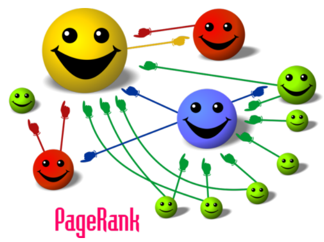
Lecture 12
Introduction to Link Analysis



Thx to Dan Weld, James Moody, Dragomir Radev

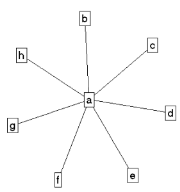
Organization

- Exams handed back at end of class
  - Model answers online later today
- Today:
  - Graph analysis (PageRank and HITS)
- Later:
  - Text handling and in-depth inverted index
  - Crawler design (Mercator)
  - More search architecture / advanced topics

2

Graphs (or Networks)

- Describe relation among items
- Symmetric or directed
- Have been around for a long time
  - Friendship networks
  - Board membership
  - Paper citations
  - US power grid
  - Web pages



3

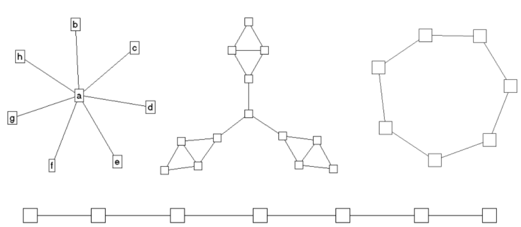
Prestige & Importance

- Which node(s) are the most important?
- How would you measure it?
  - # links?
  - # "2-deep links"?
  - position in the graph?
- This is also sometimes called determining "centrality", especially in social network research

4

Prestige & Importance

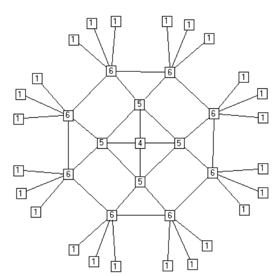
- Which node(s) are the most important?



5

Prestige & Importance

- Degree centrality is one way
  - Just count the links!



6

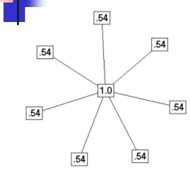
Prestige & Importance

- Another way: measure closeness
  - Node is important if it is close to all others
  - Based on inverse of distance from each node to every other node

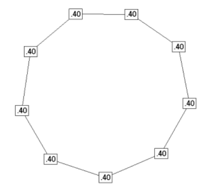
$$C_c(n_i) = \left[\sum_{j=1}^{g} d(n_i, n_j) \right]^{-1}$$

7

Closeness



Distance	Closeness	normalized
0 1 1 1 1 1 1 1	.143	1.00
1 0 2 2 2 2 2 2	.077	.538
1 2 0 2 2 2 2 2	.077	.538
1 2 2 0 2 2 2 2	.077	.538
1 2 2 2 0 2 2 2	.077	.538
1 2 2 2 2 0 2 2	.077	.538
1 2 2 2 2 2 0 2	.077	.538
1 2 2 2 2 2 2 0	.077	.538



Distance	Closeness	normalized
0 1 2 3 4 4 3 2 1	.050	.400
1 0 1 2 3 4 4 3 2	.050	.400
2 1 0 1 2 3 4 4 3	.050	.400
3 2 1 0 1 2 3 4 4	.050	.400
4 3 2 1 0 1 2 3 4	.050	.400
4 4 3 2 1 0 1 2 3	.050	.400
3 4 4 3 2 1 0 1 2	.050	.400
2 3 4 4 3 2 1 0 1	.050	.400
1 2 3 4 4 3 2 1 0	.050	.400

8

Closeness

Distance	Closeness	normalized
0 1 2 3 4 5 6	.048	.286
1 0 1 2 3 4 5	.063	.375
2 1 0 1 2 3 4	.077	.462
3 2 1 0 1 2 3	.083	.500
4 3 2 1 0 1 2	.077	.462
5 4 3 2 1 0 1	.063	.375
6 5 4 3 2 1 0	.048	.286

9

Prestige & Importance

- Other ideas:
  - Identify nodes with smallest max-distance to all other nodes
  - *Betweenness* - for what fraction of paths is the node along the path?
  - Bonacich Power Centrality, aka *Proximity-to-prestige* - a node's importance depends on the importance of its neighbors
  - Academic impact analysis
- These ideas came about before the Web, but very relevant

10

Web Link Analysis

- Search in late 1990s was pretty bad
  - Content growth outstripped human editors
- Lots of Web interest in 1997-1999 in using the hyperlink graph
  - **PageRank**, Page
  - **HITS**, Kleinberg
  - **"Silk from a sow's ear"**, Pirolli, Pitkow, Rao
- Can measure "importance", but that's not all

11

PageRank

- For first time, SEs got the right page
  - AltaVista used to rank pages by URL length
  - When PageRank hit, it was astonishing
- Intuition:
  - Web is a big directed graph
  - A "random surfer" clicks at random
  - Importance of a page = probability the surfer is on the page
  - Suppose *P* has *N* forward links; surfer clicks on link with probability *1/N*
  - Query-independent!!!

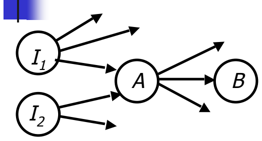
12

PageRank Intuition

- You have an adjacency matrix E where e[i,j]=1 if i cites j
  - It describes the Web
- Each node in the graph gets a PageRank score, p_u for node u
- Each site in the Web votes for important sites by linking to them
  - Weigh votes acc. to importance of sender
  - How is importance of sender determined?
  - With its PageRank score!
- PageRank is defined recursively (and computed iteratively)

13

PageRank



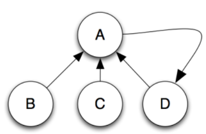
- A node with C links contributes 1/C of its PageRank to each target node

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

- Damping factor d... coming shortly...

14

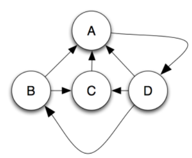
PageRank Example



- Total PR = 1, so init each node to 0.25
- PR(A) = (0.15/3) +
  0.85 * (0.25/1 + 0.25/1 + 0.25/1)
- PR(A) = 0.6875

15

PageRank Example 2



- Again, init all nodes to 0.25
- PR(A) = (0.15/3) +
  0.85 * (0.25/2 + 0.25/1 + 0.25/3)
- PR(A) = .05 + .85*(0.125 + 0.25 + 0.083)
- PR(A) = 0.4393

16

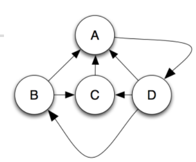
Some extra bits

- What about complicated graphs?
  - Algorithm keeps updating until it meets "stopping criteria"
- Rank sinks
  - Regions of the graph that accumulate rank, but do not distribute it externally
  - Can drain rank from the rest of the system
  - Soln: with probability (1-d), random surfer types in a random URL instead of clicking a link
- Dangling links
  - Nodes with no outlinks are disallowed

17

Let's try it

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$

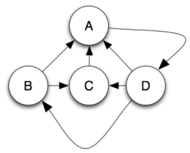


A	B	C	D
0.25	0.25	0.25	0.25

18

Let's try it

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$



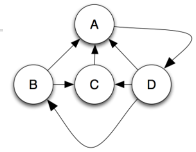
PR(A) = 0.0375 + 0.85(0.25/2 + 0.25/1 + 0.25/3)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.428			

19

Let's try it

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$



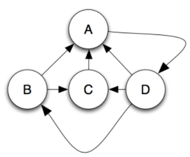
PR(B) = 0.0375 + 0.85(0.25/3)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.428	0.109		

20

Let's try it

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$



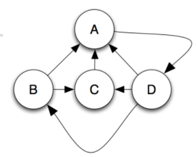
PR(C) = 0.0375 + 0.85(0.25/2 + 0.25/3)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.428	0.109	0.215	

21

Let's try it

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$



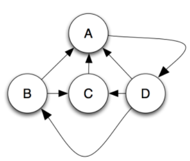
PR(D) = 0.0375 + 0.85(0.25/1)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.427	0.108	0.215	0.25

22

Let's try it

$$PR(A) = \frac{(1-d)}{N} + d\sum_i \frac{PR(I_i)}{C(I_i)}$$



PR(D) = 0.0375 + 0.85(0.25/1)

A	**B**	**C**	**D**
0.25	0.25	0.25	0.25
0.427	0.108	0.215	0.25
0.337	0.108	0.154	0.401
0.328	0.151	0.197	0.324
0.361	0.129	0.193	0.317

23

PageRank Matrix

- Every node has prestige value p[v]
  - p sometimes called the "rank vector"
  - Transition matrix E holds transition probs
  - P' = E^T P
    - PageRank is induced from the adjacency matrix

$$p'[v] = \sum_u E^T[v,u]p[u] = \sum_u E[u,v]p[u]$$

New value of p[v] is the sum of all values incoming to v

24

Adding PageRank to a SE

- Weighted sum of page importance and query-similarity
- Score(query, doc)=
  - w*sim(q, p) + (1-w) * PR(p)
    - If sim(q, p) > 0
    - Otherwise, 0
- Where:
  - 0 < w < 1
  - Values sim(q,p) and R(p) are normalized

25

Hubs and Authorities

- Due to Kleinberg, 1997
- Unlike PageRank, is query-dependent

- A page is a good ***authority*** if it is pointed-to by many good ***hubs***
- A page is a good ***hub*** if it is pointed-to by many good ***authorities***
- Good hubs and authorities reinforce each other

26

HITS algorithm

- Hyperlink-Induced Topic Search
  - Obtain root set using input query
  - Expand the root set by radius one
  - Run iterations on the hub and authority scores together
  - Report top-ranking authorities and hubs

$$auth(p) = \sum_{i=1}^{n} hub(i)$$

$$hub(p) = \sum_{i=1}^{n} auth(i)$$

27

More HITS

- Init all hub() and auth() scores to 1
- Repeat k times
- After each step, normalize the scores to prevent them from going to infinity
  - Like PR, scores will converge

28

Some exam notes

- Exams handed back shortly
  - Mean = 37.08 (out of 50)
  - Stddev = 5.77
- Model answer available online shortly
- Exam scores will be scaled/adjusted appropriately and as needed at end of class

29