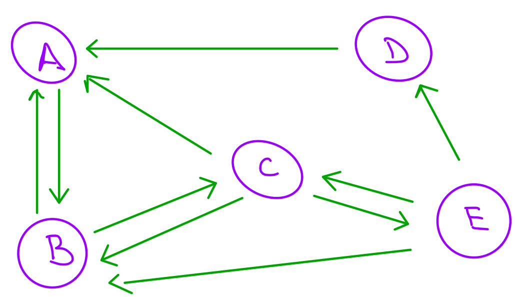
4.9 Markov Chains + Google's Page Rank Algorithm

Thanks to kleinproject.org

Goals:

1. Rough intro to Google's Page Rank Alg.
2. Intro to Markov chains

Let's explore the internet. There are pages and there are links from one page to another. Here is a simple example.



Q: How do we determine how "important" each page is so we can rank them.

Google's A: Take a random "walk", i.e. randomly click on links. Then...

more important page $\iff$ you are more likely to end up on the page after several clicks

The formal setup

1. Record the probability of transitioning from one page to another in a <u>transition matrix</u>.

$$T = \begin{array}{cc} & \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix} & \begin{bmatrix} 0 & 1/2 & 1/3 & 1 & 0 \\ 1 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 1/3 & 0 & 0 \end{bmatrix} \end{array}$$

prob. of moving from C to B

$*$ j^{th} column contains the probabilities of moving from page j to page i.

2. Record the probabilities that we are on each page after 0 clicks, 1 click, ... in a sequence of <u>state vectors</u> $\overline{x}_0, \overline{x}_1, \overline{x}_2, \ldots$.

- Suppose we start on page C. Then

$$\overline{x}_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix}$$

- Looking at the graph and where we can go from page C, we see

$$\overline{x}_1 = \begin{bmatrix} 1/3 \\ 1/3 \\ 0 \\ 0 \\ 1/3 \end{bmatrix}$$

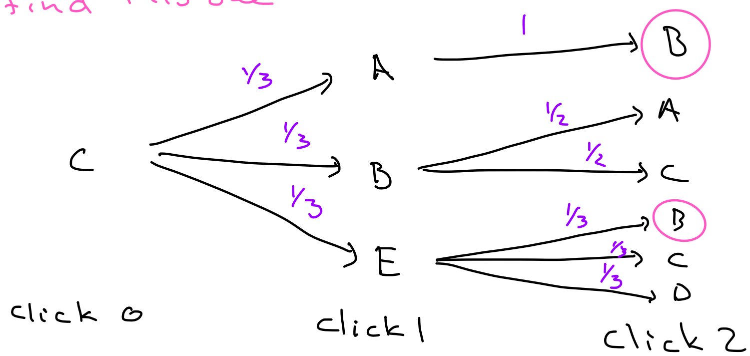
- what is $\overline{x}_2$? $\quad \overline{x}_2 = \begin{bmatrix} * \\ * \\ * \\ * \\ * \end{bmatrix}$

← prob of being at A after 2 clicks

← " " " " B " 2 "

Let's find this one

$\frac{1}{3}\cdot 1 + \frac{1}{3}\cdot\frac{1}{3} = \frac{4}{9}$



... continue ... until you notice the following

(3) $\overline{x}_k$ can be computed using T

$$\overline{x}_1 = T\cdot\overline{x}_0$$

$$\overline{x}_2 = T\cdot\overline{x}_1 = T^2\overline{x}_0$$

$$\overline{x}_3 = T\cdot\overline{x}_2 = T^3 x_0$$

$$\vdots$$

$$\boxed{\overline{x}_k = T\,\overline{x}_{k-1} \quad \text{so} \quad \overline{x}_k = T^k\,\overline{x}_0}$$

Pause for definitions

Def A vector with nonnegative entries that add up to 1 is called a probability vector. A square matrix whose columns are probability vectors, is called a stochastic matrix

* each state vector $\bar{x}_0, \bar{x}_1, \dots$ is a probability vector.
* T is a stochastic matrix.

Def A Markov chain is a sequence of prob. vectors together with a stochastic matrix T s.t. $\bar{x}_k = T\bar{x}_{k-1}$ for all $k \geq 1$.

* just like in our example.

"Solving" the Page Rank Problem

We want to see what happens to our Markov chain after lots of clicks, i.e. as $k \to \infty$.

- Remember, $\bar{x}_k = T^k \bar{x}_0$, so we need to compute T^k for large k — yikes!

1. Diagonalize T (if possible)

   (a) Eigenvalues (using wolframalpha)

$$p(\lambda) = -\lambda^5 + \frac{7}{9}\lambda^3 + \frac{2}{9}\lambda^2 - \frac{1}{18}\lambda + \frac{1}{18}$$

There are 5 different eigenvalues, so

T <u>is</u> diagonalizable. Let's call them

$\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$. Then

complex eigenvalues

$$\lambda_1 = 1$$

$$|\lambda_2| = |\lambda_3| \approx 0.7$$

$$|\lambda_4| = |\lambda_5| \approx 0.3$$

(b) Let $\bar{v}_i$ be an eigenvector assoc. to λ_i.

(we can compute these if needed)

(e.g. $\bar{v}_1 = \begin{bmatrix} 4 \\ 5.\bar{3} \\ 3 \\ 0.\bar{3} \\ 1 \end{bmatrix}$)

Then $T = PDP^{-1}$, where

$$D = \begin{bmatrix} 1 & & & & \\ & \lambda_2 & & 0 & \\ & & \lambda_3 & & \\ & 0 & & \lambda_4 & \\ & & & & \lambda_5 \end{bmatrix}, \quad P = \begin{bmatrix} \bar{v}_1 & \bar{v}_2 & \bar{v}_3 & \bar{v}_4 & \bar{v}_5 \end{bmatrix}$$

so

$$T^k = (PDP^{-1})^k = PD^kP^{-1}.$$

Now, we want to study T^k as $k \to \infty$.

$$T^{\infty} = \lim_{k\to\infty} T^k = \lim_{k\to\infty} P D^k P^{-1}$$

$$= \lim_{k\to\infty} P \begin{bmatrix} 1^k & & & & 0 \\ & \lambda_2^k & & & \\ & & \lambda_3^k & & \\ & & & \lambda_4^k & \\ 0 & & & & \lambda_5^k \end{bmatrix} P^{-1}$$

* Recall: $|\lambda_i| < 1$ for $i = 2,3,4,5$

$$= P \begin{bmatrix} 1 & & & 0 \\ & 0 & & \\ & & \ddots & \\ 0 & & & 0 \end{bmatrix} P^{-1}$$

* compute P, P^{-1}

$$\approx \begin{bmatrix} 0.293 & 0.293 & & 0.293 \\ 0.390 & 0.390 & & 0.390 \\ 0.220 & 0.220 & \cdots & 0.220 \\ 0.024 & 0.024 & & 0.024 \\ 0.073 & 0.073 & & 0.073 \end{bmatrix}$$

Thus, after lots of clicks we arrive at

$$\bar{X}_{\infty} = T^{\infty} \cdot \bar{X}_0 = T^{\infty} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \underset{\text{of } T^{\infty}}{3^{rd} \text{ col}} = \begin{bmatrix} 0.293 \\ 0.390 \\ 0.220 \\ 0.024 \\ 0.073 \end{bmatrix}$$

Recall, $\bar{X}_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ meant we started on page C, but we see that we would have gotten the same answer no matter what page we started on.

The vector $\bar{x}_\infty$ gives the prob. of ending up on each page after lots of clicks, no matter where we started. This gives the page ranking B, A, C, E, D

Another view of what happened

Recall: the eigenvectors $\bar{v}_1, \dots, \bar{v}_5$ are L.I.

so they form a basis for $\mathbb{R}^5$ since they must also span. Then

$$\bar{x}_0 = c_1 \bar{v}_1 + \dots + c_5 \bar{v}_5$$ for some

scalars $c_1, \dots, c_5$. Then,

$$T^k \bar{x}_0 = c_1 T^k \bar{v}_1 + \dots + c_5 T^k \bar{v}_5$$
$$= c_1 \lambda_1^k v_1 + \dots + c_5 \lambda_5^k v_5$$

$T\bar{v}_1 = \lambda_1 \bar{v}_1$
$T\bar{v}_2 = \lambda_2 \bar{v}_2$
$\vdots$

so as $k \to \infty$

$$T^k \bar{x}_0 = c_1 1 \cdot \bar{v}_1 + c_2 0 \bar{v}_2 + \dots + c_5 0 \bar{v}_5$$
$$= c_1 \bar{v}_1$$

Also, $c_1\bar{v}_1$ will have to be a probability vector so $T^k \bar{x}_0$ is

- an eigenvector assoc. to $\lambda = 1$
- a vector whose entries sum to 1.

Summary

Def If T is a stochastic matrix, then $\bar{q}$ is a steady state vector for T if

- $T\bar{q} = \bar{q}$, and
- $\bar{q}$ is a prob. vector.

Theorem If T is a stochastic matrix s.t. some power of T contains only positive entries, then T has a unique steady-state vector $\bar{q}$, and any Markov chain defined by $\bar{x}_k = T\bar{x}_{k-1}$ converges to $\bar{q}$ as $k \to \infty$, i.e. $\bar{x}_\infty = \bar{q}$.

"regular"

Solving our Page rank problem (again)

1. Create the transition matrix T.
   If it isn't regular, then tweak it.
   our example: all entries in T^9 are positive ✓
2. Find any eigenvector associated to $\lambda = 1$. (using a computer)
   our example:
   using wolframalpha
   "eigenvector {{0, 1/2, 1/3, 1, 0}, {1, 0, ... }"
   $$\bar{v}_1 = \begin{bmatrix} 4 \\ 5\frac{1}{3} \\ 3 \\ \frac{1}{3} \\ 1 \end{bmatrix}$$
3. Divide out the sum of entries to get the steady-state vector $\bar{q}$. Entries of $\bar{q}$ give the page rank.
   our example
   sum entries in v_1 is $41/3$
   $$\bar{q} = \frac{3}{41}\bar{v}_1 \approx \begin{bmatrix} 0.293 \\ 0.390 \\ 0.220 \\ 0.024 \\ 0.073 \end{bmatrix}$$
   this gives the ranking

#1	#2	#3	#4	#5
B	A	C	E	D