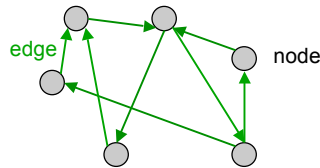


PageRank

Goal

- Given a directed graph with n nodes
- Assign each node a score that represents its importance in structure
 - Call score **PageRank**: $pr(\text{node})$



1

Conferring importance

Core ideas:

- A node should **confer** some of its importance **to the nodes to which it points**
 - If a node is important, the nodes it links to should be important
- A node should **not transfer more** importance **than it has**

2

Attempt 1

Refer to nodes by numbers $1, \dots, n$ (arbitrary numbering)
 Let t_i denote the number of edges out of *node* i (outdegree)

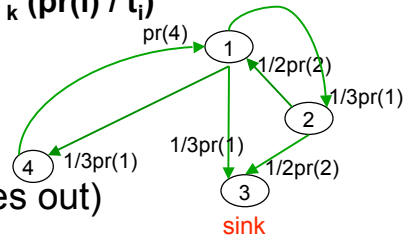
Define

$$pr_{\text{new}}(\mathbf{k}) = \sum_{i \text{ with edge from } i \text{ to } k} (pr(i) / t_i)$$

Iterate until converges

Problems

- Sinks (nodes with no edges out)
- Cyclic behavior



3

Attempt 2

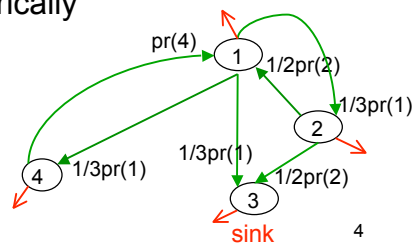
Random walk model

- Attempt 1 gives movement from node to linked neighbor with probability $1/\text{outdegree}$
- Add **random jump to any node**

$$pr_{\text{new}}(\mathbf{k}) = \alpha/n + (1-\alpha) \sum_{i \text{ with edge from } i \text{ to } k} (pr(i) / t_i)$$

– α parameter chosen empirically

- Break cycles
- Escape from sinks



4

Normalized?

- Would like $\sum_{1 \leq k \leq n} (\text{pr}(k)) = 1$
- Consider $\sum_{1 \leq k \leq n} (\text{pr}_{\text{new}}(k))$

$$= \sum_{1 \leq k \leq n} (\alpha/n + (1-\alpha) \sum_{i \text{ with edge from } i \text{ to } k} (\text{pr}(i) / t_i))$$

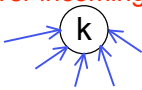
$$= \sum_{1 \leq k \leq n} (\alpha/n) + \sum_{1 \leq k \leq n} ((1-\alpha) \sum_{i \text{ with edge from } i \text{ to } k} (\text{pr}(i) / t_i))$$

$$= \alpha + (1-\alpha) \sum_{1 \leq k \leq n} \sum_{i \text{ with edge from } i \text{ to } k} (\text{pr}(i) / t_i)$$

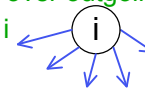
$$= \alpha + (1-\alpha) \sum_{1 \leq i \leq n} \sum_{k \text{ with edge from } i \text{ to } k} (\text{pr}(i) / t_i)$$

$$= \alpha + (1-\alpha) \sum_{i \text{ with edge from } i} \text{pr}(i)$$

*inner sum \sum_i over incoming edges for one k



*inner sum \sum_k over outgoing edges for one i



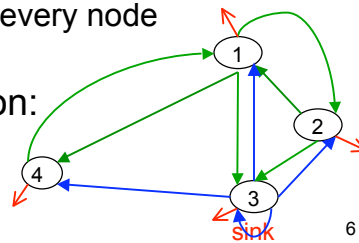
5

Problem for desired normalization

- Have $\sum_{1 \leq k \leq n} (\text{pr}_{\text{new}}(k)) = \alpha + (1-\alpha) \sum_{i \text{ with edge from } i} \text{pr}(i)$
- **Missing pr(i)** for nodes with no edges from them
 - sinks!
- **Solution:** add n edges out of every sink
 - Edge to every node including self
 - Gives 1/n contribution to every node

Gives desired normalization:

If $\sum_{1 \leq k \leq n} (\text{pr}_{\text{initial}}(k)) = 1$
 then $\sum_{1 \leq k \leq n} (\text{pr}(k)) = 1$



6

Matrix formulation

- Let E be the n by n adjacency matrix
 $E(i,k) = 1$ if there is an edge from node i to node k
 $= 0$ otherwise
- Define **new matrix L** :
For each row i of E ($1 \leq i \leq n$)
If row i contains $t_i > 0$ ones, $L(i,k) = (1/t_i) E(i,k)$, $1 \leq k \leq n$
If row i contains 0 ones, $L(i,k) = 1/n$, $1 \leq k \leq n$
- Vector **pr** of PageRank values defined by
$$pr = (\alpha/n, \alpha/n, \dots, \alpha/n)^T + (1-\alpha) L^T pr$$
- has a solution representing the **steady-state values $pr(k)$**

7

Calculation

- Choose α
 - No single best value
 - Page and Brin originally used $\alpha = .15$
- Simple iterative calculation
 - Initialize $pr_{\text{initial}}(k) = 1/n$ for each node k
so $\sum_{1 \leq k \leq n} (pr_{\text{initial}}(k)) = 1$
 - $pr_{\text{new}}(k) = \alpha/n + (1-\alpha) \sum_{1 \leq i \leq n} L(i,k) pr(i)$
- Converges
 - Has necessary mathematical properties
 - In practice, choose convergence criterion
 - Stops iteration

8