Information Networks: PageRank

Computational Social Systems I (VU) (706.616)

Elisabeth Lex

ISDS, TU Graz

June 4, 2020

Repetition

- Information Networks
- Shape of the Web
- Hubs and Authorities

PageRank

- Intuition from last time: links as votes (HITS algorithm)

- Page more important if it has many in-links

PageRank Algorithm

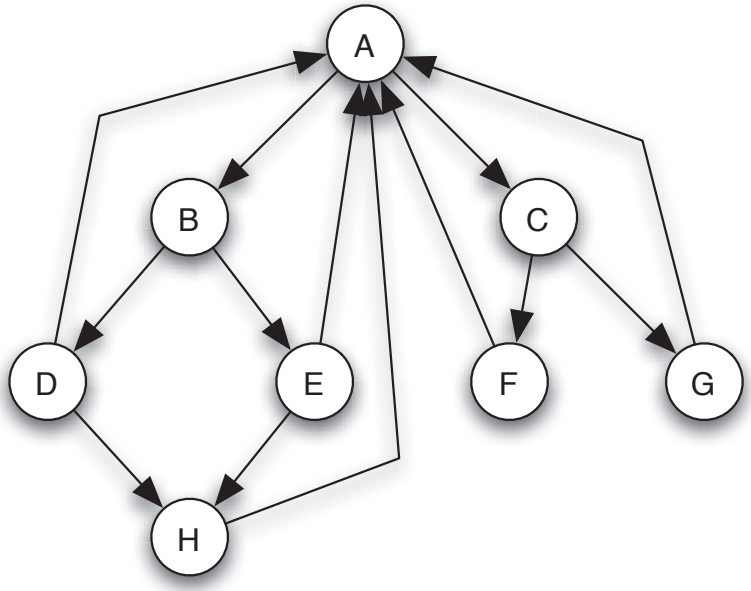
- Method to determine link popularity of a page
- The more links point to a page, the higher its weight (i.e., its PageRank)
- If page linked from page with high PageRank, the higher the PageRank of page
- Aim of algorithm: sort links according to PageRank to derive ranking
- Intuition: random surfer

The basic definition of PageRank

- We compute PageRank in the following way:

1. Given a Web graph with n nodes assign each node initial PageRank $1/n$
2. Choose a number of steps k
3. Perform a sequence of k updates where we calculate rank of each node: each page divides and passes its current PageRank equally across its out-going links. Each page updates its new PageRank to be the sum of the shares it receives.

PageRank Example: First two steps

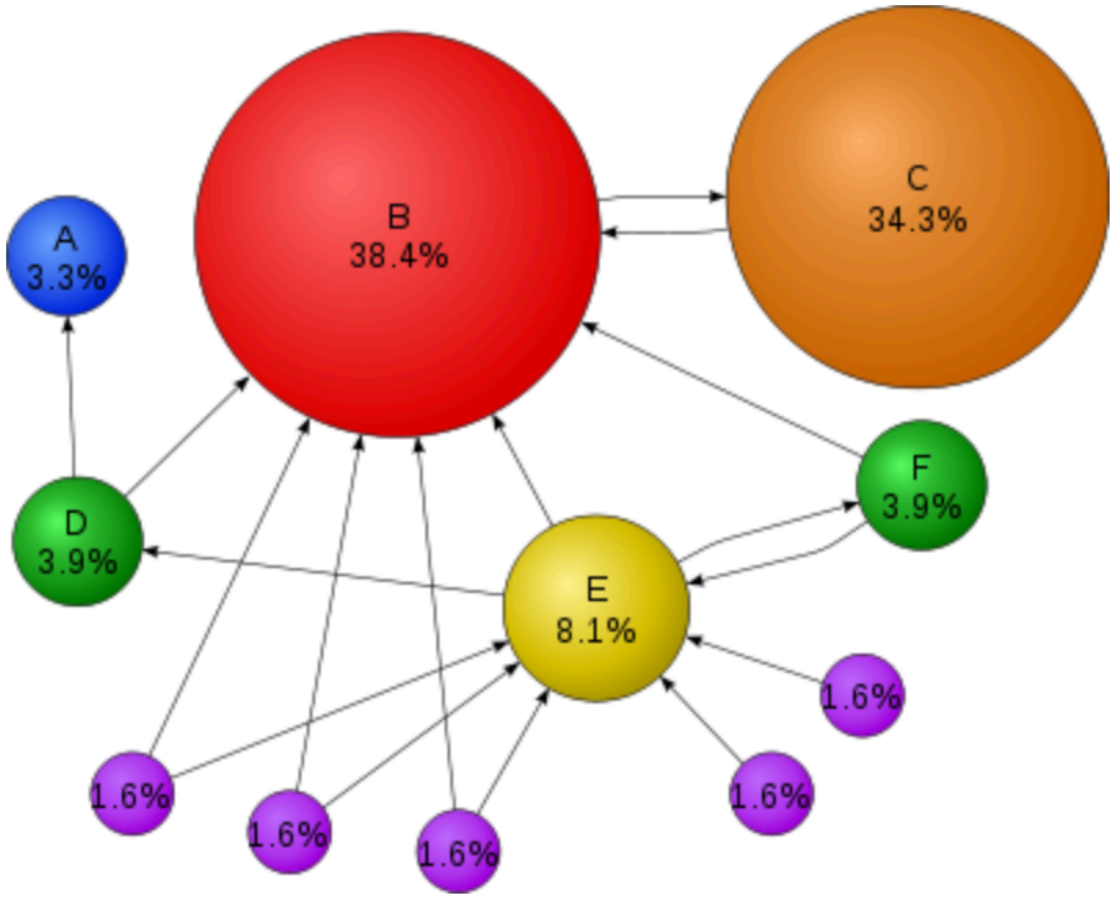


All pages start out with a PageRank of 1/8. PR(A) =1/2. It gets all of F, G, H and half of D and E. What about B and C?
PR(B) and PR(C): get half of A's PR, so only 1/16

Step \ Page	A	B	C	D	E	F	G	H
1	1/2	1/16	1/16	1/16	1/16	1/16	1/16	1/8
2	5/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

Careful: error in book, page 408!

Example for PageRank Scores



PageRank vs HITS

- Can you think of the major difference between PageRank and HITS?

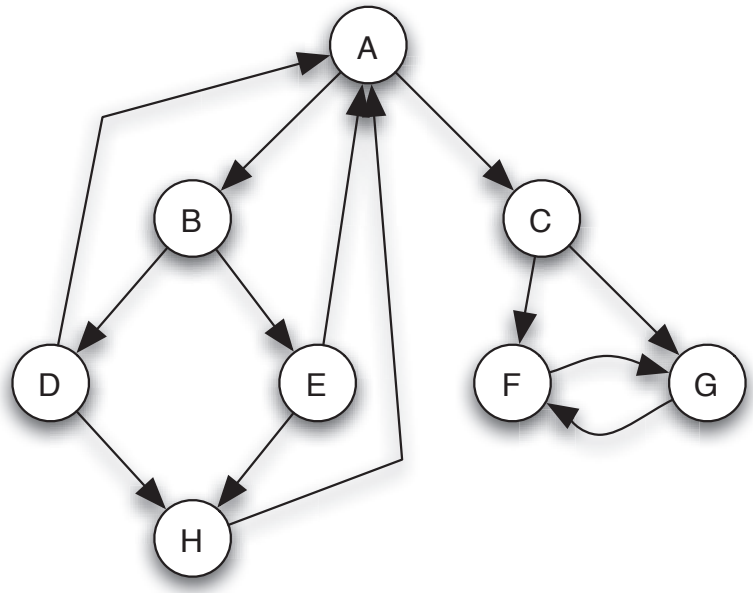
- Unlike HITS, PageRank is independent of the search query!

Equilibrium Values of PageRank

- Similarly as with hub-authority computation if we increase the number of iteration steps k the values will become stable and will not change anymore
- The calculation converges and we reach an equilibrium
- One can prove this convergence
- One can also prove that for a strongly connected network the equilibrium values are unique

A basic problem with PageRank: Example

- Now, F and G point to each other and not to A
- PageRank that flows from C to F and G can never flow back to the network
- Links out of C - “slow leak”, all the PageRank ends up at F and G



Results in convergence to PageRank of 1/2 for each of F and G
Others have PageRank of 0

A basic problem with PageRank

- Wrong nodes may end up with "all" the PageRank
- If graph is not strongly connected - complete PageRank will leak to nodes in OUT
- Therefore: scale down all values by a scaling factor s (strictly between 0 and 1)
- Divide the residual $1 - s$ equally over all nodes giving $(1 - s)/n$ to each
- Preserves the total PageRank in the network - based on redistribution
- Can be shown that this rule converges and that no PageRank is leaking

Limit of the Scaled Update

- Repeated application converges to set of limiting PageRank values as number of updates k goes to infinity
- These limiting values form the unique equilibrium: unique set of values that remains unchanged under application of update rule
- Depend on choice of scaling factor s
- In practice: scaling factor s usually between 0.8 and 0.9

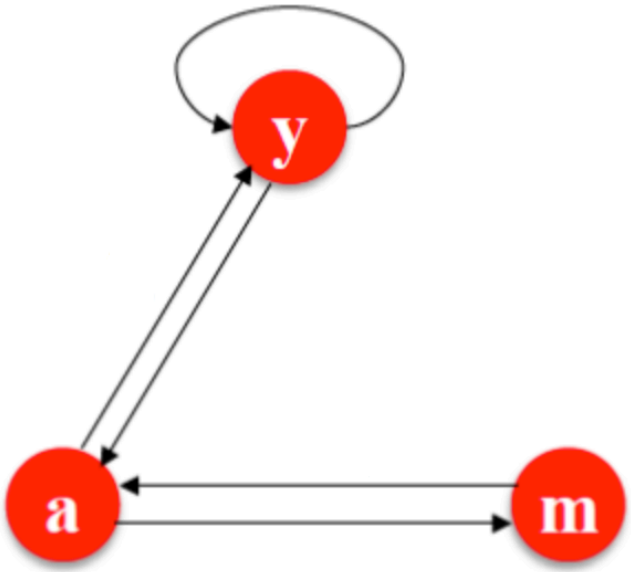
The "Flow" model of PageRank

- Remember: page important if linked from other important pages
- Plus, a "vote" (via in-link) from an important page is worth more
- Based on that, we can define the rank" r_j for a page j

$$r_j = \sum_{i \to j} \frac{r_i}{d_i} \tag{1}$$

  where d_i is the out-degree of node i
- Intuitively, we can think of PageRank as a kind of fluid that "flows" through the network
- The fluid passes from node to node across links
- It pools at the nodes that are the most important

Example



$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m/1$$
$$r_m = r_a/2$$

These are called “flow” equations

Solving the equations

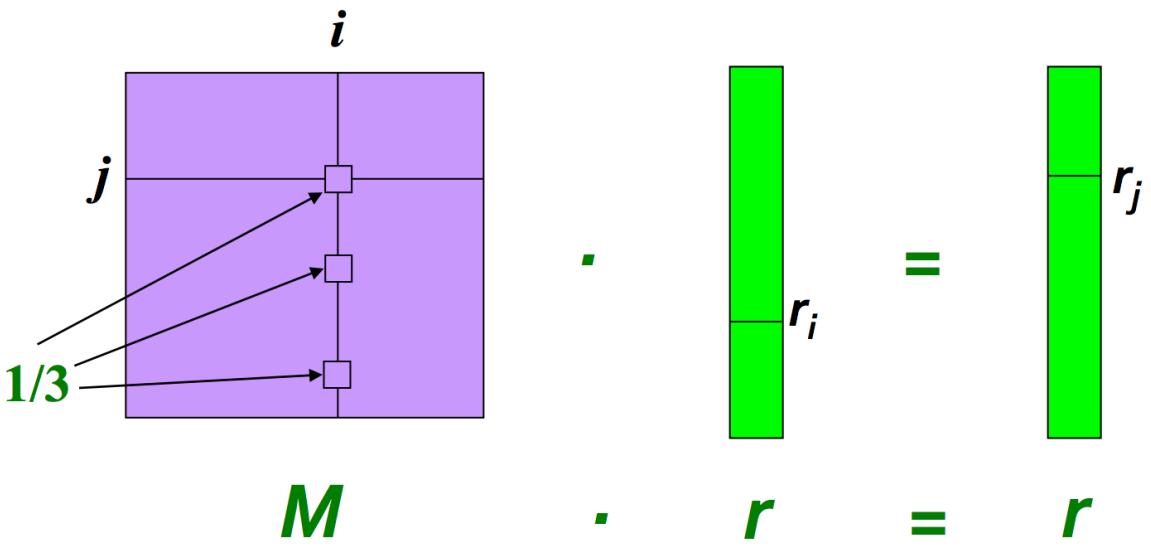
- In the last example: 3 equations, 3 unknowns, no constants
- This means, there is no unique solution to them
- We need an additional constraint to enforce unique solution
- Ranks need to sum up to 1, i.e.
- This means that for our small graph:
  $r_y = r_y/2 + r_a/2$
  $r_a = r_y/2 + r_m/1$
  $r_m = r_a/2$
  $r_y + r_a + r_m = 1$
- So we have 3 unknowns and 4 equations - solvable through elimination
- Solution: $r_y = 2/5$, $r_a = 2/4$, $r_m = 1/5$

Matrix Formulation

- Elimination does not apply to large scale graphs
- We need a different formulation of the problem
- Matrix formulation: Stochastic adjacency matrix M
  - Let page i have d_i out-links
  - if $i \rightarrow j$, then $M_{ji} = \frac{1}{d_i}$ else $M_{ji} = 0$ where M is a column stochastic matrix, i.e., columns sum up to 1
- Rank vector r: vector with an entry per page
- Length of r is the number of pages in our sample
  - r_i corresponds to pagerank score of page i
  - $\sum_i r_i = 1$ due to constraint of flow equations
- Thus, we can write the flow equations from before as vector-matrix product:
  $r = M \cdot r$

Matrix Formulation: Example

- Flow equation as sum: $r_j = \sum_{i \to j} \frac{r_i}{d_i}$
- Flow equations in matrix form as vector-matrix product: $r = M \cdot r$
- Let's assume page i, which has links to 3 pages, i.e. $d_i = 3$. One of the pages it links to is page j.



Matrix Formulation

- Recursive maxtrix equation $r = M \cdot r$ resembles an eigenvalue problem

- Eigenvalue problem definition:

Definition

Vector x is an eigenvector with the corresponding eigenvalue λ if they are a solution to the following problem: $Ax = \lambda x$

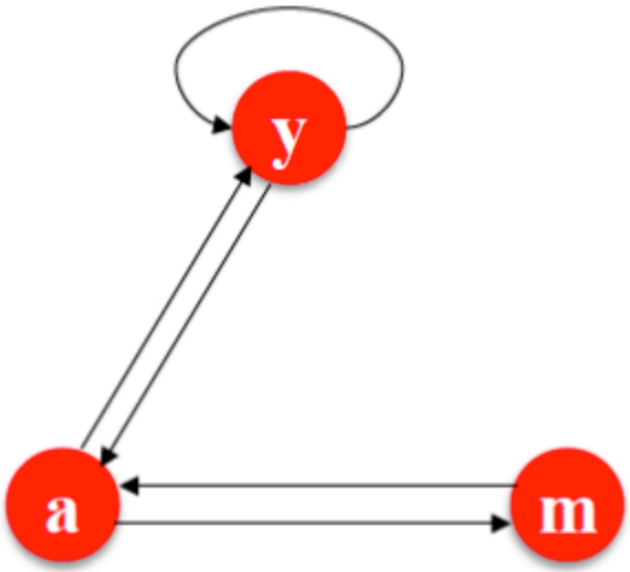
Note that A is given, and we aim to compute x and λ

Matrix Formulation

- Equation $r = M \cdot r$ looks similar to $Ax = \lambda x$
- In other words: rank vector r is an eigenvector of stochastic web matrix M
- What is the value of λ?
- Rank vector r is not any eigenvector, but its *principal* eigenvector, i.e., its corresponding eigenvalue is 1, ergo $\lambda = 1$
- Reason: vector r has unit length (its coordinates are nonnegative and sum to 1, also called "stochastic vector")
- Plus, each column of M sums up to 1 (M is "column stochastic")
- This means: $M \cdot r \leq 1$
- Hence, largest eigenvalue of $M = 1$

This can be efficiently solved for r using Power iteration method

Example:



	y	a	m
y	½	½	0
a	½	0	1
m	0	½	0

$$r = M \cdot r$$

$$\begin{bmatrix} y \\ a \\ m \end{bmatrix} = \begin{bmatrix} ½ & ½ & 0 \\ ½ & 0 & 1 \\ 0 & ½ & 0 \end{bmatrix} \begin{bmatrix} y \\ a \\ m \end{bmatrix}$$

$$r_y = r_y/2 + r_a/2$$
$$r_a = r_y/2 + r_m/1$$
$$r_m = r_a/2$$

Random Walks Interpretation: An equivalent definition of PageRank

- So far, we computed PageRank using flow equations and in terms of matrix formulation
- Now, we will look at an interpretation what PageRank scores reflect
- Random Walk Interpretation
- PageRank scores equivalent to probability distribution of a random walker on the graph

Random Walk Interpretation: An equivalent definition of PageRank

Consider someone who is randomly browsing a network of Web pages - a "random web surfer"

- Surfer starts at any time t by choosing a page i at random, picking each page with equal probability
- At time $t+1$, surfer picks uniformly at random an out-going link from page i and follows it
- Ends up on some page j linked from i
- Process repeats indefinitely
- If page j has no out-going links, surfer stays
- This is called a **Random Walk** on the network

Random Walk Interpretation

With what probability is a random walker at time t at a given page?

- Let $p(t)$ be a vector whose coordinate i denote the probability that the surfer is at page i at time t
- Thus, $p(t)$ gives us a probability distribution over pages

Random Walk Interpretation

Where is the random walker going to be at time $t+1$?

- Random walker follows an out-going link uniformly at random
- Thus: $p(t+1) = M \cdot p(t)$
- Suppose random walk reaches a state $p(t+1) = M \cdot p(t) = p(t)$, then $p(t)$ is called stationary distribution of a random walk
- Remember: rank vector $r = M \cdot p(t)$
- In other words: r is a stationary distribution for the random walk

What does that mean?

- PageRank scores correspond to probability is at a given node at a given time step
- A side note: random walks are effectively Markov processes
- Why is that important? Because for graphs that satisfy certain conditions, the stationary distribution is unique and will be reached at some point regardless of the initial probability distribution at time $t = 0$
- This means that there are conditions under which PageRank vector r is unique and will be achieved regardless of initialization

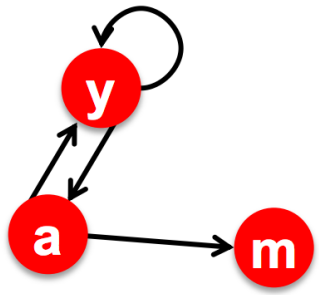
Problems with real web graphs

- Spider traps: a group of one or more pages that have no links out of the group will eventually accumulate all the importance of the Web
  - Problem: eventually, group absorbs all the PageRank scores
- Dead ends: some pages have no out-links
  - Problem: dead ends make PageRank "leak out"

Random Teleports as solution

- At each step, random surfer has 2 options:
  - With probability β, follow a link at random
  - With probability $1 - \beta$, jump to some random page
  - In practice, $\beta = 0.8, 0.9$
- This enables random walker to teleport out of dead end within few time steps

Problem with dead ends



- Problem: Pages with 0 out-degree - their PageRank does not get distributed ("leaks out")
- What is apparent if we look at matrix M?

	y	a	m
y	½	½	0
a	½	0	0
m	0	½	0

It is not stochastic anymore! Why? Node m has 0 out-degree

Solution: Teleports

- Teleports help us make matrix M **stochastic**
  - Whenever all the entries for a column in M are 0, we can replace them with $1/d_i$ where d_i is the out-degree of node i
- Teleports help us make matrix M **aperiodic**
  - Random walker can teleport out of loops
- Teleports help us make matrix M **irreducible**[1]
  - Teleports help us add random jumps to matrix M

[1] Irreducability: from any state, there is a non-zero probability of going from any one state to any other

Google's Solution: Random Jumps

Idea: combines all of this:

- Make matrix M stochastic, aperiodic, irreducible
- At each step, random surfer has 2 options:
  - With probability β, follow a link at random
  - With probability $1 - \beta$, jump to some random page
- PageRank equation [Brin-Page, 1998]

$$r_j = \sum_{i \to j} \beta \frac{r_i}{d_i} + (1 - \beta)\frac{1}{n} \tag{2}$$

Limitations of PageRank

- PageRank measures general popularity / importance of a page - Why a problem?
  - Neglects topic-specific authorities
  - Topic-specific PageRank
- Susceptible to Link spam
  - Link structures created to boost PageRank
  - Solution: TrustRank

Some Practical Examples for PageRank

- PageRank on protein interaction graphs[2]
- Social Media Analysis[3]
- Altmetrics and Analysis of Readership Data on Mendeley[4]

[2] http://rsos.royalsocietypublishing.org/content/2/4/140252.abstract

[3] http://www.cs.columbia.edu/~ecj2122/research/social_higgs/jubb_facheris_discovery_of_the_higgs.pdf

[4] http://arxiv.org/abs/1504.07482

Summary

We have learned about:

- PageRank
- Random Walks
- Problems with PageRank: dead ends
- Teleportation as solution
- Some applications beyond web search and ranking

Thanks for your attention

elisabeth.lex@tugraz.at

Slides use figures and content from Mining of Massive Datasets by Jure Leskovec, Anand Rajaraman, Jeff Ullman. See http://www.mmds.org/