius theorem gives us more information on the largest eigenvalue if the matrix has all entries real positive numbers.

Theorem 1.2 (Perron-Frobenius)**.** *If all components of a real square matrix A are positive, then we have the following properties:*

- *The largest eigenvalue r is unique (i.e $AM(r) = 1$) and positive.*
- *There exists a corresponding eigenvector with all components positive.*
- *r satisfies*
$$\min_i \sum_j A_{ij} \leq r \leq \max_i \sum_j A_{ij}$$
- *Let*
$$f(x) = \min_{\{i|x_i \neq 0\}} \frac{[Ax]_i}{x_i}$$
Then $r = \max_x f(x)$.

2 Probabilities and Markov chains

We will see a few cases where probabilities come in handy to express uncertain events. For example, the PageRank algorithm can be seen as a random walk on a directed graph of webpages, with the importance score of a webpage being derived from the overall probability of landing on that webpage during the random walk.

So let's review some basic probabilities first!

2.1 Reviews on probability

2.1.1 Basic definitions

The first thing we need in a probabilistic setting is the **universe**, i.e the space of *possible events*. We usually write it as Ω. For the random walk on webpages, it is natural to take Ω as the set of all webpages, something like $\Omega = \{w_1, \cdots, w_n\}$ if we have n pages under consideration.

Then on this universe, we need a **σ-algebra $\mathcal{A}$**, which is the set of measurable sets we can assign a probability to. When the universe is *discrete*, which is the case when $\Omega = \{w_1, \cdots, w_n\}$, then we usually take $\mathcal{A} = \mathcal{P}(\Omega)$, i.e all the possible subsets of Ω. In fact, for a discrete space, the σ in σ-algebra is a bit superfluous, but we won't expand on this. When we have a universe Ω endowed with a σ-algebra $\mathcal{A}$, then we call $(\Omega, \mathcal{A})$ a **measurable space**.

Finally, we can define a probability $\mathbb{P}$ on our σ-algebra $\mathcal{A}$: it is a function that swallows a set $A \in \mathcal{A}$ and outputs a number between 0 and 1, the **probability of the event** A. We only need this function to satisfy two properties to call it a probability:

- $\mathbb{P}(\Omega) = 1$

- $\mathbb{P}(\cup_{i=1}^{+\infty} A_i) = \sum_{i=1}^{+\infty} \mathbb{P}(A_i)$ for $\{A_i\}_i$ disjoint events, i.e $i \neq j \Rightarrow A_i \cap A_j = \emptyset$ (this is called *σ-additivity*).

We then call $(\Omega, \mathcal{A}, \mathbb{P})$ a **measured space**.

Let's give some properties of this probability measure.

Proposition 2.1 (Properties of $\mathbb{P}$)**.** *Let $A, B \in \mathcal{A}$.*

1. $\mathbb{P}(\emptyset) = 0$.
2. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ *where A^c is the* ***complement*** *of A, $A^c = \Omega \backslash A$.*
3. $\mathbb{P}(A) \geq \mathbb{P}(B)$ *if $B \subseteq A$.*
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Independence is also a crucial notion in probabilities. Intuitively, event A and B are independent if one happening does not change the probabilities of the other one happening. It is a bit deeper than this but obviously it has a precise mathematical definition which we give now.

Definition 2.1 (Independence)**.** We say events $A, B \in \mathcal{A}$ are independent if and only if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

2.1.2 Conditional probability

The next thing we usually learn in probability class is **conditional probability**. It makes precise the idea that an event happening might give us some information on the probability of another event.

Definition 2.2 (Conditional probability)**.** Let $A, B \in \mathcal{A}$. Then the conditional probability of A with respect to B is given by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

for $\mathbb{P}(B) \neq 0$.

Of course if two events are independent we have the following:

Proposition 2.2 (Indepence and conditioning)**.** *If A and B are independent then $\mathbb{P}(A|B) = \mathbb{P}(A)$.*

A useful formula is that of the **total probabilities**:

Proposition 2.3 (Total probabilities)**.** *Let $A \in \mathcal{A}$ and $\{B_i\}_i \subseteq \mathcal{A}$ a family of disjoint sets such that $\cup_{i=1}^{+\infty} B_i = \Omega$. Then*

$$\mathbb{P}(A) = \sum_{i=1}^{+\infty} \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

And finally Bayes' rule, that allows updating conditional probabilities:

Theorem 2.1 (Bayes' rule)**.** *For two events $A, B \in \mathcal{A}$, we have*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

2.1.3 Random variables

We often want to work on a simpler space than Ω. **Random variables** come in handy to express complicated events.

Definition 2.3 (Random variable)**.** A random variable X is a function from a measured space $(\Omega, \mathcal{A}, \mathbb{P})$ to a measurable space $(\Theta, \mathcal{B})$ such that

$$\forall B \in \mathcal{B},\ X^{-1}(B) \in \mathcal{A}$$

where $X^{-1}(B) = \{\omega \in \Omega; X(\omega) \in B\}$.

This is a bit technical! Not to worry: we will use random variables (and more precisely *sequences* of random variables) when we will model the random walk on webpages, so the spaces will be simple to understand. For example, we will look at probabilities such as $\mathbb{P}(X_i = w)$, the probability that at time i we are on webpage w, or $\mathbb{P}(X_{i+1} = w | X_i = \tilde{w})$, the probability that I'll land on page w if I am already on $\tilde{w}$.

2.2 Markov chains

Markov chains are one of the most widely encountered stochastic models in the literature. It has a few simple properties that make it very useful to model *memoryless* processes. What this means is that if you look at a random process $\{X_i\}$ (e.g, your position on the network of webpages), your position at time $i+1$ only depends on your position at time i, and not on times 0, 1, ..., $i-1$. It is often a simplifying assumption, but what we may lose in precision for the model we gain by having tractable computations and insights.

2.2.1 Representing the state space

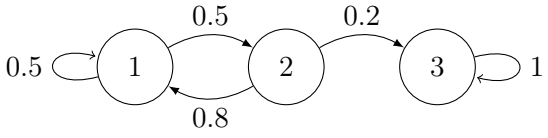
The first thing to define for a Markov process is the **state space** S: it is the space on which our stochastic process will move. These spaces come in many forms and sizes, but we only need the most simple of them: a *discrete* and *finite* state space. This means that we only have a finite number of positions that our stochastic process can take, e.g, the n webpages on our network.

Once we have the states, we define the **transition probabilities** p_{ij}: this is the probability that being in state i, I will be in state j in the next step. One requirement (which is natural) is that

$$\sum_{j \in S} p_{ij} = 1$$

This is so because from state $i \in S$, we can only go to any other state $j \in S$.

We often represent Markov chain using directed networks. Suppose we have three states, $S = \{1, 2, 3\}$ and $p_{11} = 0.5, p_{12} = 0.5, p_{21} = 0.8, p_{23} = 0.2, p_{33} = 1$. Then we can represent the chain as such:



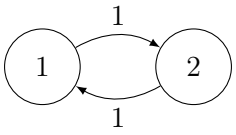
This is much clearer this way! So suppose the process starts at state 2. What is the probability that in the next step I'll be on 1? We read the number on the edge linking 2 to 1: it is 0.8. What is the probability that I'll do the following path: $\{2, 1, 1, 2, 3\}$? Simply $0.8 \times 0.5 \times 0.5 \times 0.5 \times 0.2$.

You see that not all states are equal here: if we land on 3, we are stuck here! We say that state j is **reachable** from i if we have a non-negative probability of ever reaching j from i. We write it as $i \to j$. Another definition, easier when we have the directed graph above, is that j is reachable from i if and only if there exists a directed path from i to j in the graph. If we use this definition, then we have to omit the edges that have zero probability (such as 3 to 2 in the example). 3 is reachable from any other state (including itself), but 1 and 2 are not reachable from 3. In fact we always have $i \to i$ (all states are reachable from themselves).

We say that two states i, j **communicate** if $i \to j$ and $j \to i$. It is written as $i \leftrightarrow j$. In the example, we have $1 \leftrightarrow 2$. The communication property forms an equivalence class, in the sense that $i \leftrightarrow i$, $i \leftrightarrow j \Leftrightarrow j \leftrightarrow i$ and $i \leftrightarrow j, j \leftrightarrow k \Rightarrow i \leftrightarrow k$. So two states that communicate are part of the same equivalence class.

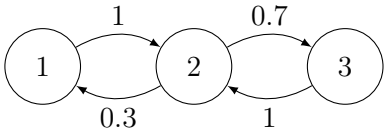
A **transient** state is one that eventually does not get visited again. In the example, 1 and 2 are transient, since there is a nonzero probability of landing on 3 at some point. When we arrive there, we never visit 1 and 2 again. On the other hand, 3 is a **recurrent** state: with probability one, after a visit to 3, we will come back to it (in that case we come back immediately).

In the following example, both states are recurrent:

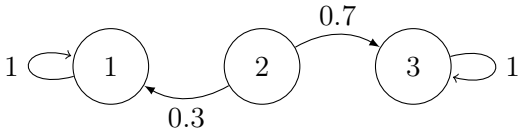


In fact in that case the chain is **periodic**: say at time $t = 1$, we start the process at state 1. Then we will only ever visit 1 at odd times, and 2 at even times. So the period of the chain is 2.

An **irreducible** Markov chain is one that is only composed of *one* communicating class. As a corollary, all states of this Markov chain are recurrent. We give below two examples, one of a one of an irreducible and one of a non irreducible Markov chain.



(a) An irreducible Markov chain



(b) A non irreducible Markov chain

2.2.2 Matrix representation of a Markov chain

The notation p_{ij} should tell us that we can also represent the Markov chain in matrix form. These matrices are called **stochastic matrices**, which simply means that the sum of the coefficients on a row adds up to one, for every row.

For example, the matrix representation of figure (a) above would be

$$P = \begin{pmatrix} 0 & 1 & 0 \\ 0.3 & 0 & 0.7 \\ 0 & 1 & 0 \end{pmatrix}$$

Since for every i, $\sum_{j \in S} p_{ij} = 1$, we have that P is indeed a stochastic matrix.

To see why the matrix representation is nice, let $\eta^{(1)}$ be a *row* vector with $|S|$ entries whose components add up to 1 ($\sum_{i \in S} \eta_i^{(1)} = 1$). We will say that $\eta_i^{(1)}$ is the probability that the chain starts ($t = 1$) at state i. In other words, if the random variable X_k is the state that the process occupies at time $t = k$, then $\mathbb{P}(X_1 = i) = \eta_i^{(1)}$.

We would like to know the distribution of X_2, call it $\eta^{(2)}$. This distribution is simply

$$\eta^{(2)} = \eta^{(1)} P$$

Again, taking figure (a) for example, let $\eta^{(1)} = (0.25, 0.5, 0.25)$. Then, we have probability 0.5 of starting at state 2, i.e $\mathbb{P}(X_1 = 2) = 0.5$. Now compute

$$\eta^{(1)} P = \begin{pmatrix} 0.25 & 0.5 & 0.25 \end{pmatrix} \times \begin{pmatrix} 0 & 1 & 0 \\ 0.3 & 0 & 0.7 \\ 0 & 1 & 0 \end{pmatrix} = (0.15, 0.5, 0.35) = \eta^{(2)}$$

Indeed, how can we arrive on state 2 at time 2? We can either start at 1 and then move to 2 (probability: $\eta_1^{(1)} \times p_{12} = 0.25 \times 1$) or start at 3 and move to 2 (probability: $\eta_3^{(1)} \times p_{32} = 0.25 \times 1$), which adds up to 0.5. In other words, $\eta_2^{(2)} = \sum_{j \in \{1,2,3\}} \eta_j^{(1)} p_{j2}$, which is simply one component of the matrix multiplication.

It seems to become trickier for larger t. Let's say after 10 iterations you want to know the probability of being on state 1. There are a lot of different paths that lead you there! Fortunately, we only have to iterate our formula above:

$$\eta^{(10)} = \eta^{(9)} \times P = \eta^{(8)} \times P \times P = \cdots = \eta^{(1)} \times P^9$$

2.2.3 The steady-state problem

Often we want to be able to answer the question: "If I sample the random process at any time, what is the probability that I will be at state s?" Another way to phrase it is "What is the frequency of being in state s?", i.e if I let the chain run for a long enough amount of time, and report the frequencies of time spent on the different states, do these frequencies stabilize?

One example where this can be useful is if you are running a machine whose state follows a Markov chain, and the cost of operating the machine is dependent of the state. You would like to know on average at each unit of time how much the machine will cost you. Say it costs 10\$ to run it for one unit of time if it is in Working condition W, 20\$ if it is Clunky C and 50\$ if it is completely broken and waiting for repairs B. Then if you know the frequencies of being in each of these states, given by say (π_W, π_C, π_B), you can get an average running cost of

$$10\pi_W + 20\pi_C + 50\pi_B$$

per unit of time.

If you read the chapter 3 in the book, you have perhaps recognized that this elusive π is none other than the importance score given by Google to a webpage, for a certain transition matrix. You can read the notes on the book for chapter 3 for more information between the two.

There are quite a few properties that the Markov chain needs to satisfy if we hope to have these "stable" frequencies π, and even more so when the state space is not finite but simply countable. We only look at the case of a finite state space though.

In that case we have our transition matrix P, and we will be looking at the following equation:

$$\pi = \pi P$$

We take π to be a row vector. If we can find such a π, we call it the **invariant measure**: if π describes the initial condition, then at the next step the probability of being on state i is the same as that of starting at i. You can also think of it as a sort of fixed point.

If the Markov chain is **irreducible** and **aperiodic** (none of the states have a period greater than 1), then there exists a unique π satisfying the equation above. It also satisfies the following property:

Proposition 2.4 (Limit of the transition matrix)**.** *For all η initial distribution of the states, if π is an invariant measure of P and P is irreducible and aperiodic then*

$$\lim_{n \to +\infty} \eta P^n = \pi$$

Another point of view worth looking at is that the equation $\pi = \pi P$ means that π is a *left-eigenvector* of P for the eigenvalue 1. In the linear algebra review, we have defined eigenvectors as *right-eigenvectors*, but most of our theorems and intuitions still hold (since left-eigenvectors of a matrix A are right-eigenvectors of the transpose of A).

In particular, there exists a version of the Perron-Froebenius theorem that states that for non-negative transition matrices that are irreducible and aperiodic, there exists a right eigenvector associated with the largest eigenvalue of the matrix (here, 1) such that all its entries are positive. This means that all components of π are positive, which we would expect since the matrix is irreducible.