Link Analysis

Leonid E. Zhukov

School of Data Analysis and Artificial Intelligence
Department of Computer Science
National Research University Higher School of Economics

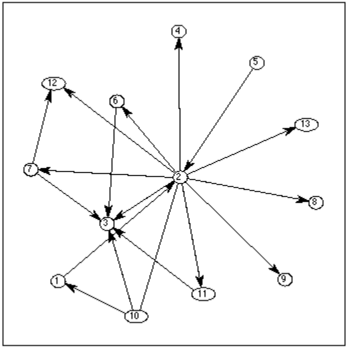
Structural Analysis and Visualization of Networks

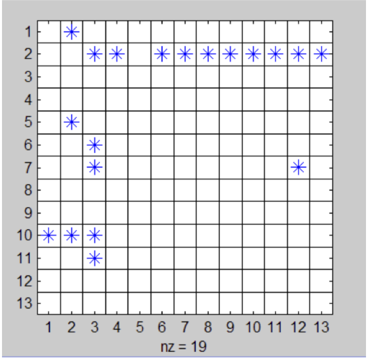
Lecture outline

1. Graph-theoretic definitions
2. Web search engines
3. Web page ranking algorithms
   – Pagerank
   – HITS
4. The Web as a graph
5. PageRank beyond the web

Graph theory

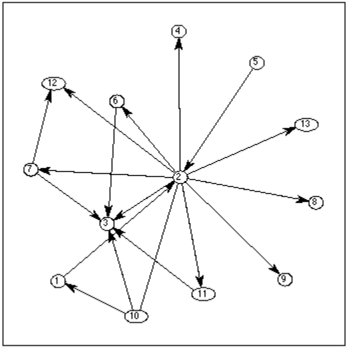
Graph $G(E, V)$, $|V| = n$, $|E| = m$
Adjacency matrix $\mathbf{A}^{n \times n}$, A_{ij}, edge $i \to j$





Graph is directed, matrix is non-symmetric: $\mathbf{A}^T \neq \mathbf{A}$, $A_{ij} \neq A_{ji}$

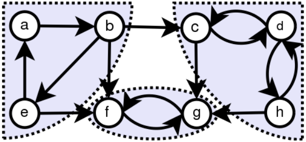
Graph theory



- sinks: zero out degree nodes, $k_{out}(i) = 0$, absorbing nodes
- sources: zero in degree nodes, $k_{in}(i) = 0$

Graph theory

- Graph is **strongly connected** if every vertex is reachable form every other vertex.
- **Strongly connected components** are partitions of the graph into subgraphs that are strongly connected



- In strongly connected graphs there is a path is each direction between any two pairs of vertices

image from Wikipedia

Graph theory

- A directed graph is **aperiodic** if the greatest common divisor of the lengths of its cycles is one (there is no integer $k > 1$ that divides the length of every cycle of the graph)

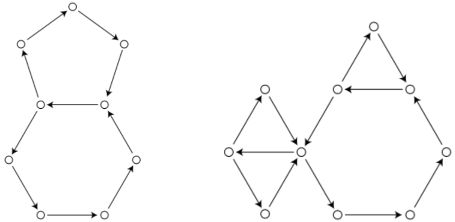
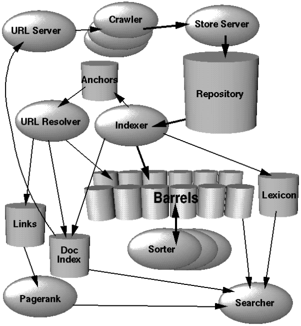


image from Wikipedia

Web search engine

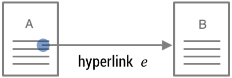
"The Anatomy of a Large-Scale Hypertextual Web Search Engine"



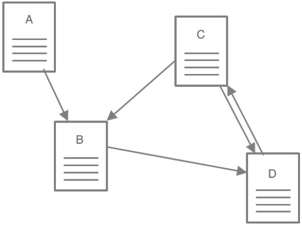
Sergey Brin and Lawrence Page, 1998

Web as a graph

- Hyperlinks - implicit endorsements



- Web graph - graph of endorsements (sometimes reciprocal)



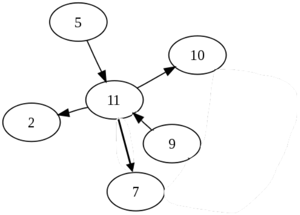
Ranking on directed graph

- iteratively update

$$r_i \leftarrow \sum_{j \in N(i)} r_j = \sum_j A_{ji} r_j$$

$$r_i^{t+1} = \sum_j A_{ji} r_j^t, \quad \text{with} \quad r_j^{t=0} = r_j^0$$

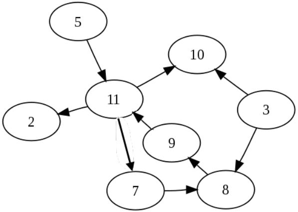
$$\mathbf{r}^{t+1} = \mathbf{A}^T \mathbf{r}^t, \qquad \mathbf{r}^{t=0} = \mathbf{r_0}$$



- norm $||\mathbf{r}^{t+1}|| \geq ||\mathbf{r}^t||$

Ranking on directed graph

- Absorbing nodes
- Source nodes
- Cycles

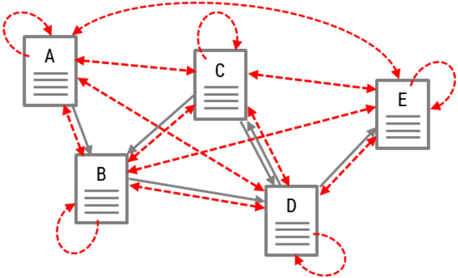


$$\mathbf{r}^{t+1} = \mathbf{A}^T \mathbf{r}^t, \quad \mathbf{r}^{t=0} = \mathbf{r_0}$$

PageRank

"PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The **probability** that the random surfer visits a page is its **PageRank**."

$$PR(A) = (1 - d) + d(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$



Sergey Brin and Larry Page, 1998

Random walk

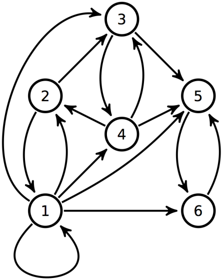
- Random walk on a directed graph

$$p_i^{t+1} = \sum_{j \in N(i)} \frac{p_j^t}{d_j^{out}} = \sum_j \frac{A_{ji}}{d_j^{out}} p_j$$

$$\mathbf{D}_{ii} = diag\{d_i^{out}\}$$

$$\mathbf{p}^{t+1} = (\mathbf{D}^{-1}\mathbf{A})^T \mathbf{p}^t$$

$$\mathbf{p}^{t+1} = \mathbf{P}^T \mathbf{p}^t$$



- Markov chain with transition probability matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$

$$\lim_{t \to \infty} \mathbf{p}^t = \pi$$

Perron-Frobenius Theorem

Perron-Frobenius theorem (Fundamental Theorem of Markov Chains)
If matrix is

- stochastic (non-negative and rows sum up to one, describes Markov chain)
- irreducible (strongly connected graph)
- aperiodic

then

$$\exists \lim_{t \to \infty} \bar{\mathbf{p}}^t = \bar{\pi}$$

and can be found as a left eigenvector

$$\bar{\pi}\mathbf{P} = \bar{\pi}, \quad \text{where} \quad ||\bar{\pi}||_1 = 1$$

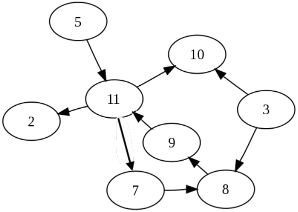
$\bar{\pi}$ - stationary distribution of Markov chain, raw vector

Oscar Perron, 1907, Georg Frobenius,1912.

PageRank

Transition matrix:

$$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$$



Stochastic matrix:

$$\mathbf{P}' = \mathbf{P} + \frac{\mathbf{s}\mathbf{e}^T}{n}$$

PageRank matrix:

$$\mathbf{P}'' = \alpha\mathbf{P}' + (1-\alpha)\frac{\mathbf{e}\mathbf{e}^T}{n}$$

Eigenvalue problem (choose solution with $\lambda = 1$):

$$\mathbf{P}''^T\mathbf{p} = \lambda\mathbf{p}$$

Notations:

e - unit column vector, **s** - absorbing nodes indicator vector (column)

PageRank computations

- Eigenvalue problem ($\lambda = 1$, $||p||_1 = \mathbf{p}^T\mathbf{e} = 1$):

$$\left(\alpha \mathbf{P}' + (1-\alpha)\frac{\mathbf{e}\mathbf{e}^T}{n}\right)^T \mathbf{p} = \lambda \mathbf{p}$$

$$\mathbf{p} = \alpha \mathbf{P}'^T \mathbf{p} + (1-\alpha)\frac{\mathbf{e}}{n}$$

- Power iterations:

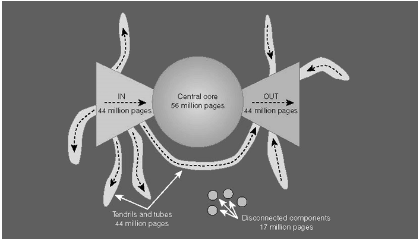
$$\mathbf{p} \leftarrow \alpha \mathbf{P}'^T \mathbf{p} + (1-\alpha)\frac{\mathbf{e}}{n}$$

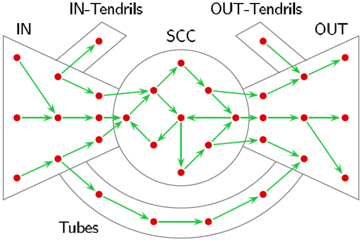
- Sparse linear system:

$$(\mathbf{I} - \alpha \mathbf{P}'^T)\mathbf{p} = (1-\alpha)\frac{\mathbf{e}}{n}$$

Graph structure of the web

Bow tie structure of the web





Andrei Broder et al, 1999

Hubs and Authorities (HITS)

Citation networks. Reviews vs original research (authoritative) papers

- authorities, contain useful information, a_i
- hubs, contains links to authorities, h_i

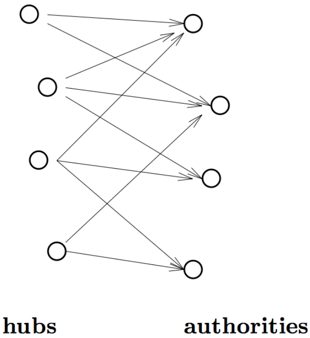
Mutual recursion

- Good authorities reffered by good hubs

$$a_i \leftarrow \sum_j A_{ji} h_j$$

- Good hubs point to good authorities

$$h_i \leftarrow \sum_j A_{ij} a_j$$



Jon Kleinberg, 1999

HITS

System of linear equations

$$\begin{aligned} \mathbf{a} &= \alpha \mathbf{A}^T \mathbf{h} \\ \mathbf{h} &= \beta \mathbf{A} \mathbf{a} \end{aligned}$$

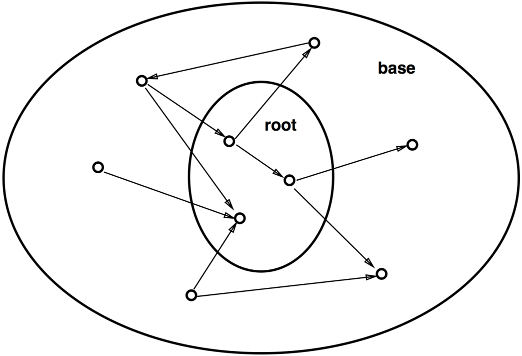
Symmetric eigenvalue problem

$$\begin{aligned} (\mathbf{A}^T \mathbf{A})\mathbf{a} &= \lambda \mathbf{a} \\ (\mathbf{A}\mathbf{A}^T)\mathbf{h} &= \lambda \mathbf{h} \end{aligned}$$

where eigenvalue $\lambda = (\alpha\beta)^{-1}$

HITS

Focused subgraph of WWW



Jon Kleinberg, 1999

PageRank beyond the Web

1. GeneRank
2. ProteinRank
3. FoodRank
4. SportsRank
5. HostRank
6. TrustRank
7. BadRank
8. ObjectRank
9. ItemRank
10. ArticleRank
11. BookRank
12. FutureRank
13. TimedPageRank
14. SocialPageRank
15. DiffusionRank
16. ImpressionRank
17. TweetRank
18. TwitterRank
19. ReversePageRank
20. PageTrust
21. PopRank
22. CiteRank
23. FactRank
24. InvestorRank
25. ImageRank
26. VisualRank
27. QueryRank
28. BookmarkRank
29. StoryRank
30. PerturbationRank
31. ChemicalRank
32. RoadRank
33. PaperRank
34. Etc...

from David Gleich

References

- The Anatomy of a Large-Scale Hypertextual Web Search Engine, Sergey Brin and Lawrence Page, 1998
- Authoritative Sources in a Hyperlinked Environment. Jon M. Kleinberg, Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1999
- Graph structure in the Web, Andrei Broder et all. Procs of the 9th international World Wide Web conference, 2000
- A Survey of Eigenvector Methods of Web Information Retrieval. Amy N. Langville and Carl D. Meyer, 2004
- PageRank beyond the Web. David F. Gleich, arXiv:1407.5107, 2014