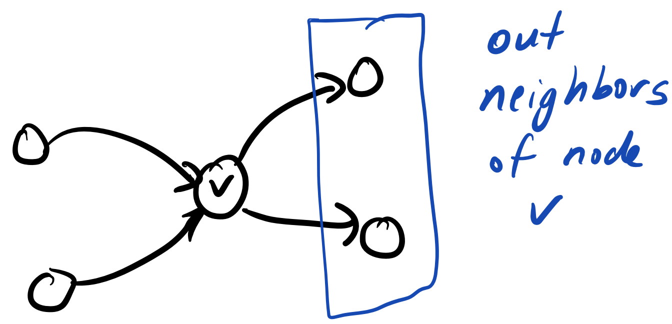
Random Walks and Centrality Scores in Graphs

<u>Definition</u>: Given a directed graph $G=(V,E)$ and a starting node $s \in V$, the standard random walk of length K is a sequence of random nodes

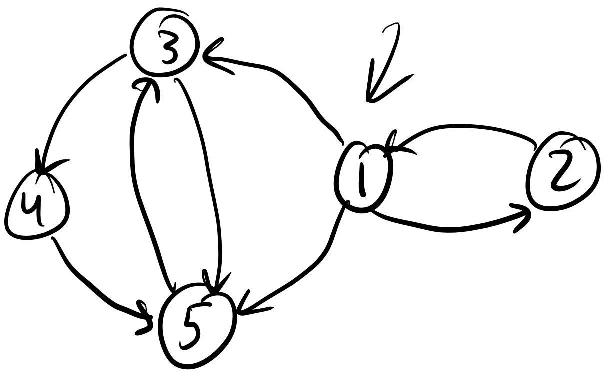
$$\underbrace{s=v_0, v_1, v_2, \ldots, v_K}_{\text{not necessarily distinct}}$$

such that node v_i is chosen uniformly at random from among the <u>out</u> neighbors of v_{i-1}.



We will use d_v to denote the <u>out</u>-degree of node v (# of out neighbors).

Math behind random walks



$$d = d_{out} = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 1 \\ 1 \end{bmatrix}$$

D is the degree matrix, $D_{ii} = d_i$

$$A = \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \end{array} \qquad D = \begin{bmatrix} 3 & & & & \\ & 1 & & & \\ & & 2 & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$A^T e_1$ gives me the set of nodes I can travel to if I start at node 1.

$$P = A^T D^{-1} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1/3 & & & & \\ & 1 & & & \\ & & \frac{1}{2} & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1/2 & 0 & 0 \\ 1/3 & 0 & 1/2 & 1 & 0 \end{bmatrix} \qquad \underline{P_{31}}$$

So $P_{i\underline{j}}$ gives me the probability of taking a step to node i if I start at node j. P is the "random walk matrix"!

$P^K x_s$ is the probability distribution for where I can be after K steps.

where $$x_s(i) = \begin{cases} 1 & \text{if } i = s \\ 0 & \text{otherwise} \end{cases}$$

If $e^T x = 1$, then $y = Px$ satisfies $y^T e = 1$.

A matrix with columns that sum to 1 (and has nonzero entries) is called a column stochastic matrix.

Observation: If $A = A^T$ then we have

$$\hat{d} = \frac{d}{e^T d}$$

$$P\hat{d} = A^T D^{-1} \hat{d}$$

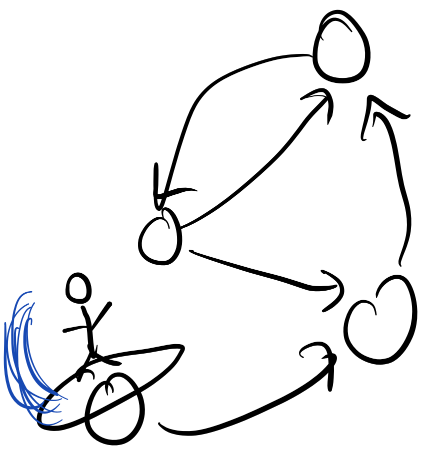
$$= A D^{-1} \hat{d} = \hat{d}$$

So $\hat{d}$ is the "steady state" vector for the graph.

PageRank

Notes follow paper "PageRank Beyond the Web" D. Gleich (SIAM Review)

Model : random surfer



At each step

- with probability $\alpha \in (0,1)$ surfer follows a uniform random out link
- with probability $(1-\alpha)$ teleport to a different node based on some distribution given by a vector $v \geq 0$ satisfying $v^T e = 1$

$v(i)$ = probability I teleport to node i.

Simple case $\quad v = \frac{e}{n}$

<u>PageRank centrality (informal)</u>

A webpage (node) is more "important" or "central" if the surfer spends a higher proportion of time there.

I.e. we want to compute the average amount of time the surfer spends at node i.

<u>Def</u>: A matrix P is column stochastic if $P \geq 0$ and $e^T P = e^T$.

It is $\underline{\text{sub}}$-stochastic if

$$e^T P \leq e^T$$

(columns sum to ≤ 1).

PageRank centrality (formal)

Let $G=(V,E)$ be a graph and let v be a column stochastic vector defining starting probabilities.

e.g. $\underline{v} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ $\quad \underline{v} = \begin{bmatrix} 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \\ 1/5 \end{bmatrix} = \frac{e}{n}$ $(n=5)$

With probability α we transition based on the standard random walk matrix

$$\boxed{P} = A^T D^{-1}$$

With probability $1-\alpha$ the surfer teleports to a node based on $\underline{v}$, i.e. teleportation matrix is

$$\underline{v}e^T = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & & \end{matrix} \\ \begin{bmatrix} v & v & v & \cdots & v \end{bmatrix} \end{array}$$

So the PageRank transition matrix is

$$M = \alpha P + (1-\alpha) v e^T$$

Let $x^{(0)} = v$ be the starting vector, then the probability distribution at step $i+1$ is given by

$$\underline{x^{(i+1)}} = \underline{M}\, x^{(i)}$$

$$= \alpha P x^{(i)} + (1-\alpha) v e^T x^{(i)}$$

1

$$= \alpha P x^{(i)} + (1-\alpha)v$$

If this procedure converges, that means $x^{(k)} \longrightarrow x$

$$x = \alpha P \underline{x} + (1-\alpha)v$$

$$\Rightarrow \quad x - \alpha P x = (1-\alpha)v$$

$$\Rightarrow \quad \boxed{(\underline{I} - \alpha \underline{P})x = (1-\alpha)v}$$

So the steady state vector x is the solution to a linear system.

This is also the solution to an eigenvector problem

$$\boxed{x = \underline{M}x.}$$

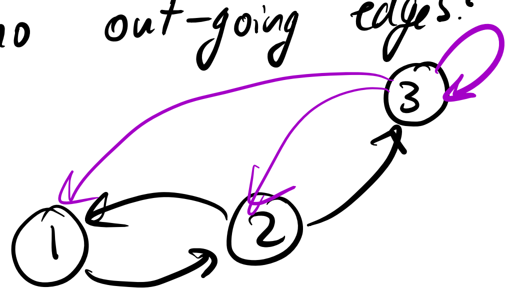
Definition: Let P be a column stochastic matrix. Assume $v \in \mathbb{R}^n$ satisfies $v^T e = 1$ and $v \geq 0$, and $\alpha \in (0,1)$. The PageRank linear system is given by

$$\boxed{(I - \alpha P)x = (1-\alpha)v} \qquad (*)$$

where x is called the PageRank vector.

Fact: As long as $\alpha \in (0,1)$ the solution x to $(*)$ exists, is unique and satisfies $x \geq 0$ and $e^T x = 1$.

Issue: What do we do if node i has no out-going edges?



$$D = \begin{bmatrix} 1 & & \\ & 2 & \\ & & 0 \end{bmatrix}$$

$$P \text{ "=" } A^T D^{-1} \qquad D^{-1} = \begin{bmatrix} 1 & & \\ & \frac{1}{2} & \\ & & \frac{1}{0}? \end{bmatrix}$$

Ideas: If I get to a dangling node:

- teleport based on v.
- add self-loop
- link it to every other node (and add self loop?)

The Pseudo-PageRank System

Definition: Let $\hat{P}$ be a column substochastic matrix ($e^T\hat{P} \le e^T$, $\hat{P} \ge 0$)
Let f be a nonnegative vector and $\alpha \in (0,1)$. The pseudo-PageRank problem is to solve the linear system

$$(I - \alpha\hat{P})y = f \quad (\star)$$

and y is called the Pseudo PageRank vector.

Theorem: Let y be the solution to $(\star)$ and define $v = \dfrac{f}{e^T f}$ and

$c = e - \hat{P}^T e \geq 0$ and $x = \dfrac{y}{e^T y}$.

Then x is the PageRank vector for the system $\boxed{(I - \alpha P)x = (1-\alpha)v}$

where $P = \hat{P} + vc^T$.

Proof: Check that v is a stochastic vector. ✓

Check that P is a column stochastic matrix:

$$\boxed{e^T P} = e^T(\hat{P} + vc^T)$$

$$= \underline{e^T \hat{P}} + \boxed{\cancel{e^T v}}^{\,1} \underline{c}^T = \boxed{\underline{e}^T}$$

because $c = e - \hat{P}^T e$

$\Rightarrow \quad e = c + \hat{P}^T e$

$\Rightarrow \quad \underline{e}^T = \underline{e^T \hat{P} + c^T}$

Note from $(I - \alpha P)x = (1-\alpha)v$ that

$$x = \alpha \underline{P} x + (1-\alpha)v$$

$$= \alpha \underline{\hat{P}} x + \alpha \underline{v} c^T x + (1-\alpha)v$$

$$= \alpha \hat{P} x + [\alpha c^T x + (1-\alpha)]\underline{v}$$

$$= \alpha \hat{P} x + \underline{\boxed{[\alpha c^T x + (1-\alpha)]} f}$$

$$\underbrace{e^T f}_{\text{call this } \gamma}$$

$$= \alpha \hat{P} x + \gamma f$$

$$\Rightarrow \quad (I - \alpha \hat{P}) x = \gamma f$$

$$\Rightarrow \quad (I - \alpha \hat{P}) \underbrace{\boxed{\frac{x}{\gamma}}}_{\underline{y}} = f$$

$$\Rightarrow \quad x = \gamma y$$

and since $x^T e = 1$

$$\Rightarrow \quad \gamma y^T e = 1 \Rightarrow \gamma = \frac{1}{y^T e}$$

$$\Rightarrow \quad x = \frac{y}{y^T e}. \qquad \blacksquare$$

Message: Every Pseudo-PageRank system corresponds to a standard PageRank system.

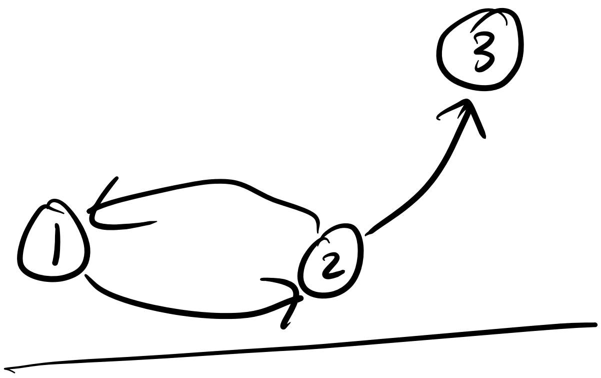
Let's fix the dangling node issue

Define D^+ to be the diagonal matrix

$$D^+_{ii} = \begin{cases} \frac{1}{d_i} & \text{if } d_i > 0 \\ 0 & \text{otherwise} \end{cases}$$

To define PageRank we use the matrix

$$\hat{P} = A^T D^+ \qquad e^T \hat{P} \leq e^T$$



$$D^+ = \begin{bmatrix} 1 & & \\ & \frac{1}{2} & \\ & & 0 \end{bmatrix}$$

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

We define a correction vector by

$$c = e - \hat{P}^T e$$

in the example:

$$c = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$c[i] = \begin{cases} 1 & \text{if } i \text{ is dangling} \\ 0 & \text{otherwise} \end{cases}$$

Options for solving dangling node issue

1. Solve pseudo-PR system

$$(I - \alpha \hat{P})y = f \quad \text{where}$$

$f = v$. By our theorem, this will correspond to the PageRank system

$$\boxed{(I - \alpha P)x = (1-\alpha)v}$$

where $\quad P = \hat{P} + vc^T$

e.g. $\quad vc^T = v[0 \ 0 \ 1]$

$$= \begin{bmatrix} \underline{0} & \underline{0} & \underline{v} \end{bmatrix}$$

$$\boxed{P = \begin{bmatrix} 0 & 1/2 & v \\ 1 & 0 & \underline{v} \\ 0 & 1/2 & \end{bmatrix}}$$

$\underline{\alpha P}$

This is called "strongly-preferential" PageRank

② "weakly preferential"

$$P = \hat{P} + cu^T \quad \text{where} \quad u = \frac{e}{n}$$

③ "sink preferential"

$$P = \hat{P} + \mathrm{diag}(c)$$

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

Personalized PageRank System

Define $v_R = \begin{cases} \frac{1}{|R|} & \text{if } i \in R \\ 0 & \text{otherwise} \end{cases}$

where $R \subseteq V$ is a subset of nodes in the graph, Then

$$(I - \alpha P)x = (1-\alpha) v_R$$

is a Personalized PageRank system and x is the PPR vector.

often $R = \{v\}$ for one node $v \in V$.