Large-Scale Networks

PageRank

Dr Vincent Gramoli | Lecturer
School of Information Technologies

THE UNIVERSITY OF SYDNEY

Introduction

- Last week we talked about:
  - Hubs whose scores depend on the authority of the nodes they point to
  - Authorities whose score depends on the hub score of the nodes pointing to them

- Today, we will see that
  - Same nodes can play both the roles of hubs and authorities
  - Nodes can play an important endorsement role without being heavily endorsed

PageRank

Basic definition

- Hub an authorities indicate multiple roles that same pages can play
- Pages can play an important endorsement role without being heavily endorsed

- Competitor companies may not endorse each other
- But most of the time prominent pages are endorsed by many others:
  - Academics
  - Government pages
  - Bloggers
  - Personal pages
  - Scientific literature

Basic definition

- This form of endorsement is at the heart of the PageRank [BP98]
  - Votes and repeated improvements are used to determine the PageRank of a page
  - Endorsement is passed through outgoing links with a weight that corresponds to the current PageRank of the source page

The PageRank can be viewed as a *fluid* circulating through the network and pooling at the nodes that are the most important

Basic definition

› PageRank is computed as follows:

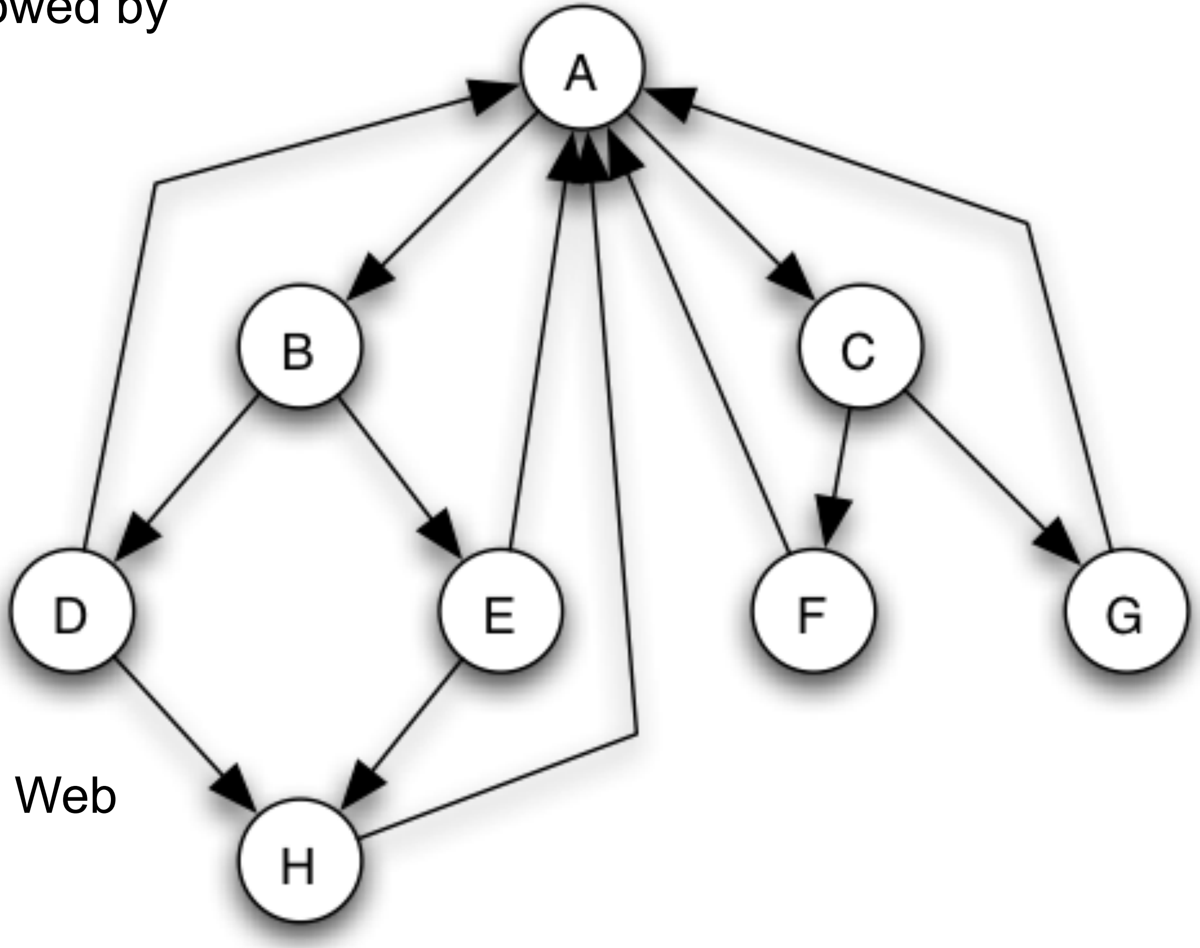
- In a network with n nodes, we assign all nodes the same initial PageRank, set to be 1/n.
- We choose a number of steps k.
- We then perform a sequence of k updates to the PageRank values, using the following rule for each update:

  Basic PageRank Update Rule: Each page divides its current PageRank equally across its outgoing links, and passes these equal shares to the pages it points to. (If a page has no outgoing links, it passes all its current PageRank to itself.) Each page updates its new PageRank to be the sum of the shares it receives.

› Since each page divides its PageRank among its outgoing link, there is no need to normalize it, the total PageRank in the network is constant

Basic definition

- A collection of eight pages: A has the largest PageRank, followed by B and C (which collect endorsements from A).



- Let's consider how this computation works on the collection of the previous 8 Web pages.

Example

- All pages start out with a PageRank of 1/8 and their PageRank values after the first two updates are given by the following table:

Step	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	3/16	1/4	1/4	1/32	1/32	1/32	1/32	1/16

- For example, A gets PageRank of ½ after the first update because it gets all of F's and G's, H's PageRank, and half each of D's and E's. On the other hand, B and C each get half of A's PageRank, so they only get 1/16 each in the first step.

- This is in keeping with the principle of repeated improvement: after the first update causes us to estimate that A is an important page, we weigh its endorsement more highly in the next update.

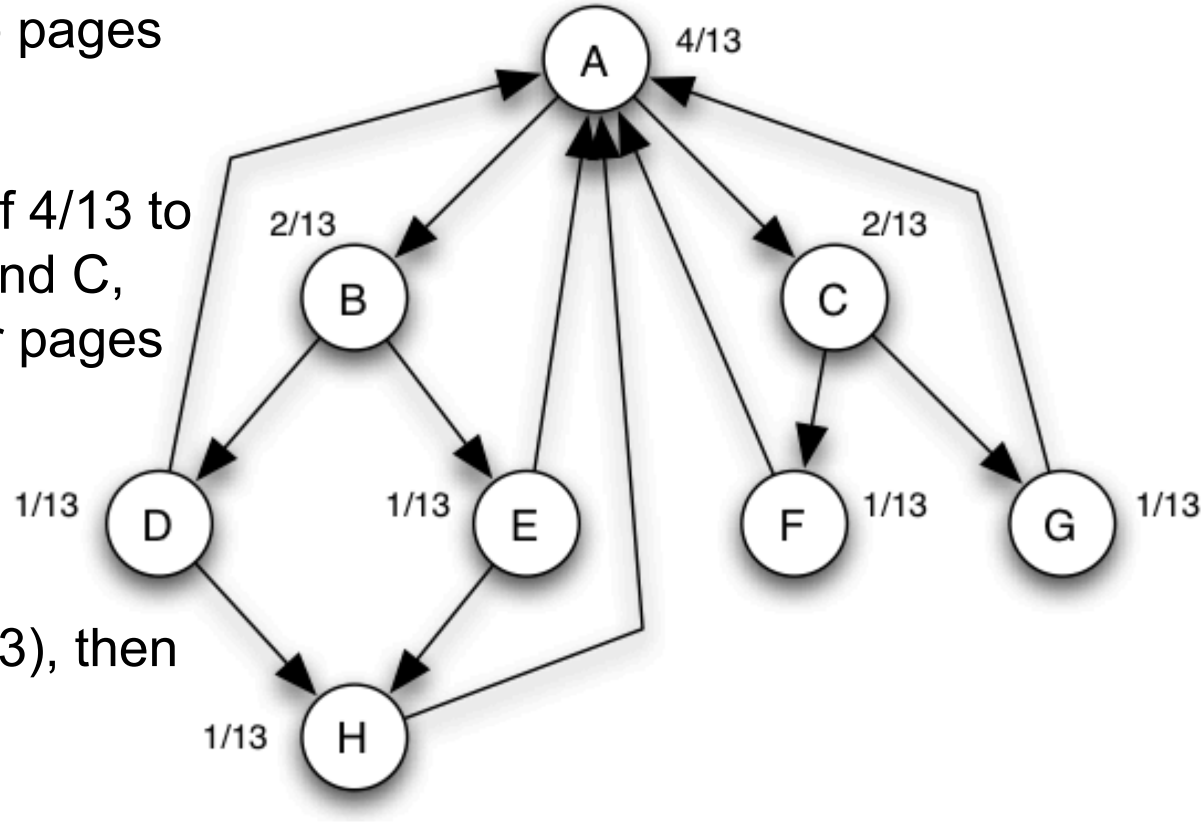
Equilibrium values of PageRank

- As with Hub and Authority and under reasonable assumptions the PageRank values of all nodes converge to limiting values as the number of update steps, k, goes to infinity

- It is easy to check that a state is an equilibrium if applying the Basic PageRank Update Rule does not update anything

Basic definition

- Equilibrium PageRank values for the network of eight Web pages
- Assigning a PageRank of 4/13 to page A, 2/13 to each B and C, and 1/13 to the five other pages achieves the equilibrium
- If the network is strongly connected (cf. Chapter 13), then there is a unique set of equilibrium values



Limitation of the basic definition

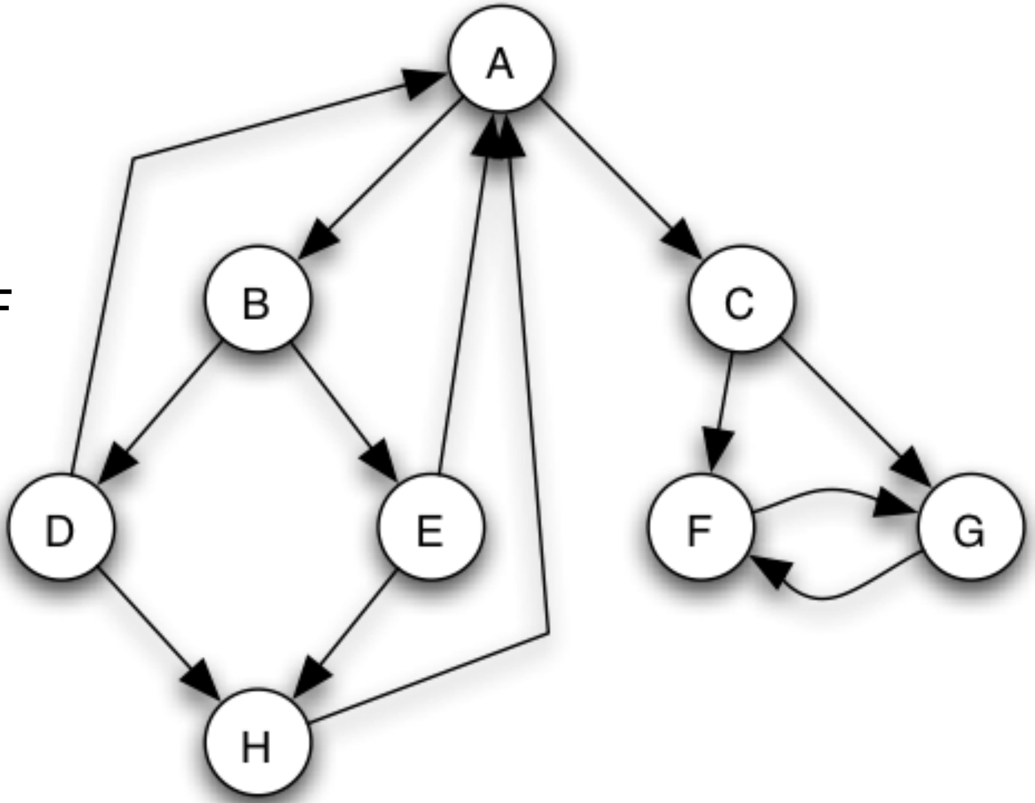
- In many networks the wrong nodes can end up with all the PageRank

- As long as there are a small set of nodes that can be reached from the network but do not have any path back to the network, then PageRank will accumulate there

- This is a problem given the bow-tie structure of the Web (cf. Chapter 13)
  - There is one giant strongly connected component (SCC)
  - Many slow leaks out of the SCC
  - All the PageRank would accumulate at the end of the downstream nodes

Limitation of the basic definition

- The same collection of eight pages, but F and G have changes their links to point to each other instead of A. Without a smoothing effect, all the PageRank would go to F and G.
- PageRank that flows from C to F and G can never circulate back into the rest of the network
- There is a kind of slow leak that causes all the PageRank to end up at F and G. We converge to ½ for F and G and 0 for others.



Scaling the definition of PageRank

- We can solve this problem similarly to the observation that rain water does not converge to the same lowest points due to a counterbalancing process of evaporation and rains on the highest points

- The idea is to pick a scaling factor s strictly between 0 and 1 and replace the Basic PageRank Update Rule by the following:

  Scaled PageRank Update Rule: First apply the Basic PageRank Update Rule. Then scale down all PageRank values by a factor of s. This means that the total PageRank in the network has shrunk from 1 to s. We divide the residual 1-s units of PageRank equally over all nodes, giving (1-s)/n to each.

Scaling the definition of PageRank

- We can solve this problem similarly to the observation that rain water does not converge to the same lowest points due to a counterbalancing process of evaporation and rains on the highest points

- The idea is to pick a scaling factor s strictly between 0 and 1 and replace the Basic PageRank Update Rule by the following:

  Scaled PageRank Update Rule: First apply the Basic PageRank Update Rule. Then scale down all PageRank values by a factor of s. This means that the total PageRank in the network has shrunk from 1 to s. We divide the residual 1-s units of PageRank equally over all nodes, giving (1-s)/n to each.

- This rule also preserves the PageRank of the network since it is based on redistribution according to a water cycle that evaporates 1-s units of PageRank in each step and rains it down uniformly across all nodes.

The limit of the scaled PageRank update rule

- Repeatedly applying the Scaled PageRank Update Rule converges to a set of limiting PageRank values as the number of updates, k, goes to infinity

- These limiting values form the unique equilibrium for the Scaled PageRank Update Rule

- But, these values depend on the value of s
  - There are different update rules for each value of s
  - In practice, PageRank uses this rule with a scaling factor s between 0.8 and 0.9

- The scaling factor makes the PageRank less sensitive to the addition or deletion of small numbers of nodes or links [LM06,ZNJ01]

Random walks

- The PageRank can be equivalently formulated using a *random walk*

- Consider someone who is randomly browsing a network of Web pages.
  - She starts by choosing a page at random, picking each page with same proba.
  - Then, she follows links for a sequence of k steps
  - In each step, she picks a random outgoing link from the current page and follows it to where it leads
  - (if the current page has no outgoing link, she stays where she is)
- This exploration of the network is called a *random walk* on the network

> *Claim:* *The probability of being at page X after k steps of this random walk is precisely the PageRank of X after k applications of the Basic PageRank UpdateRule. (cf. Section 14.6 for the proof.)*

Applying Link Analysis in Modern Web Search

Modern Web search

- This link analysis plays an important role in search engines since the 90's:
  - Google
  - Yahoo!
  - Microsoft's Bing
  - Ask

- Nowadays, link analysis is a bit different:
  - Extension with other analyses
  - More complex
  - Unknown secret ingredients

Modern Web search

- Google's search engine
  - PageRank has always been a core component of Google's search engine
  - The role of PageRank has been claimed to be declining
  - In 2003/2004, the different Hilltop link analysis was introduced [BM01]

- *Anchor text*: the highlighted bits of clickable text that activate a hyperlink
  - Analyzing anchor text helps ranking by combining text and links
  - Clicking on the link associated with "University of Sydney" in a sentence "I am a student of University of Sydney" will likely lead to a page about this university
  - This analysis can be applied to Hubs and Authorities:
    - If the link has highly relevant anchor text while others don't, then
    - We can weight the contributions of the relevant links more heavily than others
    - Example: As we pass PageRank, we multiply it by the quality of the anchor text

Search engine optimization

- Search Engine Optimization (SEO)
  - Updates to Google's ranking function that push off a company from the first screen could spell financial ruin
  - Google's most significant updates where compared to hurricanes (unpredictable damaging act of nature)
  - Guidelines for improving page ranking emerged (SEO)
  - Experts advising companies how to create sites and pages that rank highly

- The perfect ranking function is a moving target
  - If a search engine keeps the same function for too long, then
  - Experts would reverse-engineer the function
  - Function would not be effective
  - Search engine companies have thus to keep their ranking functions secret

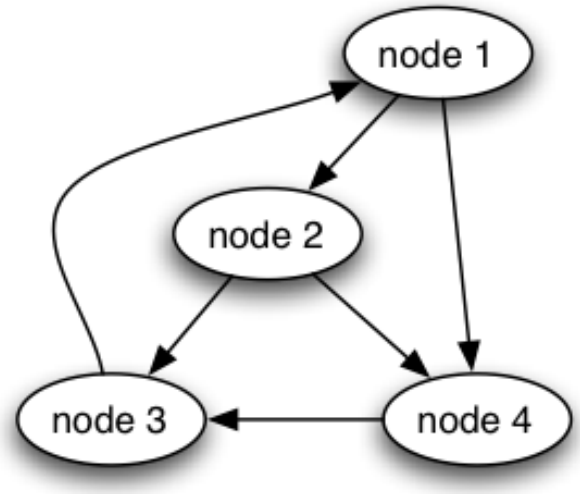
Spectral Analysis of PageRank

Spectral analysis

Let's analyze PageRank using matrix-vector multiplication and eigenvectors

- Under the Basic Update Rule, each node takes its PageRank and divides it equally over all the nodes to which it points

Spectral analysis



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

The flow of PageRank can be represented using the matrix N above

- Nij = 0 if i does not link to j
- Nij is reciprocal of the number of nodes i points to, otherwise
- Nii = 1 if i has no outgoing link (it passes its PageRank to itself)

N is similar to the adjacency matrix except when i points to j

Spectral analysis

Let's represent the PageRank of all nodes using a vector r

› The coordinate ri is the PageRank of node i

› The Basic PageRank Update Rule becomes:

$$ri = N_{1i}r1 + N_{2i}r2 + \ldots + N_{ni}rn. \quad (3)$$

This corresponds to multiplication by the transpose of the matrix, just as we saw for the Authority Update Rule. Thus, equation (3) can be rewritten as:

$$r = N^T r.$$

The Scaled PageRank can be represented in the same way

- In the Scaled Update Rule, the updated PageRank is scaled down by a factor of s and the residual 1-s units are divided equally over all nodes.
- We can defined N'_{ij} to be $sN_{ij} + (1-s)/n$ and then the Scaled Update Rule can be written:

$$ri = N'_{1i}\, r1 + N'_{2i}\, r2 + \ldots + N'_{ni}\, rn.$$

Or equivalently:

$$r = N'^{T} r.$$

Spectral analysis

The Scaled PageRank can be represented in the same way

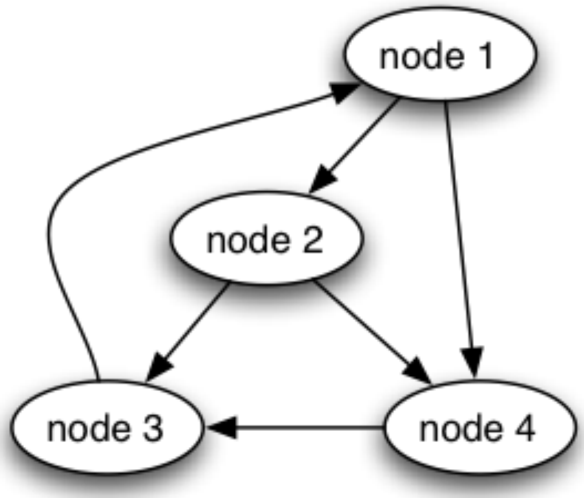
- In the Scaled Update Rule, the updated PageRank is scaled down by a factor of s and the residual 1-s units are divided equally over all nodes.
- We can defined N'_{ij} to be sN_{ij} + (1-s)/n and then the Scaled Update Rule can be written:

$$ri = N'_{1i}\, r1 + N'_{2i}\, r2 + \ldots + N'_{ni}\, rn.$$

Or equivalently:

$$r = N'^{T} r.$$

Spectral analysis



$$\begin{bmatrix} .05 & .45 & .05 & .45 \\ .05 & .05 & .45 & .45 \\ .85 & .05 & .05 & .05 \\ .05 & .05 & .85 & .05 \end{bmatrix}$$

- The flow of PageRank under the Scaled PageRank Update Rule
- Representation with matrix N’ with scaling factor s = 0.8
- The entry N'_{ij} specifies the portion of i’s PageRank that should be passed to j in one update

Spectral analysis

Repeated Improvement using the Scaled PageRank Update Rule

- As we apply the rule to an initial vector $r^{\langle 0 \rangle}$, we produce a sequence of vector $r^{\langle 1 \rangle}$, $r^{\langle 2 \rangle}$, … where each vector is obtained from the preceding one via multiplication by N'^T.

$$r^{\langle k \rangle} = (N'^T)^k r^{\langle 0 \rangle}.$$

- Since PageRank is conserved as it is updated, we don't need to normalize it
- If the Scaled PageRank tends to a limiting vector $r^{\langle * \rangle}$, then this limit should satisfy:

$$N'^T r^{\langle * \rangle} = r^{\langle * \rangle}.$$

Convergence of the Scaled PageRank Update Rule

- Matrices involved are not symmetric (as opposed to MM^T and M^TM)
- N' is a *positive matrix* (i.e., every N_{ij} is positive)
- So we can apply Perron's Theorem [LM06], hence:
  - P has a real eigenvalue c > 0 such that c > c' for all other eigenvalues c'.
  - There is an eigenvector y with positive real coordinates corresponding to the largest eigenvalue c, and y is unique up to multiplication by a constant
  - If the largest eigenvalue c is equal to 1, then, for any starting positive vector x ≠ 0 the sequence of vectors $P^k x$ converges to a vector in the direction of y as k goes to infinity
- Applying the Scaled PageRank Update Rule from any starting point converges to a unique vector y

Conclusion

Conclusion

- Link analysis of networks is important
  - Relies on
    - Hubs and authorities
    - Repeated improvement technique
    - Additional weights
  - Allows to rank pages, journals, cases… (any information that is networked)

- Search engine is probably the mostly used application
  - Some adjustments are necessary for some network structures (scale factor)
  - Search engines use link analysis but also other techniques

References

- [BP98] Sergey Brin and Lawrence Page – The anatomy of large-scale hypertextual Web search engine. Proc. Of 7th Intl World Wide Web Conference, p.107-117, 1998.

- [BM01] Krishna, Mihaila. When experts agree: Using non-affiliated experts to rank popular topics. *Proc. Of the 10th Int'l World Wide Web Conference*, p. 597-602, 2001.

- [LM06] Langville, Meyer. Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, 2006.

- [ZNJ01] Zheng, Ng, Jordan. Stable Algorithms for Link Analysis. *Proc. of 24th ACM SIGIR Conference on Research and Development in information Retrieval*, p.258-266, 2001.