Eigenvectors of Adjacency Matrices and Markov Chains

Barum Park

Department of Sociology
Cornell University
b.park@cornell.edu

March 8, 2021

Introductory Notes

In this note, we'll go through eigenvalues/eigenvectors and Markov chains. The treatment of these topics will be quite superficial and focuses only on the elements that are needed to understand further materials that will be discussed in our class. In particular, we consider mainly simple graphs that are undirected and connected. This implies that we won't need to deal with transient or absorbing states, communication classes, and recurrence when discussing Markov chains (at least for our purposes).

The note will become "a bit" more formal towards the end in order to keep track of the results. I hope students who have only limited exposure to mathematical texts will not be driven away by the terms "lemma" or "corollary." You can think of these as bookkeeping devices, to prevent statements that start with "as we have discussed in third line of the first paragraph that appeared two pages ago..." In the end, if you have understood the note up to the section "Stationary and Limiting Distributions," you can just jump to Theorem 3 in case you don't want to go into the nitty-gritty of Markov chains.

Most statements are accompanied with a proof in the appendix that I believe are accessible without knowledge of complex numbers, analysis/measure theory, or linear algebra beyond the superficial level. So, I believe most students will be able to follow them if they invest enough time. Given that, the proofs are by no means the main point of this note. They are there for the curious and can be skipped. The same is true for the `R` code: it is included for those who might be interested but is not essential and can be skipped.

In preparing the notes, I've consulted Levin et al. 2008. *Markov Chains and Mixing Times* for the discussion of Markov chains. However, I do not recommend this book to sociology students. The authors say that the material requires only undergraudate-level knowledge, but the undergraduates they have in mind are majoring in mathematics not sociology. In general, I've tried to make the

material more accessible to students, which basically means that the discussion will be much more verbose and lengthy (e.g., irreducibility and aperiodicity cover not even a single page in the book).

All matrices and vectors are written in **bold** font to distinguish them from scalars (1×1 matrices; you can think of them as real numbers). All vectors, $\mathbf{v}$, will be assumed to be column vectors; when we want to denote a row vector, we'll use $\mathbf{v}^\top$, where $^\top$ stands for the transpose. All matrices will be assumed to be real, i.e., all elements a_{ij} of the matrix $\mathbf{A}$ are assumed to be real numbers.

It is assumed that students know what an matrix inverse is. The (multipicative) inverse of a non-zero real number x is the "reciprocal" of x—i.e., x^{-1} or $1/x$. If you multiply x by it's inverse, you'll get the (multiplicative) identity—i.e., $x \times x^{-1} = x^{-1} \times x = 1$. Similarly, if you have a square matrix $\mathbf{A}$, the inverse of $\mathbf{A}$ is a matrix $\mathbf{A}^{-1}$, such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Not all matrices have an inverse. If the inverse of $\mathbf{A}$ exists, $\mathbf{A}$ is called *invertible*, otherwise $\mathbf{A}$ is *singular*. Two properties of the matrix inverse that'll be used below are the following:

1. A square matrix $\mathbf{A}$ is invertible if and only if its determinant is non-zero.
2. If $\mathbf{A}$ and $\mathbf{B}$ are both invertible square matrices of the same dimension, then the matrix $\mathbf{C} = (\mathbf{AB})$ is invertible as well and is equal to $\mathbf{C}^{-1} = (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.

Eigenvectors

Let $\mathbf{A}$ be a $n \times n$ (square) matrix of real numbers. For every such matrix $\mathbf{A}$, we can find n scalars λ_i and non-zero vectors $\mathbf{x}_i$, such that

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i, \qquad i = 1, 2, ..., n. \tag{1}$$

The vectors $\mathbf{x}_i$ are called *eigenvectors* of $\mathbf{A}$ and λ_i's are called the corresponding *eigenvalues*.[1]

Notice that eigenvectors are not uniquely defined: when $\mathbf{x}$ is an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ, then the pair $(\lambda, c\mathbf{x})$ for any non-zero c will also form a valid *eigenpair*, since it satisfies equation (1). Out of this reason, we often require that the eigenvectors have a *length* (or *norm*) of 1, which means that all eigenvectors satisfy $\|\mathbf{x}_i\|_2 = \sqrt{\mathbf{x}_i^\top \mathbf{x}_i} = 1$. When we talk about eigenvectors of a matrix $\mathbf{A}$, these "normalized" versions are often meant. But when talking about Markov chains, which we will do later in this note, eigenvectors (assuming that they have non-negative elements) are usually normalized to sum to one in order to represent probability vectors.[2] What normalization is implied will be evident from the context.

To understand what eigenvectors are, it might be good to recap how matrices transform vectors. Whenever you multiply a vector $\mathbf{v}$ by the matrix $\mathbf{A}$, you will get another vector $\mathbf{v}^* = \mathbf{A}\mathbf{v}$, where $\mathbf{v}$

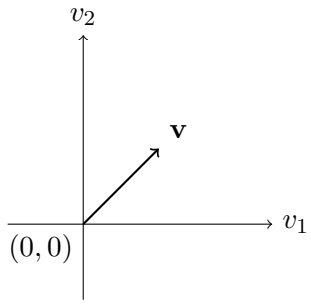
[1] Even when $\mathbf{A}$ is a real matrix, λ_i and $\mathbf{x}_i$ will be, in general, complex. But we won't deal with this issue in this note.

[2] A probability vector $\mathbf{p}$ is a vector that consists of non-negative elements the sum of which is equal to one. Hence, it represents the probability distribution over a finite set.

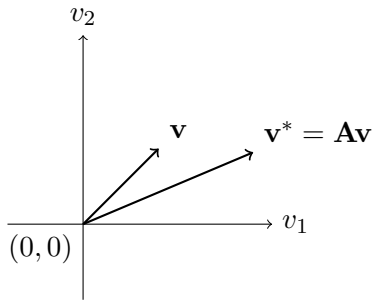
and $\mathbf{v}^*$ are of the same dimension. In general, $\mathbf{v}$ and $\mathbf{v}^*$ will point in different directions, which means that the matrix $\mathbf{A}$ stretches and rotates the vector $\mathbf{v}$ into $\mathbf{v}^*$. Consider, for example,

$$\mathbf{v} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \text{ and } \mathbf{A} = \begin{bmatrix} 1 & 1.25 \\ 0.2 & 0.75 \end{bmatrix}.$$

The vector $\mathbf{v}$ will point from the origin $(0, 0)$ to $(1, 1)$ and multiplying $\mathbf{v}$ from the right by $\mathbf{A}$ will transform this vector into $\mathbf{v}^*$, which points to $(2.25, 0.95)$ (see plot below).



Vector $\mathbf{v}$ points from $(0, 0)$ to $(1, 1)$



Vector $\mathbf{v}^*$ points from $(0, 0)$ to $(2.25, 0.95)$.

We see that the length of $\mathbf{v}$ is stretched by $\mathbf{A}$ and the direction it points to is changed as well.

A quite interesting question to ask is the following: "are there any vectors the direction of which is *not* changed when multiplied by $\mathbf{A}$?" These directions are what eigenvectors stand for. If you look again into equation (1), you'll discover that transforming the eigenvector $\mathbf{x}$ by the matrix $\mathbf{A}$ (i.e., $\mathbf{Ax}$) is the same as multiplying $\mathbf{A}$ by a scalar λ. But multiplying a vector by a scalar can only stretch the vector $\mathbf{x}$; it will not rotate it in any way and, hence, not change its direction (except for flipping it 180 degrees).

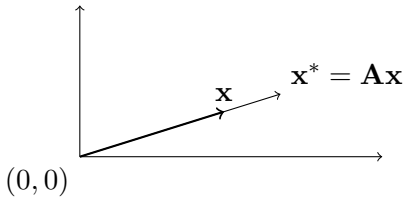
For example, the first eigenvector of $\mathbf{A}$ and its corresponding eigenvalue are

$$\mathbf{x} = \begin{bmatrix} 0.9545 \\ 0.2981 \end{bmatrix} \text{ and } \lambda = 1.3904.$$

So,

$$\mathbf{x}^* = \mathbf{Ax} = (1.3904)\mathbf{x} = \begin{bmatrix} 1.3272 \\ 0.4145 \end{bmatrix}$$

which means that $\mathbf{A}$ is just stretching $\mathbf{x}$ by about 40% further in the same direction, as shown below:



Multiplying an eigenvector $\mathbf{x}$ of $\mathbf{A}$ by the matrix $\mathbf{A}$ itself will not change the direction of $\mathbf{x}$.

Right and Left-eigenvectors

What we have defined as *the* eigenvectors of $\mathbf{A}$ in equation (1) are sometimes called the *right-eigenvectors*. On the other hand, the *left-eigenvectors* are defined by the equations

$$\mathbf{x}_i^\top \mathbf{A} = \lambda_i \mathbf{x}_i^\top, \qquad i = 1, 2, ..., n. \tag{2}$$

Transposing both sides of (2), it is easy to see that the left-eigenvectors of $\mathbf{A}$ are the right-eigenvectors of $\mathbf{A}^\top$ and vice versa. A result that is not so obvious, but which we'll need later, is the following:

Lemma 1. *Let $\mathbf{A}$ be a square matrix. Then $\mathbf{A}$ and $\mathbf{A}^\top$ have the same eigenvalues. In other words if λ is an eigenvalue of $\mathbf{A}$, then it is an eigenvalue of $\mathbf{A}^\top$ as well.*

The set of all eigenvalues of a matrix $\mathbf{A}$, $\{\lambda_1, ..., \lambda_n\}$ is called the *spectrum* of $\mathbf{A}$. So, whenever you hear a something with spectrum or spectral (e.g., "spectral clustering," "spectral graph theory," etc.) it often deals with the eigenvalues/eigenvectors of matrices.

Why are Eigenvectors Useful?

Eigenvectors and eigenvalues turn up everywhere in applications of linear algebra. Here we will consider one important property that will be useful in considering Markov chains.

It is often convenient to consider all n eigenvector-eigenvalue pairs simultaneously as follows:

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{\Lambda} \tag{3}$$

where

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n] \quad \text{and} \quad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix},$$

i.e., $\mathbf{X}$ is a matrix where the ith column is the ith eigenvector and $\mathbf{\Lambda}$ is a diagonal matrix where the (i, i)th entry is λ_i and the rest is zero (you should verify that the ith column of both sides of the equation is $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$.)

Now, if the matrix of eigenvectors $\mathbf{X}$ is *invertible* (i.e., the inverse-matrix $\mathbf{X}^{-1}$ exists), then we can write the matrix $\mathbf{A}$ in terms of its eigenvectors and eigenvalues by post-multiplying both sides of (3) by $\mathbf{X}^{-1}$:

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}, \tag{4}$$

which is called the *eigendecomposition* of the matrix $\mathbf{A}$. Square matrices for which the eigendecomposition exists (i.e., for which $\mathbf{X}$ is invertible) are called *diagonalizable*.

The eigendecomposition has lots of merits. For example, we can write the inverse of $\mathbf{A}$ as

$$\mathbf{A}^{-1} = \Big(\mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1})^{-1} = \mathbf{X}\mathbf{\Lambda}^{-1}\mathbf{X}^{-1},$$

where $\mathbf{\Lambda}$ will have λ_i^{-1} in the ith diagonal. Also, raising the matrix $\mathbf{A}$ to powers becomes very easy. For example:

$$\mathbf{A}^2 = \mathbf{A}\mathbf{A} = (\mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1})(\mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}) = \mathbf{X}\mathbf{\Lambda}\mathbf{I}\mathbf{\Lambda}\mathbf{X}^{-1} = \mathbf{X}\mathbf{\Lambda}^2\mathbf{X}^{-1},$$

where $\mathbf{I}$ is the $n \times n$ identity matrix and

$$\mathbf{\Lambda}^2 = \mathbf{\Lambda}\mathbf{\Lambda} = \begin{bmatrix} \lambda_1^2 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n^2 \end{bmatrix}.$$

Multiplying A one more time gives

$$\mathbf{A}^3 = \mathbf{A}^2\mathbf{A} = (\mathbf{X}\mathbf{\Lambda}^2\mathbf{X}^{-1})(\mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}) = \mathbf{X}\mathbf{\Lambda}^3\mathbf{X}^{-1}$$

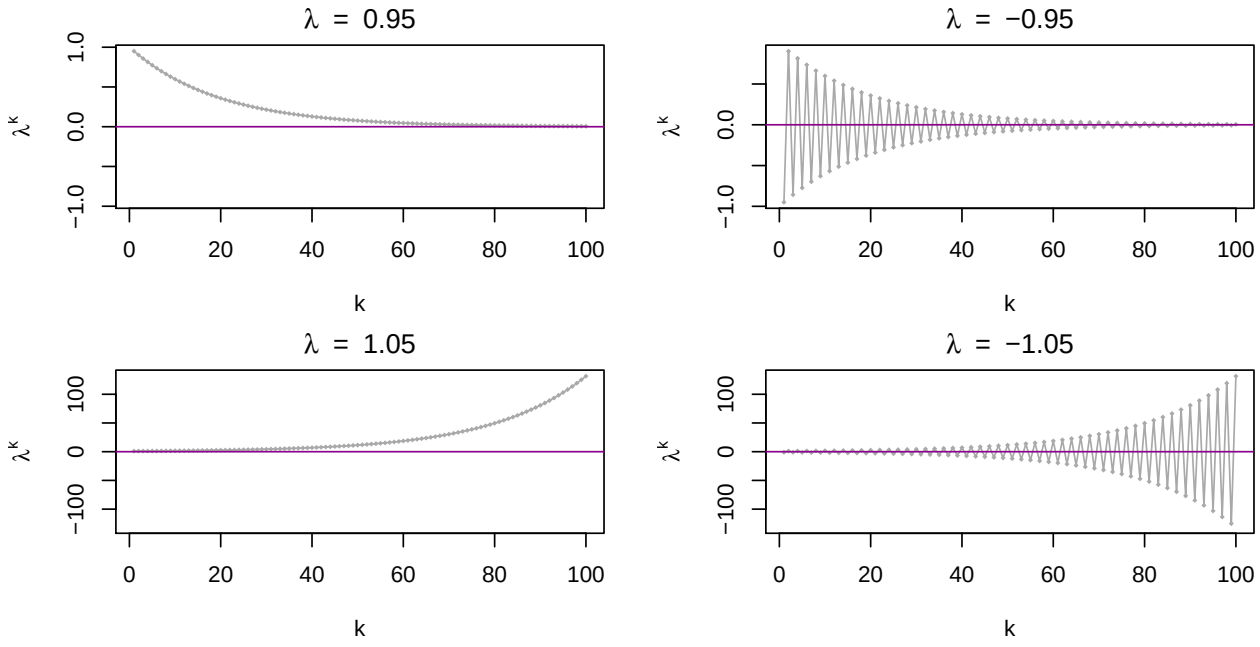
and, in general, we see that

$$\mathbf{A}^k = \mathbf{X}\mathbf{\Lambda}^k\mathbf{X}^{-1}.$$

So, using the decomposition in (4), raising the matrix $\mathbf{X}$ to a power k amounts to nothing else then raising the eigenvalues of $\mathbf{A}$ to that power.

Now, suppose an eigenvalue λ_i is strictly smaller in magnitude then 1, i.e., $|\lambda_i| < 0$. Recall from basic algebra that each time you multiply a number smaller in magnitude than 1 by itself, the result will get closer and closer to zero until it converges to 0. That is, $|\lambda_i|^{k+1} < |\lambda_i|^k$ for $k > 0$. On the other hand, if $|\lambda_i| > 1$, then $|\lambda_i|^{k+1} > |\lambda_i|^k$ for $k > 0$.[3] Below is a plot that shows how eigenvalues will behave when raised to larger and larger powers:

[3] Of course, this holds also for complex numbers, but let us assume that all eigenvalues are real.



Hence, we see that if we self-multiply $\mathbf{A}$ an infinite amount of time, the result will depend on the magnitude of the eigenvalues of $\mathbf{A}$, i.e.,

$$\begin{aligned}\lim_{k\to\infty} \mathbf{A}^k &= \mathbf{X}\left(\lim_{k\to\infty} \mathbf{\Lambda}^k\right)\mathbf{X}^{-1} \\ &= \mathbf{X}\begin{bmatrix} \lim_{k\to\infty}\lambda_1^k & 0 & \cdots & 0 \\ 0 & \lim_{k\to\infty}\lambda_2^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lim_{k\to\infty}\lambda_n^k \end{bmatrix}\mathbf{X}^{-1}. \end{aligned} \tag{5}$$

For eigenvalues that are greater in magnitude than 1, the corresponding diagonal element of $\mathbf{\Lambda}$ will "explode," while eigenvalues with smaller magnitude than 1 will "vanish" to zero. For example, if all eigenvalues are smaller than 1 in magnitude, then $\lim_{k\to\infty}\mathbf{A}^k = \mathbf{0}$, where $\mathbf{0}$ is a $n \times n$ square matrix of zeros. Any "interesting" convergence result will, therefore, be achieved only if the matrix has a subset of eigenvalues that are equal to 1 in absolute value, while other eigenvalues are smaller than 1 in magnitude. These considerations will be important when we discuss Markov chains on networks.

An Application: Eigenvector Centrality

One application in which powers of $\mathbf{A}$ play a major role is *eigenvector centrality*, first proposed by Philip Bonacich in "Factoring and weighting approaches to status scores and clique identification" published in the *Journal of Mathematical Sociology* in 1972. Let $G(V, E)$ be a graph and $\mathbf{A}$ its associated adjacency matrix. We assume that the graph is connected (each node is reachable from all others via a path) and undirected (ties a symmetric). Before introducing the measure, we state

an important theorem from linear algebra, which we will not prove here.[4]

Theorem 1 (Spectral Theorem)**.** *Let* $\mathbf{A}$ *be real symmetric square matrix of dimension* $n \times n$*. Then all eigenvalues of* $\mathbf{A}$ *are real and there exists an orthonormal basis of* $\mathbb{R}^n$ *consisting of eigenvectors of* $\mathbf{A}$*.*

For our current purpose, the term "orthonormal basis" can be understood as saying that the matrix of eigenvectors $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]$ has an inverse, $\mathbf{X}^{-1}$, and that this inverse is equal to the transpose of $\mathbf{X}$, i.e., $\mathbf{X}^{-1} = \mathbf{X}^\top$. So, if $\mathbf{A}$ is symmetric (which it will be for undirected graphs), we can rewrite the eigendecomposition in equation (3) as

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^\top = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^\top,$$

which is referred to as the *spectral decomposition* of symmetric square matrices (you should verify that the last equality works; notice that $\mathbf{x}_i\mathbf{x}_i^\top$ is a $n \times n$ matrix).

Bonacich was concerned with measures of popularity. He starts with the first-order popularity measure defined as

$$\rho_i^{(1)} = \sum_{j=1}^{n} a_{ij}$$

where a_{ij} is the (i, j)th entry of the adjacency matrix $\mathbf{A}$. So we see that this measure is nothing but the degree of node i (or how many other nodes in the network are connected to i by a walk of length 1). The second-order popularity measure is about the degree of each nodes' neighbors (how many other nodes are connected to i in walks of length 2)

$$\rho_i^{(2)} = \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ik} a_{kj} = \sum_{j=1}^{n} a_{ij}^{(2)},$$

where $a_{ij}^{(2)}$ is the (i, j)th entry of $\mathbf{A}^2 = \mathbf{A}\mathbf{A}$. These equations get quickly out of control as we move to higher order measures of popularity. So let us switch to matrix notation, where we have for the first- and second-order popularity

$$\boldsymbol{\rho}^{(1)} = \begin{bmatrix} \rho_1^{(1)} \\ \rho_2^{(1)} \\ \vdots \\ \rho_n^{(1)} \end{bmatrix} = \mathbf{A}\mathbf{1} \quad \text{and} \quad \boldsymbol{\rho}^{(2)} = \begin{bmatrix} \rho_1^{(2)} \\ \rho_2^{(2)} \\ \vdots \\ \rho_n^{(2)} \end{bmatrix} = \mathbf{A}^2\mathbf{1}$$

[4] After adding the proof, I thought it's too much. For the interested student, you can find a accessible proof in Lemma 1 (eigenvalues are real) and Lemma 4 (eigenvectors form orthonormal basis) of Professor David P. Williamson's lecture notes on spectral graph theory.

where $\mathbf{1}$ is a vector of ones. The kth-order popularity measure is, thus,

$$\boldsymbol{\rho}^{(k)} = \mathbf{A}^k \mathbf{1}$$

Bonacich goes on to define the ∞-th order popularity measure as

$$\boldsymbol{\rho}^{(\infty)} = \lim_{k \to \infty} \mathbf{A}^k \mathbf{1}.$$

Of course, an infinte-length walk will connect all nodes to all others an infinite number of times, given that the graph is connected. This implies $\mathbf{A}^k$ will "blow up," and all entries of $\boldsymbol{\rho}^{(\infty)}$ will be simply ∞.

However, we can use a simple "trick" to still obtain a meaningful measure that reflects the relative importance of each node. Consider the matrix $\mathbf{B} = \lambda_1^{-1}\mathbf{A}$, where λ_1 is the largest eigenvalue of $\mathbf{A}$. The matrix $\mathbf{B}$ will have the same structure as $\mathbf{A}$ and, thus, contains the same relational information, but every entry that is equal to 1 will be replaced by $1/\lambda_1$. Notice as well that the eigenvectors of $\mathbf{B}$ are the same as those of $\mathbf{A}$ and only the eigenvalues change (if we define the eigenvectors to have length equal to one). To see this, let $\mathbf{x}_i$ be an eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ_i. Then,

$$\mathbf{B}\mathbf{x}_i = (\lambda_1^{-1}\mathbf{A})\mathbf{x}_i = \lambda_1^{-1}(\mathbf{A}\mathbf{x}_i) = \lambda_1^{-1}\lambda_i\mathbf{x}_i = (\lambda_i/\lambda_1)\mathbf{x}_i,$$

which shows $\mathbf{x}_i$ is an eigenvector of $\mathbf{B}$ with corresponding eigenvalue $\mu_i = \lambda_i/\lambda_1$.

So, instead of deriving the ∞-order popularity measure of based on $\mathbf{A}$, Bonacich uses $\mathbf{B}$ for this purpose. Using the spectral decomposition of $\mathbf{B}$, we see that

$$\begin{aligned}
\lim_{k\to\infty} \mathbf{B}^k &= \lim_{k\to\infty} \sum_{i=1}^{n} \mu_i^k \mathbf{x}_i \mathbf{x}_i^\top \\
&= \lim_{k\to\infty} \left[\mu_1^k \mathbf{x}_1\mathbf{x}_1^\top + \mu_2^k \mathbf{x}_2\mathbf{x}_2^\top + \cdots + \mu_n^k \mathbf{x}_n\mathbf{x}_n^\top\right] \\
&= \lim_{k\to\infty} \left[\left(\frac{\lambda_1}{\lambda_1}\right)^k \mathbf{x}_1\mathbf{x}_1^\top + \left(\frac{\lambda_2}{\lambda_1}\right)^k \mathbf{x}_2\mathbf{x}_2^\top + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^k \mathbf{x}_n\mathbf{x}_n^\top\right] \\
&= \left[\lim_{k\to\infty} \left(\frac{\lambda_1}{\lambda_1}\right)^k\right] \mathbf{x}_1\mathbf{x}_1^\top + \left[\lim_{k\to\infty} \left(\frac{\lambda_2}{\lambda_1}\right)^k\right] \mathbf{x}_2\mathbf{x}_2^\top + \cdots + \left[\lim_{k\to\infty} \left(\frac{\lambda_n}{\lambda_1}\right)^k\right] \mathbf{x}_n\mathbf{x}_n^\top \\
&= \mathbf{x}_1\mathbf{x}_1^\top.
\end{aligned}$$

The last step follows from $\lim_{k\to\infty}(\lambda_i/\lambda_1)^k = 0$ for all other terms except the first, because $\lambda_1 > |\lambda_i|$ for all i. So, the *eigenvector centrality* is defined as

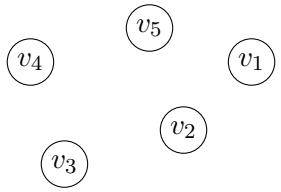
$$\mathbf{e} = \mathbf{B}^\infty \mathbf{1} = \mathbf{x}_1\mathbf{x}_1^\top\mathbf{1} = \mathbf{x}_1$$

which is just the eigenvector corresponding to the largest eigenvalue of the adjacency matrix $\mathbf{A}$. The only problem with this derivation is that the largest eigenvalue of $\mathbf{A}$ might not be unique (i.e.,

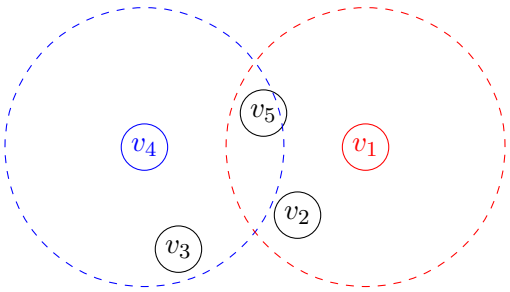
there might be multiple eigenvalues equal to λ_1). Bonacich just "assumes" that λ_1 is largest in magnitude and positive. Below, we'll see that this is a reaosnable assumption, given that the graph G is connected and contains at least one odd-length cycle.

Markov Chains: Throwing Frisbee at Stewart Park

With eigenvalues/eigenvectors at hand, let us move on to Markov chains. Consider a group of 5 network scholars that play frisbee. They visit Stewart Park and arrange themselves in the following positions:

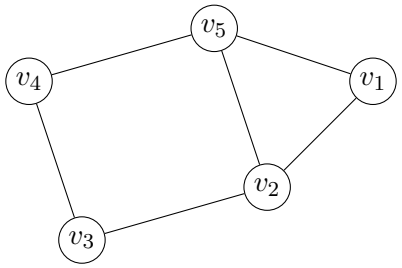


Spending too much time on research and not enough on frisbee practice, these scholars are extremely bad in throwing the disk around. The maximal distance they can throw the disk is given in the following plot (only for v_1 and v_4).



So, v_1 is able to throw the disk to v_2 and v_5 but not to v_3 or v_4; v_4, on the other hand, is able to reach v_3 and v_5, but not the rest. Assuming that all of these scholars are equally bad in frisbee, we can represent who can pass the disk to whom by a graph of the following form:

Figure 1: Network representation of who can throw the frisbee to whom



and the corresponding adjacency matrix

$$\mathbf{A} = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{array} \begin{array}{c} \begin{array}{ccccc} v_1 & v_2 & v_3 & v_4 & v_5 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \end{bmatrix} \end{array}$$

Now, suppose that each time someone receives the frisbee, she'll throw it to one of her neighbors at random. For example, if v_4 has the frisbee, she'll throw it to nodes v_3 or v_5 with probability 0.5 and 0.5; if v_2 has the frisbee, he'll throw it to v_3, v_5 or v_1 with probability $1/3, 1/3$, and $1/3$, respectively. We see that the probability that a node v_i will throw the frisbee to any of its neighbors is inversely proportional to its degree: the more neighbors a node has, the less likely that each of the neighbors will receive the frisbee on the next turn.

Transition Matrix

Let's formalize these statements a bit more. Let Y_t be a random variable that takes on values in the vertex set of the graph $G(V, E)$—$Y_t \in V = \{v_1, ..., v_n\}$—where $\{Y_t = v_k\}$, $v_k \in V$, represents the event that node v_k receives the frisbee at time t. We will also use the shorthand notation $\{Y_t = k\}$ or $\{Y_t = k_t\}$ for this event. The sequence of random variables $\{Y_0, Y_1, ..., Y_t, ...\}$ form a stochastic process that describes the movement of the frisbee. An important assumption that we've made is that the location of the frisbee at time $t + 1$ depends only on where it is at time t, instead of the whole history preceding t. In other words, we have assumed that

$$\Pr[Y_{t+1} = k_{t+1} \,|\, Y_t = k_t, Y_{t-1} = k_{t-1}, ..., Y_0 = k_0] = \Pr[Y_{t+1} = k_{t+1} \,|\, Y_t = k_t].$$

Stochastic processes that have this structure are said to have the *Markov property* and are called *Markov chains*.

The *conditional* probabilities of where the frisbee will move at time $t + 1$ given that v_i has it in her hands at time t are called *transition probabilities*. For the frisbee-throwing example we have

$$p_{ij} = \Pr[Y_{t+1} = v_j \,|\, Y_t = v_i] = \begin{cases} d(v_i)^{-1}, & \text{if } a_{ij} = 1 \\ 0, & \text{if } a_{ij} = 0, \end{cases} \tag{6}$$

where $d(v_i)$ is the degree of node v_i. In other words, if $\{v_i, v_j\}$ is an edge in the graph, $\{v_i, v_j\} \in E$, the probability that v_i will pass on the frisbee to v_j is inversely-proportional to the degree of v_i; if v_i and v_j do not have an edge, then the probability of passing on the frisbee from v_i to v_j is zero (the scholars cannot throw that far).

The set of values that Y_t can take on (here the vertex set V) is called the *state space* of the Markov chain. As the state-space is finite, we can represent all the transition probabilities compactly as a matrix

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}, \tag{7}$$

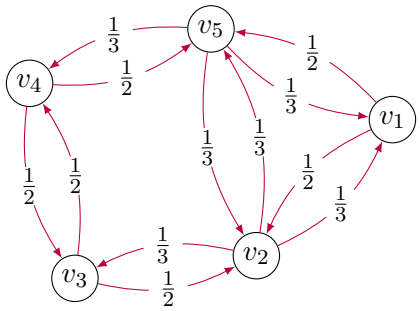
where $\mathbf{D}$ is a diagonal matrix with (i,i)th entry equal to the degree of node v_i.[5] The matrix $\mathbf{P}$ is called the *transition matrix* of the Markov chain and, for our example graph, it will be

$$\mathbf{P} = \begin{array}{c} \\ v_1 \\ v_1 \\ v_1 \\ v_1 \\ v_1 \end{array} \begin{array}{c} \begin{array}{ccccc} v_1 & v_2 & v_3 & v_4 & v_5 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \end{bmatrix} \end{array}$$

Notice that each row of $\mathbf{P}$ sums to one, which should be the case since the ith row gives us the probability distribution that the frisbee will be thrown from v_i to any of the other nodes.

Another way to represent $\mathbf{P}$ is to use a directed graph, where we label the directed edges with the probabilities of transition:

Figure 2: Graph representation of Markov chain



There are two important things to notice about the transition matrix of this example. First, we note that the matrix $\mathbf{P}$ does not change over time. This means that the probability that v_i will pass the frisbee to v_j remains the same, even after infinite passages of the frisbee. A Markov chain where the transition probabilities do not change with t are called *time homogeneous*. Second, notice that the network in Figure 1 is undirected, which reflects our assumption that whenever v_i can throw

[5] If this does not make immediate sense to you, convince yourself by creating multiplying these matrices by hand. You should get that p_{ij} is the probability of moving from i to j. The intuition is as follows: pre-multiplying $\mathbf{A}$ by $\mathbf{D}^{-1}$ has the effect of multiplying the ith row of $\mathbf{A}$ by the ith diagonal entry of $\mathbf{D}^{-1}$, which is d_i^{-1}. Whenever the jth column of the ith row of $\mathbf{A}$—i.e., a_{ij}—is equal to zero, it will stay zero; when $a_{ij} = 1$, on the other hand, this value will be switched to the d_i^{-1}.

the frisbee to v_j, v_j can throw it to v_i as well. Hence, even though $\mathbf{P}$ is *not* a symmetric matrix, $p_{ij} > 0$ implies that $p_{ji} > 0$. This will simplify our analysis of the Markov chains considerably.

Marginal Probabilities

Now, suppose that the first scholar to throw the frisbee is node v_1. As v_1 will hold the frisbee with probability one, the probability distribution of where the frisbee is located at time 0 can be represented by the vector $\boldsymbol{\pi}_0^\top = [1, 0, 0, 0, 0]$. Next, we know that at time 1, the node to receive the frisbee is either v_2 or v_5, each with probability 1/2. Using $\mathbf{P}$, we can calculate this as[6]

$$\boldsymbol{\pi}_1^\top = \boldsymbol{\pi}_0^\top \mathbf{P}.$$

Of course, at this point ($t = 1$), we are not sure anymore whether v_2 or v_5 holds the frisbee, which makes calculations of where it will be at time $t + 2$ not as straightforward. To calculate the vector $\boldsymbol{\pi}_2$, we use the Law of Total Probability. For example, the probability that the frisbee will be in the hands of v_1 is

$$\begin{aligned}\Pr[Y_2 = 1] &= \sum_{k_1=1}^{n} \underbrace{\Pr[Y_2 = 1 \,|\, Y_1 = k_1]}_{\text{the } (k_1, 1)\text{th entry of } \mathbf{P}} \times \underbrace{\Pr[Y_1 = k_1]}_{\text{the } k_1\text{th entry of } \boldsymbol{\pi}_1} \\ &= \boldsymbol{\pi}_1^\top \mathbf{p}_1,\end{aligned}$$

where $\mathbf{p}_1$ is the ith column of $\mathbf{P}$. Similarly, $\Pr[Y_2 = 2] = \boldsymbol{\pi}_1^\top \mathbf{p}_2$, $\Pr[Y_2 = 3] = \boldsymbol{\pi}_1^\top \mathbf{p}_3$, and so on, and row-wise stacking all these results together, we obtain

$$\boldsymbol{\pi}_2^\top = \boldsymbol{\pi}_1^\top \mathbf{P}.$$

But we have just discovered that $\boldsymbol{\pi}_1^\top = \boldsymbol{\pi}_0 \mathbf{P}$. So, we could rewrite this as

$$\boldsymbol{\pi}_2^\top = \boldsymbol{\pi}_1^\top \mathbf{P} = (\boldsymbol{\pi}_0^\top \mathbf{P})\mathbf{P} = \boldsymbol{\pi}_0^\top \mathbf{P}^2.$$

In fact, because the past trajectory of the frisbee doesn't matter when determining the probability distribution of the next step, we can write the relationship between the marginal probabilities and the transition probabilities as

$$\boldsymbol{\pi}_t = \Big(\text{Prob. location at } t-1\Big) \times \Big(\text{given location at } t-1\text{, prob. of location in the next step}\Big),$$

or

$$\boldsymbol{\pi}_t^\top = \boldsymbol{\pi}_{t-1}^\top \mathbf{P}. \tag{8}$$

[6] Again, if this result doesn't make immediate sense, try to do the calculation by hand and check that you get indeed the vector $\boldsymbol{\pi}_1^\top = [0, .5, 0, 0, .5]$.

and, further,

$$\begin{aligned}\boldsymbol{\pi}_t^\top &= \boldsymbol{\pi}_{t-1}^\top \mathbf{P} \\ &= \boldsymbol{\pi}_{t-2}^\top \mathbf{P}^2 \\ &= \boldsymbol{\pi}_{t-3}^\top \mathbf{P}^3 \\ &\vdots \\ &= \boldsymbol{\pi}_0^\top \mathbf{P}^t.\end{aligned}$$

So, we have the result

$$\boldsymbol{\pi}_t^\top = \boldsymbol{\pi}_0^\top \mathbf{P}^t. \tag{9}$$

Stationary and Limiting Distribution

Now, we'll try to answer the question: after an infinite number of throws, what is the probability that v_i would hold the frisbee in her hands? Would this probability depend on the initial state (i.e., the scholar who throws the frisbee first)? These questions lead us to consider the limiting behavior of Markov chains. As it is always the case with limits, the discussion will be necessarily a bit more formal.

Sometimes a subtle distinction is made between a *stationary distribution* (or *steady state*) of a Markov chain and its *limiting distribution*. This might confuse you in the future (at least it was confusing to me when I first encountered these concepts), so let us make the distinction clear.

Suppose there is a probability vector $\boldsymbol{\pi}$ defined over the (finite) state space of a Markov chain that satisfies the balance equation

$$\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \mathbf{P}. \tag{10}$$

Then $\boldsymbol{\pi}$ is called a *stationary (or steady state) distribution* of the Markov chain. The name "stationary" makes intuitive sense: equation (10) tells us that if we start the chain with distribution $\boldsymbol{\pi}_0 = \boldsymbol{\pi}$, all further iterations of the process will leave the distribution unchanged. From our discussion on eigenvectors, we see immediately that $\boldsymbol{\pi}$ must be a left-eigenvector of $\mathbf{P}$ corresponding to an eigenvalue of $\lambda = 1$. So a vector that satisfies the balance equation always exists (although it might be complex or not non-negative and, hence, not a probability vector). The important question is whether

1. $\boldsymbol{\pi}$ defines a probability distribution: are all elements are non-negative real numbers that sum to one?
2. whether $\boldsymbol{\pi}$ is *unique*: are there multiple $\boldsymbol{\pi}$ that satisfy $\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \mathbf{P}$? Will the vector $\boldsymbol{\pi}$ depend on the initial distribution?
3. whether $\boldsymbol{\pi}_t$ converges to a unique $\boldsymbol{\pi}$ as $t \to \infty$: if $\boldsymbol{\pi}$ is unique, do we have $\lim_{t\to\infty} \boldsymbol{\pi}_t = \boldsymbol{\pi}$ regardless of where we start the chain?

In 3., we considered the notion of a vector to which $\boldsymbol{\pi}_t$ converges as $t \to \infty$. This vector, if it exists, is called the *limiting distribution* of the Markov chain can be expressed as:

$$\boldsymbol{\pi}_\infty^\top = \boldsymbol{\pi}_0^\top \mathbf{P}^\infty. \tag{11}$$

At this point, you might wonder why the distinction is made between stationary and limiting distributions. After all, if $\boldsymbol{\pi}$ doesn't change after reaching the steady state, it must be the case that $\boldsymbol{\pi} = \boldsymbol{\pi}_\infty$. The problem is that $\boldsymbol{\pi}_t$ might never reach the steady state (even if the distribution $\boldsymbol{\pi}$ is unique). Here is a very simple counter example: consider a Markov chain on the state space $\{1, 2\}$ with transition matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{12}$$

Hence, whenever $Y_t = 1$, then $Y_{t+1} = 2$ with probability 1, and when $Y_t = 2$, then $Y_{t+1} = 1$ with probability 1. It is not difficult to see that $\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \mathbf{P}$ for the vector $\boldsymbol{\pi} = [0.5, 0.5]^\top$. So $\boldsymbol{\pi}$ is a stationary distribution. However, $\lim_{t\to\infty} \mathbf{P}^t$ will not converge, but oscillate between

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \text{ and } \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and $\boldsymbol{\pi}$ will oscilate between $[0, 1]^\top$ and $[1, 0]^\top$. This implies that $\lim_{t\to\infty} \boldsymbol{\pi}_t$ doesn't converge. So, the limiting distribution does not exist in this case.[7]

When is the Stationary Distribution the Limiting Distribution?

It turns out that all conditions 1. to 3. hold if the transition matrix $\mathbf{P}$ has a specific structure. Important to this discussion is the concept of a primitive matrix.[8]

Definition (Primitive Matrix)**.** *Let* $\mathbf{A}$ *be a non-negative square matrix. That is,* $\mathbf{A} = (a_{ij})$ *with* $a_{ij} \geq 0$ *for* i, j*. Let* $a_{ij}^{(k)}$ *be the* (i, j)*th element of the matrix* $\mathbf{A}^k$*.* $\mathbf{A}$ *is called* primitive *if there exists a positive integer* k *such that* $a_{ij}^{(k)} > 0$ *for all* i, j*.*

For a transition matrix $\mathbf{P}$ to be primitive, there has to be some time point t at which all elements of $\mathbf{P}^t = (p_{ij}^{(t)})$ are strictly positive. Intuitively speaking, this means that any node v_i must have a non-zero probability to throw the frisbee to any other node v_j at the tth iteration of the process. The reason why primitive matrices are of importance is two fold: first, if there is a time point t for which $p_{ij}^{(t)} > 0$ holds for all $v_i, v_j \in V$, then the transition matrix $\mathbf{P}$ will stay positive (i.e., all elements of $\mathbf{P}$ are strictly positive) for all subsequent periods.[9] So, primitive matrices are non-zero

[7] Notice, however, that the proportion of visits to each state over time, $\mu_i = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}(Y_t = v_i)$, where $\mathbb{I}(A)$ is an indicator function that is 1 if A is true an 0 otherwise, will converge to $\boldsymbol{\pi} = [1/2, 1/2]^\top$. This vector of proportions is sometimes called the *asymptotic frequency distribution* (which is indeed a "probability" if you are a Frequentist). We will, however, not dwell on this concept in this note.

[8] In the context of Markov chains, primitive matrices are sometimes called *regular matrices*.

[9] Note that the (i, j)th element of the matrix $\mathbf{P}^{t+1} = \mathbf{P}\mathbf{P}^t$ is $\sum_{k=1}^{n} p_{ik} p_{kj}^{(t)}$. All of the $p_{kj}^{(t)}$ terms are strictly

square matrices that turn to *positive matrices* when raised to a sufficiently large power. The second reason is that primitive matrices are the object of the following fundamental theorem, which we will not be able to prove here:[10]

Theorem 2 (Perron-Frobenius, Primitive Matrix, Abridged)**.** *Let* $\mathbf{A}$ *a primitive matrix. Then,*

1. *There exists a real number* $\rho > 0$*, called the* Perron root*, that is an eigenvalue of* $\mathbf{A}$ *and is strictly larger, in absolute value, than any other eigenvalue of* $\mathbf{A}$*. That is,* $\rho > |\lambda_i|$ *for all* i*.*
2. *The eigenvalue* ρ *is simple, i.e., there are no other eigenvalues of* $\mathbf{A}$ *equal to* ρ*.*
3. *All elements of the eigenvector corresponding to* ρ*, called the* Perron vector*, are strictly positive.*

The rest of this note will consider conditions that a graph $G(V, E)$, such as the one in Figure 1, has to satisfy in order for the associated transition matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$ to be primitive. Further, we'll show that the Perron root is equal to 1 for transition matrices. Together this will establish that $\mathbf{P}^t$ converges to a limit $\mathbf{P}^\infty$, which implies that $\boldsymbol{\pi}_t$ will converge to $\boldsymbol{\pi}_\infty$, which is unique and stationary.

We start by showing that the transition matrix of a Markov chain on a finite state space is primitive if it is *irreducible* and *aperiodic*.

Irreducibility

Definition (Irreducibility)**.** *Let* $\mathbf{P}$ *be the transition matrix of a Markov chain on a finite state space* V*.* $\mathbf{P}$ *is* irreducible *if for any two states* $v_i, v_j \in V$*, there exists an integer* $k = k(v_i, v_j)$ *such that* $p_{ij}^{(k)} > 0$*.*

Notice we have written $k = k(v_i, v_j)$ to emphasize that this integer depends on the specific states v_i and v_j—i.e., $k(v_i, v_j)$ and $k(v_k, v_l)$ are not necessarily equal. So, a Markov chain is irreducible if you can reach any state from any other state in finite steps with positive probability. This condition is weaker than primitivity, since for primitivity all states have to be reachable from all others after t iterations, while for irreducibility, the time t to reach another state can depend on the pair (v_i, v_j) under consideration.[11]

The Markov chain of the frisbee-throwing example is indeed irreducible, since each pair of nodes can be reached through a path in Figure 2 with positive probability. For example, v_1 can reach v_3 through the path (or "steps") $\{v_1, v_2, v_3\}$ which has probability $P(\{v_1, v_2, v_3\}) = 1/2 \times 1/3 > 0$, among several other paths. An example of a Markov chain that is not irreducible is shown in Figure 3.

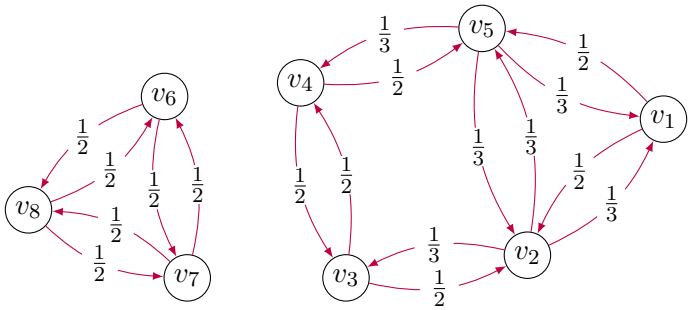
---

positive while $p_{ik}, k = 1, 2, ..., n$ are all non-negative with at least one element strictly greater than zero. Hence, $\sum_{k=1}^{n} p_{ik} p_{kj}^{(t)} > 0$ for all i, j implying that all elements of $\mathbf{P}^{t+1}$ must be positive.

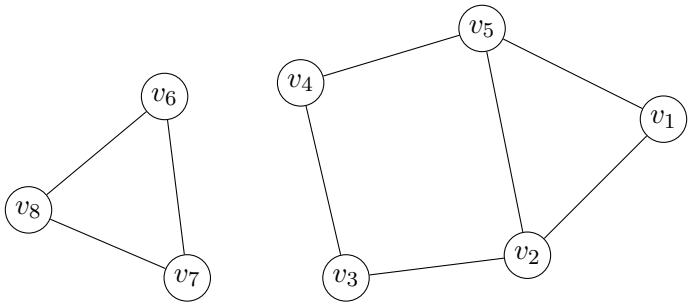
[10] Interested students might consult the last chapter of Mayer's textbook *Matrix Analysis and Applied Linear Algebra.*

[11] This becomes clear when you compare the definitions. Primitivity states: "there exists a k such that for all v_i, v_j, $p_{ij}^{(k)} > 0$." Irreducibility states: "for all v_i, v_j, there exists a $k = k(v_i, v_j)$ such that $p_{ij}^{(k)} > 0$." Hence, primitivity implies irreducibility but the converse doesn't hold.

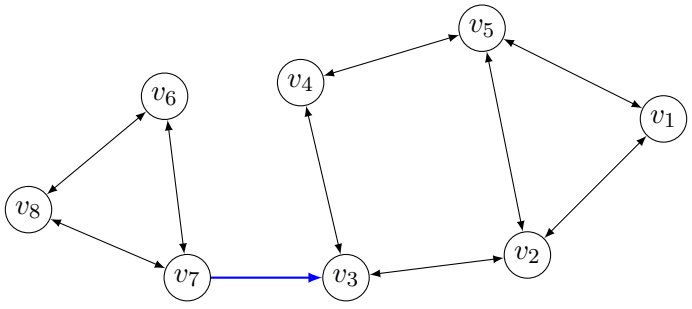
Figure 3: Markov chain that is not irreducible



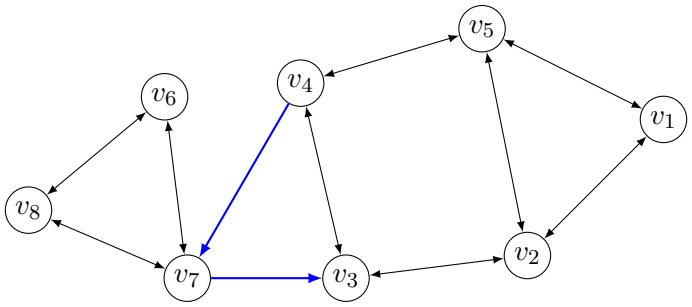
Here we see that node v_1 is not accessible from the either v_6, v_7, or v_8 and vice versa. Hence, there exists no integer k such that $p_{16}^{(k)} > 0$. In general, since $p_{ij} > 0$ whenever $p_{ji} > 0$ for random walks on undirected graphs, the transition matrix $\mathbf{P}$ will be irreducible if there exists a path between all pairs of nodes $v_i, v_j \in V$—in short if G is connected. This is certainly not the case in the underlying graph of the Markov chain shown in Figure 3, which would look like the following:



It's important to differentiate between *weak and strong connectivity* when dealing with Markov chains on directed networks, where the chain (i.e., frisbee) moves only in the direction of the edges. For example, consider the following weakly connected graph, where we have highlighted in blue the only edge that is not reciprocated.



The Markov chain on this network will be able to move from any node in $\{v_6, v_7, v_8\}$ to any node in $\{v_1, ..., v_5\}$ but not the other way around. So, what is needed for irreducibility is that the graph is *strongly connected*, such as one in the figure below:



We'll discuss a little bit what scholars do with directed networks at the end of the note. But for now, we'll discuss only undirected networks.

Aperiodicity

Another concept that is important in the discussion of stationary distributions of Markov chains is *aperiodicity*. Consider the set $\mathcal{T}(v_i) = \mathcal{T}_i = \{t \geq 1 : p_{ii}^{(t)} > 0\}$, which consists of all time steps (which are integers) at which the chain returns to v_i after a starting at v_i. The *period* of the state v_i is defined as $\text{gdc}(\mathcal{T}_i)$, where $\gcd(A)$ denotes the greatest common divisor of the set A. For example, in the simple Markov chain below in Figure 4, we see that it will take exactly 4 time steps until the chain returns to its starting point, regardless of the node. So for all $v_i \in V$, $\mathcal{T}_i = \{4, 8, 12, 16, ...\}$ and the period of all v_i's is $\gcd(\mathcal{T}_i) = 4$.

Figure 4: Markov chain where each state has period 4

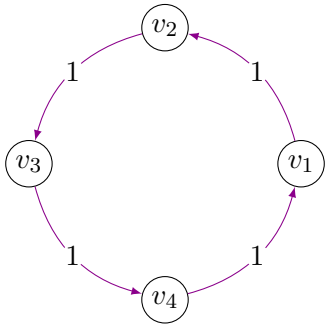
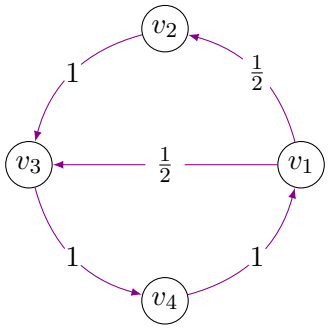


Figure 5: Markov chain where each state has period 1



For another example, consider the Markov chain in Figure 5, where we have added one additional transition path from v_1 to v_3 to Figure 4. Now, there's a cycle of length 3 from v_i back to v_i—$\{v_1, v_3, v_5, v_1\}$—and a cycle of length 4—$\{v_1, v_2, v_3, v_4, v_1\}$—and all other integers in the set $\mathcal{T}_i$ will be of the form $4a + 3b$ where a, b are non-negative integers. Hence, we see that the period of v_1 is $\gcd(T_1) = 1$. For v_2, we see that there's a cycle of length 4 (the old path from Figure 4) and, in addition, a walk of length 7: $\{v_2, v_3, v_4, v_1, v_3, v_4, v_1, v_2\}$. Hence, again, we have a period of $\gcd(\mathcal{T}_2) = 1$. You can check that the period of v_3 and v_4 is 1 as well (you should verify this).

In fact, that all states have the same period in both examples is not a coincidence. And it will hold for all irreducible Markov chains.

Lemma 2. *If* $\mathbf{P}$ *is irreducible, then all states have the same period. That is,* $gdc(\mathcal{T}_i) = gdc(\mathcal{T}_j)$ *for all* $v_i, v_j \in V$*.*

So, for irreducible Markov chains, we can talk about the *period of the Markov chain* itself, instead of individual states. For example, the Markov chain represented in Figure 4 has a period of 4, while that in Figure 5 has a period of 1. Markov chains with a period of 1 are called *aperiodic*.

For the purpose of this note, the following corollary will be useful.

Corollary 1. *Let* $G(V, E)$ *be a connected (undirected) graph with adacency matrix* $\mathbf{A}$*. Then the transition matrix* $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$ *is aperiodic if* G *contains at least one cycle of odd length.*

Now, that we have the concepts of aperidoicity and irreducibility, we introduce a lemma that connects transition matrices of Markov chains to the Perron-Frobenious Theorem.

Lemma 3. *Let* $\mathbf{P}$ *be the transition matrix of a Markov chain on a finite state space. If* $\mathbf{P}$ *is irreducible and aperiodic, then* $\mathbf{P}$ *is primitive.*

With Lemma 3, we can apply the Perron-Frobenius Theorem (Theorem 2) to Markov chains on networks, given that the graph is undirected, connected, and contains a cycle of odd length (which can be also a self-loop, which has length 1). There remains, however, one last piece that we need to rule out. We know that $\mathbf{P}$ has an eigenvalue of 1 from the balance equation in (10). But we have not shown that this is the largest eigenvalue of $\mathbf{P}$, which we need to prevent the Markov chain from "blowing up." Further, if we can show that 1 is the largest eigenvalue of $\mathbf{P}$, then it would mean that the stationary distribution is *unique*, since by the Perron-Frobenius Theorem there are no two eigenvalues equal to 1.[12] Fortunately, it turns out that $\lambda = 1$ is indeed the largest eigenvalue of *any* transition matrix of Markov chains on finite state spaces (not necessarily aperiodic and irreducible).

Lemma 4. *Let* $\mathbf{P}$ *a transition matrix of a Markov chain with finite state space. Then, for all eigenvalues* λ *of* $\mathbf{P}$*, we have* $|\lambda| \leq 1$*.*

[12] This follows from the fact that the geometric multiplicity (which can be understood as the number of independent eigenvectors associated with an eigenvalue) cannot exceed the algebraic multiplicity of eigenvalues (the number of eigenvalues that have the same value). But, again, we won't go too deep here.

Putting it All Together

Finally, we have arrived at the following theorem, the proof of which is now relatively easy if we assume that $\mathbf{P}$ is diagonalizable (recall that the matrix of eigenvectors of $\mathbf{P}$ might not be invertible). Even if $\mathbf{P}$ is not diagonalizable, the limiting distribution is unique and equal to the stationary distribution; but the proof is a bit more complex.

Theorem 3. *Let $G(V, E)$ be a undirected graph that is connected and contains a cycle of odd length. Then, a unique limiting distribution of the Markov chain on G exists and is equal to the stationary distribution.*

By Lemma 3, Lemma 4, and the Perron-Frobenius Theorem, the transition matrix $\mathbf{P}$ of the Markov chain has a unique largest eigenvalue $\rho = 1$ and corresponding (right-)eigenvector $\mathbf{x}$ with strictly positive entries. As each row of $\mathbf{P}$ sums to one, $\mathbf{P1} = \mathbf{1}$. Hence, the eigenvector corresponding to ρ is simply $\mathbf{1}$. Further, all other eigenvalues of $\mathbf{P}$ are strictly less than 1 in magnitude. So,

$$\lim_{t\to\infty} \mathbf{P}^t = \mathbf{X}\left(\lim_{t\to\infty} \mathbf{\Lambda}^t\right)\mathbf{X}^{-1} = \mathbf{X}\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}\mathbf{X}^{-1} = \mathbf{x}_1\mathbf{r}_1^\top = \mathbf{1}\mathbf{r}_1^\top$$

where $\mathbf{x}_1$ is the first column of $\mathbf{X}$ and $\mathbf{r}_1^\top$ is the first row of $\mathbf{X}^{-1}$. This shows that $\mathbf{P}^\infty$ will be a matrix with each row equal to $\mathbf{r}_1^\top$. Further, we have

$$\lim_{t\to\infty} \boldsymbol{\pi}_0^\top \mathbf{P}^t = \boldsymbol{\pi}_0^\top \mathbf{P}^\infty = \boldsymbol{\pi}_0^\top \mathbf{1}\mathbf{r}_1^\top = \mathbf{r}_1^\top, \tag{13}$$

which doesn't depend on the initial state $\boldsymbol{\pi}_0$. Hence, *regardless of where we start the chain*, we'll arrive at $\mathbf{r}_1^\top$ as $t \to \infty$, which shows that the limiting distribution is unique.

Further, multiplying both sides of equation 13 from the right by $\mathbf{P}$, we obtain

$$\left(\lim_{t\to\infty} \boldsymbol{\pi}_0^\top \mathbf{P}^t\right)\mathbf{P} = \left(\lim_{t\to\infty} \boldsymbol{\pi}_0^\top \mathbf{P}^{t+1}\right) = \lim_{t\to\infty} \boldsymbol{\pi}_1^\top \mathbf{P}^t = \mathbf{r}_1^\top \mathbf{P},$$

where $\boldsymbol{\pi}_1^\top = \boldsymbol{\pi}_0^\top \mathbf{P}$. But $\lim_{t\to\infty} \boldsymbol{\pi}_1^\top \mathbf{P}^t = \mathbf{r}_1^\top$ implying that $\mathbf{r}_1^\top \mathbf{P} = \mathbf{r}_1^\top$. Hence, the limiting distribution must be the stationary distribution. Further, since the left-eigenvector corresponding to $\rho = 1$ is unique, $\mathbf{r}_1$ must be the left-eigenvector of $\mathbf{P}$ corresponding to ρ.

We can also check that $\mathbf{r}_1$ is indeed a probability vector: first, all entries have to be non-negative as $\lim_{t\to\infty} \boldsymbol{\pi}_0 \mathbf{P}^t$ is the sum of non-negative real numbers. Second, as $\mathbf{X}^{-1}\mathbf{X} = \mathbf{I}$, $\mathbf{r}_1^\top \mathbf{x}_1 = 1$ (since this is the $(1, 1)$th entry of the identity matrix). But $\mathbf{x}_1 = \mathbf{1}$, so $\mathbf{r}_1^\top \mathbf{1} = \sum_{i=1}^n r_i = 1$. So, $\mathbf{r}_1^\top$ is a probability distribution.

Some code

We can check this also on our computer. Feel free to skip this part and jump to the graphs that follow, as coding is not the main point of this note. I've just left the code here since I thought some of you might be interested (the code uses the `R` language; for students who come from a `C/C++` or `Python` background, notice that `R` indexing starts from `1` not `0`).

First, we create the adjacency matrix $\mathbf{A}$.

R

```
# creating the adjacency matrix of the frisbee-throwing example
A = matrix(
  c(0, 1, 0, 0, 1,
    1, 0, 1, 0, 1,
    0, 1, 0, 1, 0,
    0, 0, 1, 0, 1,
    1, 1, 0, 1, 0),
  nrow = 5,
  byrow = TRUE
)
```

We can also have a look into the adjacency matrix:

R

```
print(A)
```

```
#R>      [,1] [,2] [,3] [,4] [,5]
#R> [1,]    0    1    0    0    1
#R> [2,]    1    0    1    0    1
#R> [3,]    0    1    0    1    0
#R> [4,]    0    0    1    0    1
#R> [5,]    1    1    0    1    0
```

Next, we create the transition matrix $\mathbf{P}$ and the initial vector $\boldsymbol{\pi}_0$:

R

```
# degree vector
# notes: rowSums sums all elements of A along the rows
d = rowSums(A)

# diagonal matrix with inverse degrees
# notes : 1) / is element-wise division (1 is boradcasted)
#         2) diag creates a diagonal matrix out of a vector
D_inv = diag(1/d)
```

```r
# transition matrix
# notes: %*% stands for matrix multiplication
P = D_inv %*% A

# initial distribution
# notes: as the first element of the 1 X n vector is 1, it means that v_1
#        starts throwing the frisbee
p0 = matrix(c(1, 0, 0, 0, 0), nrow = 1)
```

Next, we define the function `MCforward` that runs the Markov chain with transition matrix $\mathbf{P}$ and initial distribution $\boldsymbol{\pi}_0$ for t iterations (the `#'` lines are just comments describing the parameters and return values of the function).

R

```r
#' Iterate Markov chain for t time periods
#'
#' @param p0 initial distribution
#' @param P  transition matrix
#' @param t  number of iterations
#' @param final_P if TRUE, return also P raised to the tth power
#' @return returns a matrix where each row is the probability distribution at time t
MCforward = function(p0, P, t, return_P = c(TRUE, FALSE)) {

  # get dimnsions of P
  dP = dim(P)

  # check arguments
  stopifnot(
      "matrix" %in% class(P), # check if P is a matrix
      dP[1] == dP[2],         # check if P is square (dim[1]: # rows, dim[2]: # of cols)
      t %% 1 == 0,            # check if t is integer (%% is the modulo operator in R)
      dim(p0)[2] == dP[2]     # p0 and P have the same # of columns
  )

  # create empty matrix to store results (NA stands for "missing")
  probs = matrix(NA, nrow = t + 1, ncol = dP[2])

  # add initial vector into first row
  probs[1, ] = p0

  # if we want to calculate the final P, we need to create a new object
  P_final = P

  # iterate Markov chain forwards
  for (i in 2:(t + 1)) {
```

```
        # update probability vector
        probs[i, ] = probs[i - 1, ] %*% P

        if (return_P)
            P_final = P %*% P_final  # update P_final

    }

    # return results
    # note: if a function encounters the "return" function, it will exit immediately
    #       and return the argument. Hence, if return_P == TRUE the next return
    #       function---i.e., return(probs)---will never be executed
    if (return_P)
        return(list(probs = probs, P = P_final))

    return(probs)

}
```

With this function, we can calculate the history of the Markov chain, using the intial state `p0` and the transition matrix `P` we've created above. Let's run `t = 50` iterations, which should be enough to see convergence in both $\boldsymbol{\pi}_\infty$ and $\mathbf{P}^\infty$:

R

```
# run MC chain for 30 iterations
res = MCforward(p0 = p0, P = P, t = 50, return_P = TRUE)

# get eigendecomposition of P
# note: this returns a list object with two elements, "values" and "vectors"
eigenP = eigen(P)

# get limiting distribution (first row of X^{-1})
# note: 1) the "solve" function solves linear equations; when provided with only one
#          argument, it will invert the matrix
#       2) by default, eigenvectors are normalized to have length equal to 1. So,
#          we need to renormalize plim_raw to a probability vector
plim_raw = solve(eigenP$vectors)[1, ]
plim = plim_raw / sum(plim_raw)
```

We can also check that the `plim` object is equal to the first left-eigenvector of $\mathbf{P}$ and that $\mathbf{P}^\infty$ is a matrix in which each row contains the same probability vector, which is, again, equal to the left-eigenvector.

R

```r
# first left-eigenvector
plim2 = eigen(t(P))$vector[, 1]
# rescale
plim2 = plim2 / sum(plim2)

# check for approximate equality of first row of X^{-1}, first left-eigenvector
# and results of the function forwardMC
all.equal(plim, plim2, res$probs)
```

```
#R> [1] TRUE
```

```r
# look at stationary dist.
print(plim)
```

```
#R> [1] 0.1666667 0.2500000 0.1666667 0.1666667 0.2500000
```

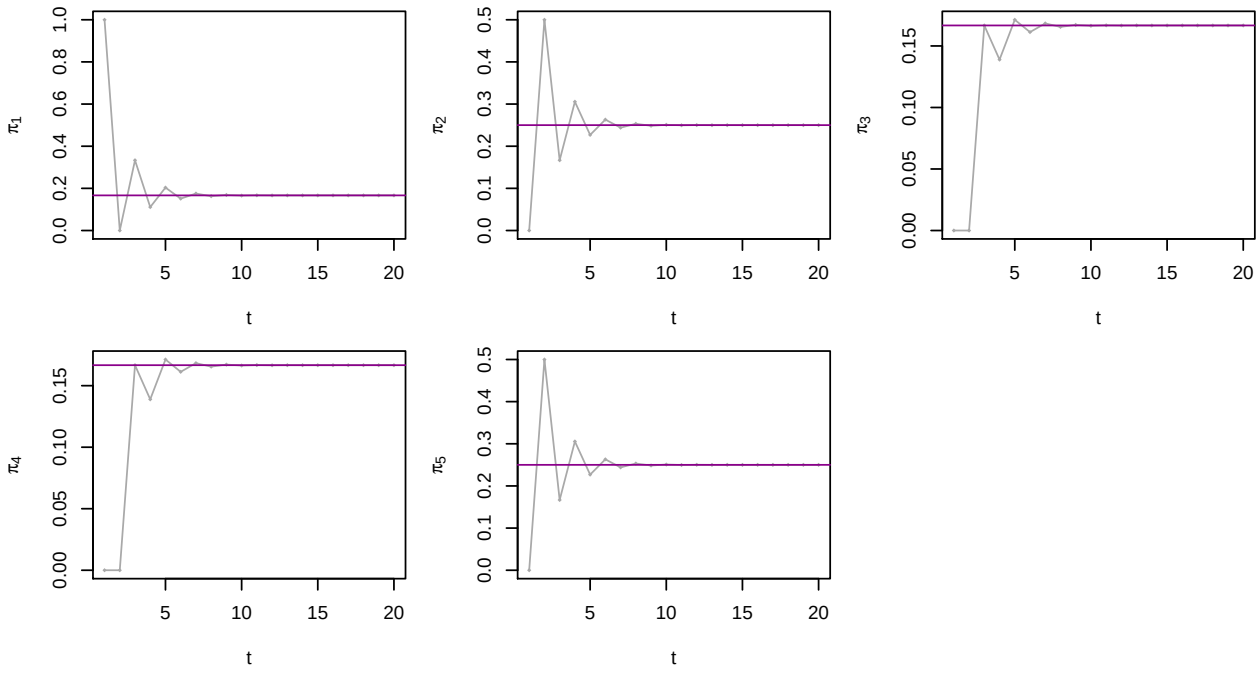
```r
# look at P^{infinity}
print(res$P)
```

```
#R>           [,1]      [,2]      [,3]      [,4]      [,5]
#R> [1,] 0.1666667 0.2500000 0.1666667 0.1666667 0.2500000
#R> [2,] 0.1666667 0.2499817 0.1666850 0.1666484 0.2500183
#R> [3,] 0.1666667 0.2500275 0.1666392 0.1666941 0.2499725
#R> [4,] 0.1666667 0.2499725 0.1666941 0.1666392 0.2500275
#R> [5,] 0.1666667 0.2500183 0.1666484 0.1666850 0.2499817
```

---

where lines following the `#R>` symbol, shows the output of the `R` console.

When we plot the history of the Markov chain, we see that the chain converges quite fast (the horizontal purple line represent the elements in $\mathbf{r}_i^\top$, i.e., left-eigenvector corresponding to $\rho = 1$):



So, we see that $\boldsymbol{\pi}_t$ indeed converges to the first left-eigenvector. How fast the chain will converge to the limiting distribution will depend on the magnitudes of the remaining eigenvalues of $\mathbf{P}$, especially the second largest in magnitude. We know that $|\lambda_i| \leq 1$ for all of them. But the closer they are, in magnitude, to 1, the slower they will "vanish" over the iterations and, hence, the slower the convergence to the stationary distribution.

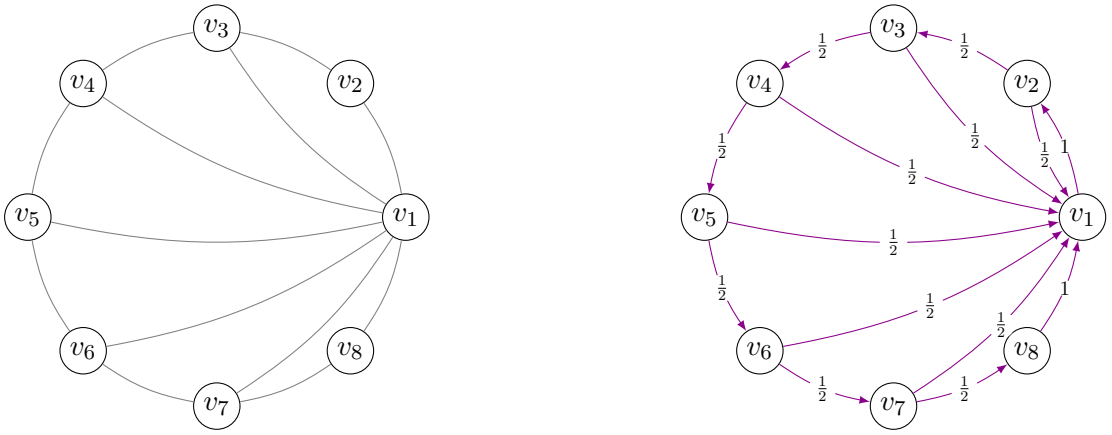
A Somewhat Disappointing Result and PageRank

Now, here is a somewhat disappointing result that makes us wonder why we have spent all this time on Markov chains on graphs.

> **Disappointing Result** *Let* $\boldsymbol{\pi} = [\pi_1, ..., \pi_n]^\top$ *be the limiting distribution of a Markov chain on a connected, undirected graph* $G(V, E)$ *that contains at least one odd-length cycle. Then,* $\pi_i = \frac{d(v_i)}{2|E|}$*, where* $d(v_i)$ *is the degree of node* v_i *and* $|E|$ *the number of edges in* G*.*

This result makes sense: the more ties a node has, the more often a random walk would visit the node. On the other hand, it implies that the limiting distribution of the Markov chain doesn't give us any information about the connectedness of the nodes beyond their degree. Hence, using a random walk measure of centrality on undirected graphs doesn't make sense.[13] On the other hand, such a measure *does* make sense on directed networks. In fact, the same network can have quite different limiting distributions depending on the directions of ties. To give you an example, consider the following two graphs:

Figure 6: Same network structure with different edge characteristics



(1) Undirected Graph
(Trans. probs. are inversely proportional to $d(v_i)$)

(2) Directed Graph
(Numbers represent trans. probs.)

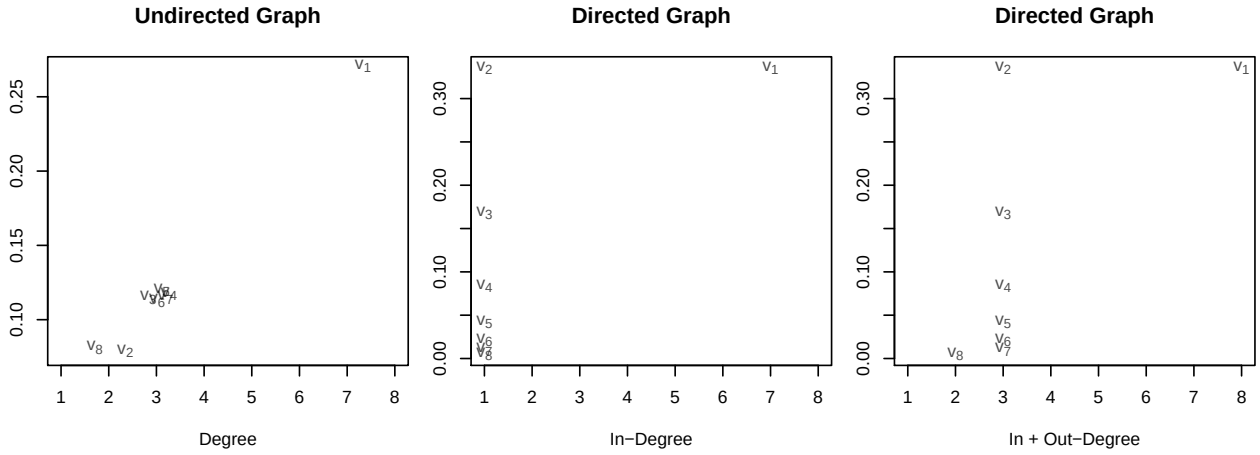
For the undirected graph, we know that the Markov chain will be irreducible and aperiodic (it's connected and has a odd-length cycle). Both of these properties hold for the directed graph as

[13] Whether the edeges are weighted doesn't change this result as shown in the appendix.

well: it is strongly connected, which implies irreducibility, and it contains a at least two cycles with lengths that are relatively prime—e.g., $\{v_1, v_2, v_1\}$ is a cycle of length 2 and $\{v_1, v_2, v_3, v_1\}$ has length 3—implying aperiodicity.

In the undirected graph, see that the vector $\boldsymbol{\pi}_\infty$ will have only three unique values, one probability for the nodes $\{v_3, ..., v_7\}$, one for $\{v_2, v_8\}$, and one for $\{v_1\}$. In fact, we can calculate these probabilities by hand as 3/26, 2/26, and 7/26, respectively. For the directed graph, on the other hand, the probabilities will vary much more across the nodes. Starting at v_1, the chain will move counter-clockwise along the circle, where at each time step, it has a probability of 1/2 to return to v_1. Hence, the probability of reaching the next node of on the circle between v_3 and v_8 will decrease exponentially with it's distance from v_1, even though $v_3, ..., v_7$ have the exact same in- and out-degree. On the other hand, v_2 will have exactly the same probability as v_1, as the chain must visit v_2 whenever it visits v_1. Hence, in the directed graph, the limiting vector $\boldsymbol{\pi}_\infty$ will provide us much richer information about the structure of the graph than the node degrees.

We can use the `R` code from above to examine the relation between the node degrees and their stationary probabilities. (To make the first plot more readable, the horizontal and vertical positions are jittered by adding a little bit of noise. The nodes of each of the two clusters in the lower-left corner would have the exact same positions.)



The graph shows clearly that the in-degree of nodes in a directed graph is a poor predictor of the stationary probabilities, except for the node v_1. Combining in- and out-degree doesn't help either. Of course, the example that I've picked is quite extreme. But it shows that the vector $\boldsymbol{\pi}_\infty$ inform us how "central'' each node is in a *directed* network.

The most often utilized centrality measure based on these intuitions is called *PageRank*. The PageRank is just the vector $\boldsymbol{\pi}_\infty$ calculated on a directed network, i.e., the ith element of this vector will be the PageRank score of the node v_i in the network. As we've discussed above, this vector is not well-behaved if the Markov chain created by the random walk on the network is either reducible

or periodic. To ensure that irreducibility and aperiodicity hold, PageRank uses a common "trick" of turning the "random walker" on the network into a "random surfer."

Without the modification, the transition matrix on a directed network is

$$p_{ij} = a_{ij} d_{\text{out}}(v_i)^{-1} \tag{14}$$

where $d_{\text{out}}(v_i) = \sum_{j=1}^{n} a_{ij}$ denotes the out-degree of node v_i. Hence, the random walker (i.e., Markov chain) moves from the node it is currently located to the adjacent nodes, where the probability of moving to v_j is inversely-proportional to the out-degree of node v_i if $(v_i, v_j) \in E$ and zero otherwise.

PageRank adds two modification to this transition probability. First, to overcome the problem of "dangling nodes"—nodes with zero out-degree—it adds an out-going link from all dangling nodes to all other nodes. Otherwise, the chain will nowhere to go after reaching these dangling nodes (they are "absorbing states" in the laguage of Markov chains). Notice that adding these links will make the transition matrix irreducible. The second "trick," which makes the Markov chain aperiodic, is to add a small "teleportation probability" to the transition probabilities for nodes that have some outgoing ties. This turns the random walker into a "random surfer" which follows the links of the graph with probability $(1 - \tau)$ and "teleport" to any other node with probability τ. Adding these two modifications to (14), we get the following transition probabilities:

$$p_{ij} = (1 - \tau) a_{ij} d_{\text{out}}(v_i)^{-1} + \frac{(1 - \tau)\mathbb{I}\left(\sum_{j=1}^{n} a_{ij} = 0\right) + \tau}{n}, \tag{15}$$

where $\mathbb{I}(A)$ is an indicator function that is equal to 1 if A is true and zero otherwise. To make sense of this equation, it's useful to break it down into three scenarios:

- **Case 1**: if $(v_i, v_j) \in E$, then the random surfer will move to v_j with probability $(1 - \tau)$ and teleport to a random node in the network with probability τ; but since there are n nodes in the network, the probability that she will move to the node v_j "by teleportation" is τ/n. So, $p_{ij} = (1 - \tau) d_{\text{out}}(v_i)^{-1} + \tau n^{-1}$.
- **Case 2**: if $(v_i, v_j) \notin E$ but the vertex v_i has some outgoing edge(s)—i.e., $\sum_{j=1}^{n} a_{ij} > 0$—then the random surfer will go to a node v_k connected to v_i, $k \neq j$, with probability $(1 - \tau)$. Hence, the only way she reaches v_j is through teleportation, the probability of which is τ/n. Hence, $p_{ij} = \tau n^{-1}$ in this case.
- **Case 3**: if the node v_i has no outgoing ties—i.e., $\sum_{j=1}^{n} a_{ij} = 0$—then the random surfer will randomly jump to any other node in the network with probability 1. Hence, the probability of hitting v_j is $p_{ij} = 1/n$

Notice that each row of the transition matrix will sum to one:

$$\sum_{j=1}^{n} p_{ij} = \begin{cases} (1-\tau)\frac{\sum_{i=j}^{n} a_{ij}}{d_{\text{out}}(v_i)} + \sum_{j=1}^{n} \frac{\tau}{n} = 1, & \text{if } \sum_j a_{ij} > 0 \\ \sum_{j=1}^{n} \frac{1}{n} = 1, & \text{if } \sum_j a_{ij} = 0. \end{cases}$$

So $\mathbf{P} = (p_{ij})$ is a valid transition matrix which is irreducible and aperiodic, implying that the Markov chain has a unique limiting distribution. The PageRank is just this vector of limiting probabilities. For the graph on the right of Figure 6, for example, the PageRank vector is

R

```
# create igraph object out of directed graph in Figure 6
# note: pkg::fun means that we use the function "fun" from the package "pkg".
#       Of course, to use this function, you'd need to install the package "igraph,"
#       which can be done by running
#       install.packages("igraph", repos = "https://cran.rstudio.com", dependencies = T)
g = igraph::graph_from_adjacency_matrix(A_dir, mode = "directed")

# teleportation prob.
tau = .15

# calculate PageRank
# note: the algorithm in the igraph package uses the "damping factor" as an option
#       which is just 1 minus the teleportation probability
#       the "algo" option specifies which algorithm to use
pagerank = igraph::page_rank(g, damping = 1 - tau, algo = "prpack")$vector
print(pagerank)
```

```
#R> [1] 0.32158497 0.29209723 0.14289132 0.07947881 0.05252849 0.04107461 0.03620671
#R> [8] 0.03413785
```

Comparing that with hand-calculations shows that this is indeed the first right-eigenvector (i.e., the stationary distribution) of the transition matrix $\mathbf{P}$ defined in (15):

R

```
# trans. matrix without teleportation
P_raw = diag(1 / rowSums(A_dir)) %*% A_dir
# add modification for teleportation (no dangling nodes in graph)
P = (1 - tau) * P_raw +
    (tau / nrow(A_dir)) * matrix(1, nrow = nrow(A_dir), ncol = ncol(A_dir))

# calculate get left-eigenvector and normalize
pagerank2 = as.double(eigen(t(P))$vectors[, 1])
pagerank2 = pagerank2 / sum(pagerank2)
print(pagerank2)
```

```
#R> [1] 0.32158497 0.29209723 0.14289132 0.07947881 0.05252849 0.04107461 0.03620671
#R> [8] 0.03413785

# check for approx. equality
all.equal(pagerank, pagerank2)

#R> [1] TRUE
```

Hence, if you have a directed graph (it might be weighted) an desire a eigenvector-like centrality measure, the PageRank will do the job.

Appendix

- **Lemma 1**: The eigenvalues of a square matrix $\mathbf{A}$ are found as follows. From (1), we have

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad \Longrightarrow \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

  for a eigen-pair of $\mathbf{A}$ (the symbol $\Longrightarrow$ means "implies"). This shows that $(\mathbf{A} - \lambda\mathbf{I})$ cannot have an inverse, since if it were invertible we would conclude that $\mathbf{x} = (\mathbf{A} - \lambda\mathbf{I})^{-1}\mathbf{0} = \mathbf{0}$, contradicting the definition of eigenvectors as nonzero vectors. Hence, $(\mathbf{A} - \lambda\mathbf{I})$ must be singular implying that

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0.$$

  This defines an nth order polynomial in λ called the *characteristic equation*. The eigenvalues of $\mathbf{A}$ can be found by finding all values of λ that satisfy this equation. In general, the solutions are not necessarily distinct.

  Now, recall that for any square matrix $\mathbf{B}$, $\det(\mathbf{B}) = \det(\mathbf{B}^\top)$. So,

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \det\left([\mathbf{A} - \lambda\mathbf{I}]^\top\right) = \det(\mathbf{A}^\top - \lambda\mathbf{I}),$$

  since the identity matrix is symmetric. In other words, the characteristic equation for $\mathbf{A}$ and $\mathbf{A}^\top$ identical. And since the eigenvalues are completely determined by the characteristic equation, the eigenvalues of $\mathbf{A}$ and $\mathbf{A}^\top$ must be the same as well.

- **Lemma 2**: Fix two arbitrary states $v_i, v_j \in V$. Since $\mathbf{P}$ is irreducible, there exist natural numbers n, m such that $p_{ij}^{(n)} > 0$ and $p_{ji}^{(m)} > 0$. Let $k = n + m$. Then, $p_{ii}^{(k)} \geq p_{ij}^{(n)} p_{ji}^{(m)} > 0$.[14] which shows that k is a multiple of $\gcd(\mathcal{T}_i)$. Now, consider an arbitrary $l \in \mathcal{T}_j$ and notice that $p_{ii}^{(k+l)} > 0$, since you can go from v_i to v_j, move around in a length-l cycle until you come back to v_j, and then move back to v_i. Hence, $k + l$ is a multiple of $\gcd(\mathcal{T}_i)$ and so is l.[15] Thus,

[14] That is, we take the length-n walk from v_i to v_j, and then the length-m walk back to v_i, which results in a length-k cycle with positive probability. However, since there might be other length-k cycles starting at v_i with higher probability, $p_{ii}^{(k)} \geq p_{ij}^{(n)} p_{ji}^{(m)}$.

[15] Let $x = \gcd(\mathbf{T}_i)$. Then $k = ax$ and $k + l = bx$ as x divides all elements in $\mathcal{T}_i$. But this implies that

$\gcd(\mathcal{T}_i)$ divides all elements in $\mathcal{T}_j$, implying that $\text{gdc}(\mathcal{T}_i) \leq \text{gdc}(\mathcal{T}_j)$. Using the same argument but interchanging the role of v_i and v_j, we obtain $\text{gdc}(\mathcal{T}_i) \geq \text{gdc}(\mathcal{T}_j)$, which completes the proof.

- **Corollary 1**: Suppose that the graph G contains a cycle C of odd length. Pick any node in C and call it v_i, and let v_j be any of its neighbors. As the G is undirected, $p_{ij} > 0$ and $p_{ji} > 0$, so all even numbers are $\mathcal{T}(v_i)$. But v_i is contained in a cycle of odd length, which implies that there is at least one odd number in $\mathcal{T}(v_i)$. Hence, $\text{gdc}\mathcal{T}(v_i) = 1$.[16] Lastly, as the graph is connected, the Markov chain is irreducible, from which it follows that all nodes (i.e., states) have the same period by Lemma 2.

- **Lemma 3**: Unfortunately, there is no short proof for this lemma :(But it can be considerably shortened if we accept as a fact the following claim (for a proof, see Lemma 1.27 in Levin et al. (2008) *Markov Chains and Mixing Times*, RI: American Mathematical Society):

  > *Let $A \subset \mathbb{Z}_+$, where $\mathbb{Z}_+$ denotes that set of all non-negative integers. If A is closed under addition and $gcd(A) = 1$, then A contains all but finitely many elements of $\mathbb{Z}_+$. That is, $|B| < n$ for some integer $n < \infty$ where $B = \mathbb{Z}_+ \setminus A$.*

  What is of importance here is the "all but finitely many" in this statement. This implies that there B has a maximal integer (a maximum will not exist if B is infinite in size). Now, consider an arbitrary state v_i of the Markov chain. As the chain is aperiodic, $\gcd(\mathcal{T}_i) = 1$. To show that $\mathcal{T}_i$ is closed under addition, consider $n, m \in \mathcal{T}_i$ and note that $p_{ii}^{n+m} \geq p_{ii}^n p_{ii}^m > 0$. Hence $n + m \in \mathcal{T}_i$. By the preceding lemma there are only finitely many non-negative integers that are not in $\mathcal{T}_i$. Let $t(v_i)$ be the largest of them and note that for all $t \geq t(v_i)$ we have $t \in \mathcal{T}_i$. By irreducibility, for all $v_j \in V$, there exists an integer $r = r(v_i, v_j)$ such that $p_{ij}^r > 0$. Hence, for $t \geq t(v_i) + r$, we have $p_{ij}^{(t)} \geq p_{ii}^{(t-r)} p_{ij}^{(r)} > 0$ (i.e., cycle from v_i back to v_i, and then walk from v_i to v_j). Let $t^*(v_i) = t(v_i) + \max_{v_j \in V} r(v_i, v_j)$. Then, for all $t > t^*(v_i)$, $p_{ij}^{(t)} > 0$ for all $v_j \in V$. Lastly, for $t \geq \max_{v_i \in V} t^*(v_i)$, $p_{ij}^{(t)} > 0$ for all $v_i, v_j \in V$. We are done.

- **Lemma 4**: Let us denote the largest eigenvalue (in magnitude) of $\mathbf{P}$ by[17]

  $$\rho = \max\{|\lambda_1|, ..., |\lambda_n|\}.$$

  We want to show that $\rho \leq 1$. Let λ be an eigenvalue of $\mathbf{P}$ and $\mathbf{x}$ its corresponding eigenvector.[18]

$l = (k + l) - k = bx - ax = (b - a)x$ is a multiple of x as well.

[16] In fact, we don't even need *all* even numbers to be elements of $\mathcal{T}_i$. $2 \in \mathcal{T}_i$ will suffice for this argument, since the gdc between any odd number and 2 must be 1.

[17] This eigenvalue is called the *spectral radius* of $\mathbf{P}$.

[18] In general, λ might be a complex number and $\mathbf{x}$ a complex vector. But let us assume that both are real for simplicity. The exact same proof holds for complex eigen-pairs.

By definition of eigenvectors and eigenvalues, we have

$$\mathbf{Px} = \begin{bmatrix} \sum_{j=1}^n p_{1j}x_j \\ \vdots \\ \sum_{j=1}^n p_{ij}x_j \\ \vdots \\ \sum_{j=1}^n p_{nj}x_n \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \vdots \\ \lambda x_i \\ \vdots \\ \lambda x_n \end{bmatrix} = \lambda \mathbf{x}$$

Let x_k be the element of $\mathbf{x}$ with largest magnitude. Using the kth row of the equation, we find that

$$|\lambda x_k| = \left| \sum_{j=1}^n p_{kj}x_j \right| \leq \sum_{j=1}^n p_{jk}|x_j| \leq \sum_{j=1}^n p_{jk}|x_k|$$

since $|x_k| \geq |x_j|$ for all $j = 1, 2, ..., n$. Now, note that $|x_k|$ is a constant and $\sum_{j=1}^n p_{kj} = 1$ (we sum the kth row of the matrix $\mathbf{P}$). Hence, $\sum_{j=1}^n p_{jk}|x_k| = |x_k| \sum_{j=1}^n p_{jk} = |x_k|$ and

$$|\lambda x_k| \leq |x_k|.$$

The result follows from dividing both sides by $|x_k|$.

- **Disappointing Result**: The proof of this result is actually quite simple. If $\boldsymbol{\pi}$ is a stationary distribution, then $\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \mathbf{P}$. So, considering only the jth node, we have $\pi_j = \sum_{i=1}^n \pi_i p_{ij}$. Recall that $p_{ij} = a_{ij} d(v_i)^{-1}$. In other words,

  1. if $(v_i, v_j) \in E$—i.e., v_i and v_j have an edge in the graph G—the (i, j)th entry of the adjacency matrix is $a_{ij} = 1$ and probability of moving from i to j is $p_{ij} = a_{ij} d(v_i)^{-1} = d(v_i)^{-1}$.
  2. If $(v_i, v_j) \notin E$, then $a_{ij} = 0$ and so $p_{ij} = 0$

  which corresponds with the way we have constructed $\mathbf{P}$. Now notice that the following holds:

$$\sum_{i=1}^n d(v_i) p_{ij} = \sum_{i=1}^n d(v_i) \left(\frac{a_{ij}}{d(v_i)} \right) = \sum_{i:(v_i,v_j)\in E} \frac{d(v_i)}{d(v_i)} = \sum_{i:(v_i,v_j)\in E} 1 = d(v_j). \tag{16}$$

  Since $\sum_{i=1}^n d(v_i) = 2|E|$, dividing all the degrees by $2|E|$ turns the degree vector $\mathbf{d} = [d(v_1), ..., d(v_n)]^\top$ into a valid probability vector. Hence, dividing both sides of (16), we conclude that

$$\pi_j = \frac{d(v_j)}{2|E|}, \qquad j = 1, 2, ..., n$$

  must be the unique stationary distribution of the Markov Chain on G.

  The situation is no different on a *weighted graph* or *weighted network*.

  Definition (Weighted graph)**.** *Let $G(V, E)$ be a graph. A weighted graph, $G(V, E, w)$, is the*

graph G together with a weight function $w : V \times V \to [0, \infty)$ that is $w(v_i, v_j) = 0$ if $(v_i, v_j) \notin E$ and strictly positive otherwise.

So, a weighted graph is nothing but a usual graph with a function w that attaches to each edge a positive weight.[19] Let w_{ij} the weight of edge $(v_i, v_j) \in E$ and $w_i = \sum_{j=1}^{n} w_{ij}$. Notice that $w_{ij} = w_{ji}$ if $(v_i, v_j) \in E$ and 0 otherwise. The transition probabilities can be then defined as $p_{ij} = w_{ij}/w_i$, i.e., given that the random walker is at v_i it will move to its neighbors with probability proportional to the weights of the edges. But then

$$\sum_{i=1}^{n} w_i p_{ij} = \sum_{i=1}^{n} w_i \left(\frac{w_{ij}}{w_i} \right) = w_j$$

and normalizing both sides by $\sum_{i=1}^{n} w_i$, we conclude

$$\pi_j = \frac{w_j}{\sum_{i=1}^{n} w_i}, \qquad j = 1, 2, ..., n,$$

which shows that the limiting probabilities are proportional to the weight attached to each node's edge (sometimes called the *node strength.*)

[19] Of course, we can also define a weighted graph with weight functions that can be negative. But it is often the case that the weights are taken to be non-negative.