SOCIAL NETWORK ANALYTICS

Page Ranking

Prakash C O

Department of Computer Science and Engineering

SOCIAL NETWORK ANALYTICS

Page Ranking

Prakash C O
Department of Computer Science and Engineering

Page Ranking

- The heart of Google's searching software is PageRank, a system for ranking web pages developed by Google founders(Larry Page and Sergey Brin).
- PageRank is a way of measuring the importance of webpages from the hyperlink network structure.
- PageRank assigns a *score of importance* to each node/page.
  Important Assumption: More important nodes/pages are likely to receive more in-links from other important nodes/pages.
- PageRank can be used for any type of network, but it is mainly useful for directed networks.
- A node's PageRank depends on the PageRank of other nodes. (Circular definition?).

Sergey Brin and Larry Page

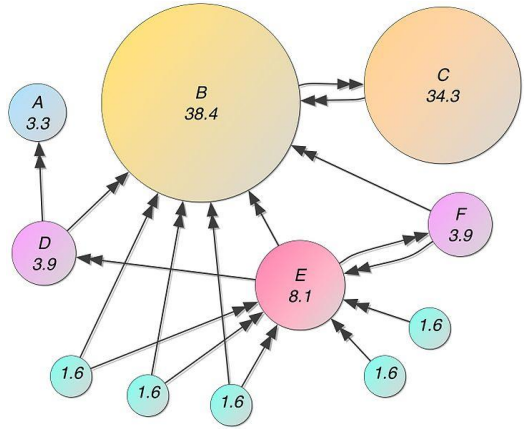
SOCIAL NETWORK ANALYTICS

Page Ranking

- Pagerank concept is developed to rank pages in the web.
- In general, when page A links to page B, this means
  - A's author thinks that B's content is interesting or important
  - So a link from A to B, adds to B's reputation

But not all links are equal..

- If A is very important, then A → B "counts more"
- If A is not important, then A → B "counts less"

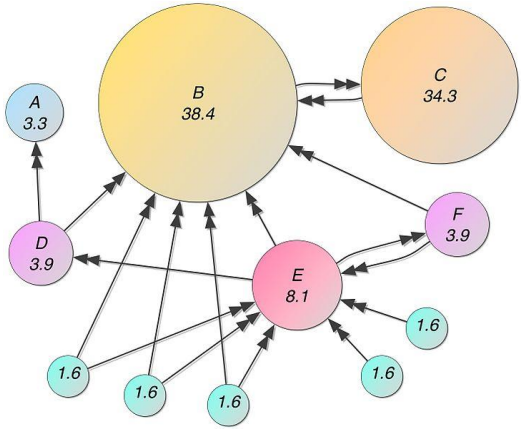


Source:https://en.wikipedia.org/wiki/File:PageRanks-Example.jpg

Page Ranking

➢ **The basic definition of PageRank.**

- Intuitively, we can think of PageRank as a kind of "fluid" that circulates through the network, passing from node to node across edges, and pooling at the nodes that are most important.



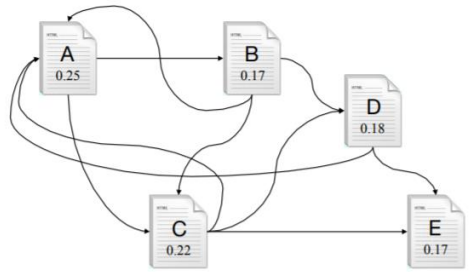
Source:https://en.wikipedia.org/wiki/File:PageRanks-Example.jpg

Page Ranking

➢The PageRank is computed as follows.

1. In a network with n nodes, we assign all nodes the same initial PageRank, set to be 1 or 1/n.
2. We choose a number of steps k.
3. 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 out-going links, and passes these equal shares to the pages it points to. (If a page has no out-going links, it passes all its current PageRank to itself.)
   - Each page updates its new PageRank to be the sum of the shares it receives.

Page Ranking



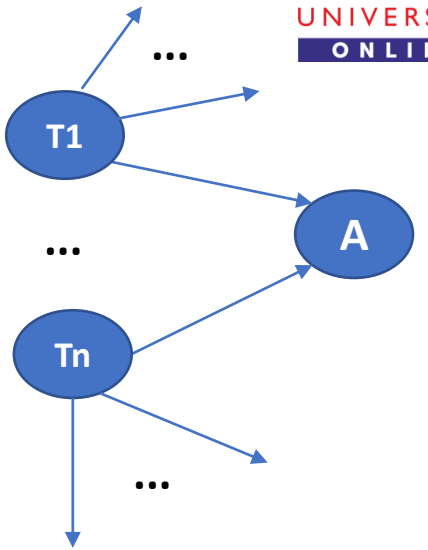
- Notice that the total PageRank in the network will remain constant as we apply PageRank update rule steps.
  Since each page takes its PageRank, divides it up, and passes it along links, PageRank is never created nor destroyed, just moved around from one node to another.

  As a result, we don't need to do any normalizing of the numbers to prevent them from growing.

Page Ranking

- Contrary to the concept of **link popularity**, **PageRank is not simply based upon the total number of inbound links**.
- **The basic approach of PageRank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally.**
- First of all, **a document ranks high in terms of PageRank, if other high-ranking documents link to it**.

PageRank Algorithm



- The original PageRank algorithm was described by Larry Page and Sergey Brin in several publications.
- It is given by

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

  where
  - PR(A) is the PageRank of page A,
  - PR(Ti) is the PageRank of pages Ti which link to page A,
  - C(Ti) is the number of outbound links on page Ti and
  - d is a damping factor which can be set between 0 and 1.
- The PageRank does not rank web sites as a whole but is determined for each page individually.
- Further, the PageRank of page A is recursively defined by the PageRanks of those pages which link to page A.

PageRank Algorithm

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

- The PageRank of pages Ti which link to page A does not influence the PageRank of page A uniformly.
- Within the PageRank algorithm, the PageRank of a page T is always weighted by the number of outbound links C(T) on page T.

  This means that the more outbound links a page T has, the less will page A benefit from a link to it on page T.

PageRank Algorithm

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

- The weighted PageRank of pages Ti is then added up.

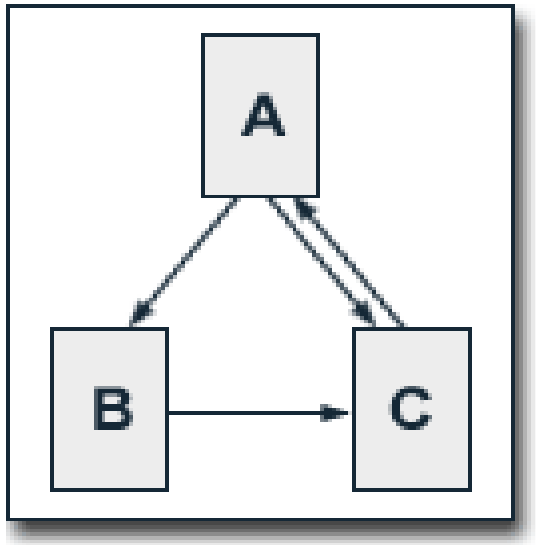
  The outcome of this is that an additional inbound link for page A will always increase page A's PageRank.

- Finally, the sum of the weighted PageRanks of all pages Ti is multiplied with a damping factor **d** which can be set between 0 and 1.

  Thereby, the extend of PageRank benefit for a page by another page linking to it is reduced.

PageRank Algorithm

➢ The characteristics of PageRank shall be illustrated by a small example.

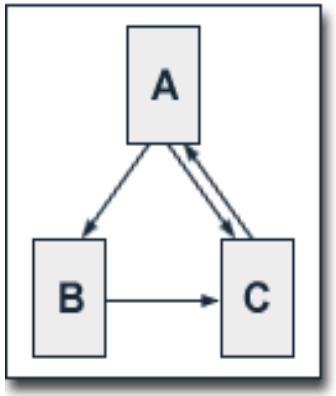


- Consider a small web consisting of three pages A, B and C, whereby page A links to the pages B and C, page B links to page C and page C links to page A.
- According to Page and Brin, the damping factor d is usually set to 0.85, but to keep the calculation simple we set it to 0.5.

Page Ranking

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

➢ So, we get the following equations for the PageRank calculation:



- PR(A) = 0.5 + 0.5 PR(C)

  PR(B) = 0.5 + 0.5 (PR(A) / 2)

  PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

  These equations can easily be solved. We get the following PageRank values:

  PR(A) = 14/13 = 1.07692308

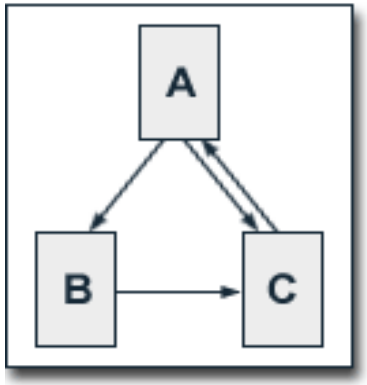
  PR(B) = 10/13 = 0.76923077

  PR(C) = 15/13 = 1.15384615

- In practice, the web consists of billions of documents and it is not possible to find a solution by inspection.

The Iterative Computation of PageRank

- Because of the size of the actual web, the Google search engine uses an approximative, iterative computation of PageRank values.
- The iterative calculation shall again be illustrated by the example, whereby each page is assigned a starting PageRank value of 1.



The Iterative Computation of PageRank

PR(A) = 0.5 + 0.5 PR(C)

PR(B) = 0.5 + 0.5 (PR(A) / 2)

PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

Iteration 1:

PR(A) = 0.5 + 0.5 * 1 = **1**

PR(B) = 0.5 + 0.5 (**1** /2) = **0.75**

PR(C) = 0.5 + 0.5 (**1** /2 + **0.75**) = **1.125**

Iteration 2:

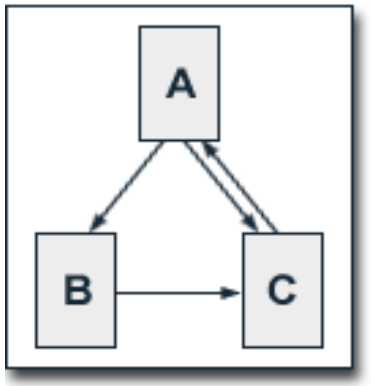
PR(A) = 0.5 + 0.5 * **1.125** = **1.0625**

PR(B) = 0.5 + 0.5 (**1.0625**/2) = **0.765625**

PR(C) = 0.5 + 0.5 (**1.0625** /2 + **0.765625**)

=**1.1484375**

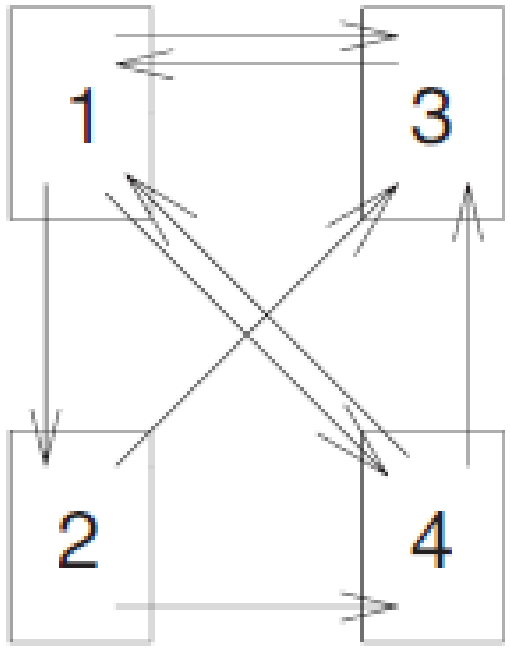
Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615



Page Ranking

➢ **Exercise: Compute PageRank for the following web pages**

(Given d=0.5 and Initial PageRank=1)



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

PR(1) =

PR(2) =

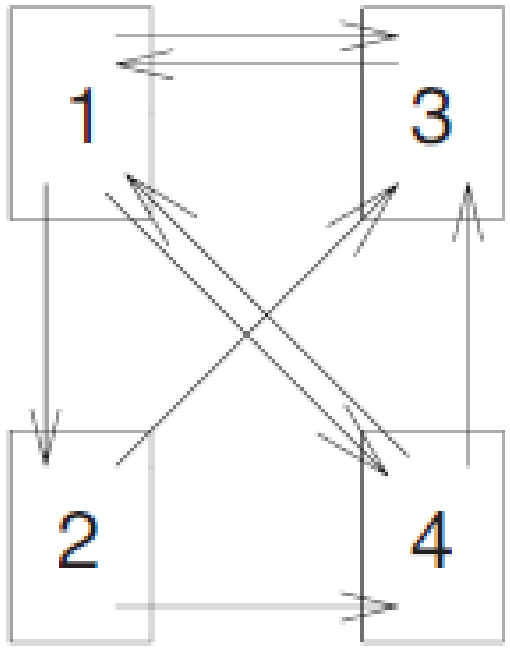
PR(3) =

PR(4) =

Page Ranking

➢ **Exercise: Compute PageRank for the following web pages**

(Given d=0.5 and Initial PageRank=1)



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

PR(1) =0.5 + 0.5 (PR(3)/1 + PR(4)/2)

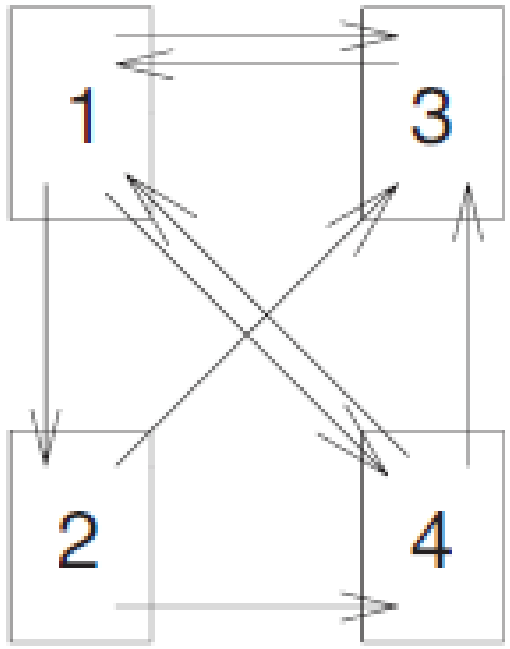
PR(2) =0.5 + 0.5 (PR(1)/3)

PR(3) =0.5 + 0.5 (PR(1)/3 + PR(2)/2+ PR(4)/2)

PR(4) =0.5 + 0.5 (PR(1)/3 + PR(2)/2)

➢ **Exercise: Compute PageRank for the following web pages**

(Given d=0.5 and Initial PageRank=1)



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

PR(1) =0.5 + 0.5 (PR(3)/1 + PR(4)/2)

PR(2) =0.5 + 0.5 (PR(1)/3)

PR(3) =0.5 + 0.5 (PR(1)/3 + PR(2)/2+ PR(4)/2)

PR(4) =0.5 + 0.5 (PR(1)/3 + PR(2)/2)

It.	PR(1)	PR(2)	PR(3)	PR(4)
0	1	1	1	1
1	1.25	0.70833	1.13541	0.88541
2				
3				

SOCIAL NETWORK ANALYTICS

Page Ranking

The Effect of Inbound Links
The Influence of the Damping Factor
The Effect of Outbound Links
The Effect of the Number of Pages

Prakash C O
Department of Computer Science and Engineering

Page Ranking - The Effect of Inbound Links

- **Each additional inbound link for a web page always increases that page's PageRank.**
- Taking a look at the PageRank algorithm, which is given by

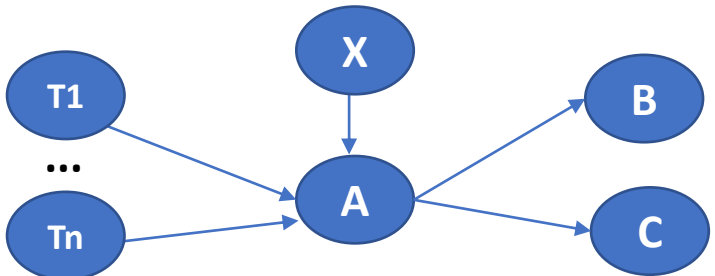
$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

An additional inbound link from page X increases the PageRank of page A by

$$d \times (PR(X)/C(X))$$

- But page A usually links to other pages itself. Thus, these pages get a PageRank benefit also. If these pages link back to page A, page A will have an even higher PageRank benefit from its additional inbound link.

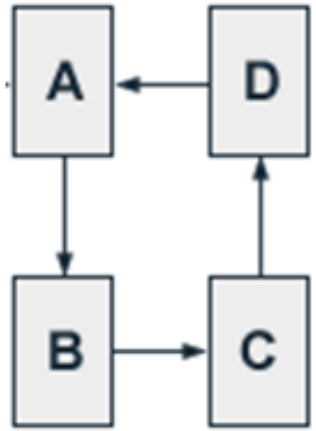
$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn)) + d \times (PR(X)/C(X))$$



Page Ranking - The Effect of Inbound Links

Example:

- Consider a website consisting of 4-pages A, B, C and D which are linked to each other in circle.



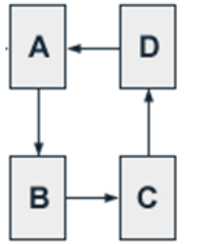
- Initially: PR(A) = 1, PR(B) = 1, PR(C) = 1, PR(D) = 1

- Without external inbound links to one of these pages, each of them obviously has a PageRank of 1.

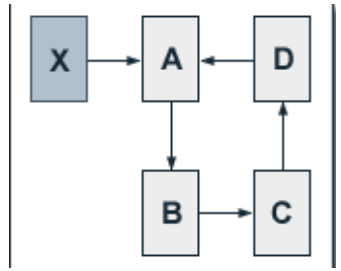
```
PR(A) = 0.5 + 0.5*PR(D) = 0.5 + 0.5*(1) = 1
PR(B) = 0.5 + 0.5*PR(A) = 0.5 + 0.5*(1) = 1
PR(C) = 0.5 + 0.5*PR(B) = 0.5 + 0.5*(1) = 1
PR(D) = 0.5 + 0.5*PR(C) = 0.5 + 0.5*(1) = 1
```

Page Ranking - The Effect of Inbound Links

Example:



- We now add a page X to our example, for which we presume **a constant Pagerank PR(X) of 10**. Further, **page X links to page A by its only outbound link.**



- **Setting d=0.5, we get the following equations**

```
PR(A) = 0.5 + 0.5 (PR(X) + PR(D)) = 5.5 + 0.5 PR(D)
PR(B) = 0.5 + 0.5 PR(A)
PR(C) = 0.5 + 0.5 PR(B)
PR(D) = 0.5 + 0.5 PR(C)
```

- Solving them gives us the following PageRank values:

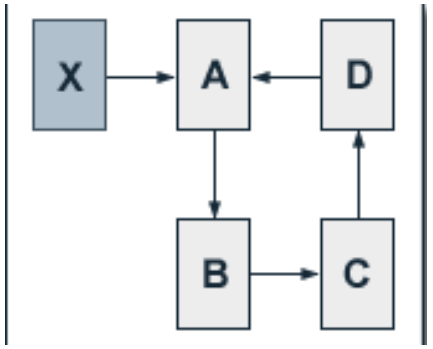
```
PR(A) = 19/3 = 6.33
PR(B) = 11/3 = 3.67
PR(C) = 7/3 = 2.33
PR(D) = 5/3 = 1.67
```

Increase in PageRank values

Each additional inbound link for a web page always increases that page's PageRank

Page Ranking - The Influence of the Damping Factor

- **The degree of PageRank propagation from one page to another by a link is primarily determined by the damping factor d.**



- If we set **d=0.75** we get the following equations for our example:

```
PR(A) = 0.25 + 0.75 (PR(X) + PR(D)) = 7.75 + 0.75 PR(D)
PR(B) = 0.25 + 0.75 PR(A)
PR(C) = 0.25 + 0.75 PR(B)
PR(D) = 0.25 + 0.75 PR(C)
```

- Solving these equations gives us the following PageRank values:

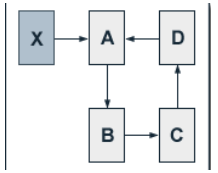
```
PR(A) = 419/35 = 11.97
PR(B) = 323/35 = 9.23
PR(C) = 251/35 = 7.17
PR(D) = 197/35 = 5.63
```

```
When d=0.5
PR(A) = 19/3 = 6.33
PR(B) = 11/3 = 3.67
PR(C) = 7/3 = 2.33
PR(D) = 5/3 = 1.67
```

Page Ranking - The Influence of the Damping Factor

- We see that there is a significantly higher initial effect of additional inbound link for page A which is given by

$$d \times PR(X)/C(X) = 0.75 \times (10/1) = 7.5$$



- This initial effect is then propagated even stronger by the links on site.
- **In this way, the PageRank of page A is almost twice as high at a damping factor of 0.75 than it is at a damping factor of 0.5.**
- At d=0.5 the PageRank of page A is almost 4-times superior to the PageRank of page D, while at d=0.75 it is only a little more than twice as high.
- **So, the higher the damping factor,**
  - **the larger is the effect of an additional inbound link for the PageRank of the page that receives the link** and
  - **the more evenly distributes PageRank over the other pages of a site.**

When d=0.5

PR(A) = 19/3 = 6.33
PR(B) = 11/3 = **3.67**
PR(C) = 7/3 = **2.33**
PR(D) = 5/3 = 1.67

14

When d=0.75

PR(A) = 419/35 = 11.97
PR(B) = 323/35 = **9.23**
PR(C) = 251/35 = **7.17**
PR(D) = 197/35 = 5.63

34

Page Ranking - The Influence of the Damping Factor

➢ **At d=0.5**, the accumulated PageRank of all pages of our site is given by

PR(A) + PR(B) + PR(C) + PR(D) = 14

- Hence, by a page with a PageRank of 10 linking to one page of our example site by its only outbound link, the accumulated PageRank of all pages of the site is increased by 10. (i.e., 4 to 14).

When d=0.5

PR(A) = 19/3 = 6.33
PR(B) = 11/3 = 3.67
PR(C) = 7/3 = 2.33
PR(D) = 5/3 = 1.67

14

➢ **At d=0.75,** the accumulated PageRank of all pages of the site is given by

PR(A) + PR(B) + PR(C) + PR(D) = 34

This time the accumulated PageRank increases by 30 (i.e., 4 to 34).

When d=0.75

PR(A) = 419/35 = 11.97
PR(B) = 323/35 = 9.23
PR(C) = 251/35 = 7.17
PR(D) = 197/35 = 5.63

34

➢ The accumulated PageRank of all pages of a site always increases by

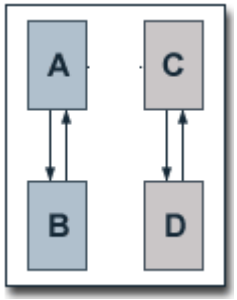
$(d/(1-d)) \times (PR(X)/C(X))$

when d=0.5, $0.5/(1-0.5) \times (10/1) = 10$

when d=0.75, $0.75/(1-0.75) \times (10/1) = 30$

Page Ranking - The Effect of Outbound Links

- Consider a web consisting of two websites, each having two web pages. One site consists of pages A and B, the other consists of pages C and D.
  Initially, both pages of each site solely link to each other.
  It is obvious that each page then has a PageRank of one.
  The total PageRank for both sites is 4.

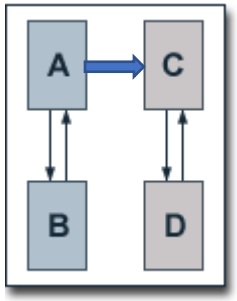


PR(A) = 0.25 + 0.75 PR(B) = 1
PR(B) = 0.25 + 0.75 PR(A) = 1
2

PR(C) = 0.25 + 0.75 PR(D) = 1
PR(D) = 0.25 + 0.75 PR(C) = 1
2

4

- Now we **add a link which points from page A to page C.**

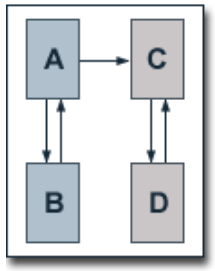


- At d=0.75, we therefore get the following equations:

$$PR(A) = 0.25 + 0.75\ PR(B)$$
$$PR(B) = 0.25 + 0.75\ PR(A)/2$$
$$PR(C) = 0.25 + 0.75\ PR(D) + 0.75\ PR(A)/2$$
$$PR(D) = 0.25 + 0.75\ PR(C)$$

Page Ranking - The Effect of Outbound Links

PR(A) = 0.25 + 0.75 PR(B)
PR(B) = 0.25 + 0.75 PR(A)/2
PR(C) = 0.25 + 0.75 PR(D) + 0.75 PR(A)/2
PR(D) = 0.25 + 0.75 PR(C)



- Solving the equations gives us the following PageRank values

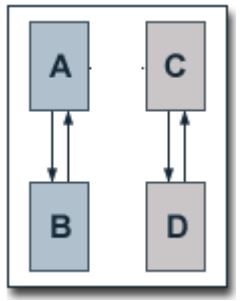
PR(A) = 14/23
PR(B) = 11/23
25/23

PR(C) = 35/23
PR(D) = 32/23
67/23

92/23 = 4

We therefore get an accumulated PageRank of 25/23 for the first site and 67/23 for the second site.

- The total PageRank for both sites is 92/23 = 4. Hence, **adding a link has no effect on the total PageRank of the web**. Additionally, **the PageRank benefit for one site equals the PageRank loss of the other.**



PR(A) = 0.25 + 0.75 PR(B) = 1
PR(B) = 0.25 + 0.75 PR(A) = 1
2

PR(C) = 0.25 + 0.75 PR(D) = 1
PR(D) = 0.25 + 0.75 PR(C) = 1
2

4

Page Ranking - The Effect of the Number of Pages

- Consider a web site consisting of three pages A, B and C, which are joined by an additional page D on the hierarchically lower level of the site.
  A link from page X which has no other outbound links and a PageRank of 10 points to page A.

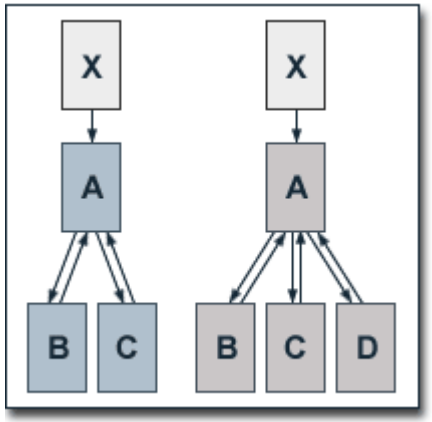
- At d=0.75, the equations before adding page D are given by

  PR(A) = 0.25 + 0.75 (10 + PR(B) + PR(C))

  PR(B) = PR(C) = 0.25 + 0.75 (**PR(A) / 2**)

- Solving the equations gives us the following PageRank values:

  PR(A) = 260/14 PR(B) = 101/14 PR(C) = 101/14



PR(A) = 260/14
PR(B) = 101/14
PR(C) = 101/14

462/14 =33

Page Ranking - The Effect of the Number of Pages

- After adding page D, the equations for the pages' PageRank values are given by

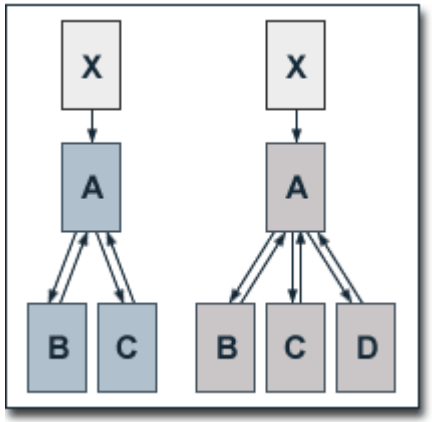
  PR(A) = 0.25 + 0.75 (10 + PR(B) + PR(C) + **PR(D)**)

  PR(B) = PR(C) = PR(D) = 0.25 + 0.75 (**PR(A) / 3**)

- Solving these equations gives us the following PageRank values:

  PR(A) = 266/14 PR(B) = 70/14 PR(C) = 70/14 PR(D) = 70/14

- Our example site has no outbound links, after adding page D, the **accumulated PageRank of all pages** increases by one from 33 (462/14) to 34 (476/14) .
  Further, **the PageRank of page A rises marginally. In contrast, the PageRank of pages B and C depletes substantially.**



PR(A) = 260/14
PR(B) = 101/14
PR(C) = 101/14
462/14 =33

PR(A) = 266/14
PR(B) = 70/14
PR(C) = 70/14
PR(D) = 70/14
476/14 =34

Page Ranking - Use-cases

- **Personalized PageRank is used by Twitter to present users with recommendations of other accounts that they may wish to follow.**
  The algorithm is run over a graph which contains shared interests and common connections.
  Their approach is described in more detail in "WTF: The Who to Follow Service at Twitter".

- **PageRank can be used as part of an anomaly or fraud detection system in the healthcare and insurance industries.**
  It can help find doctors or providers that are behaving in an unusual manner, and then feed the score into a machine learning algorithm.

Page Ranking - Use-cases

- **PageRank has been used to rank public spaces or streets, predicting traffic flow and human movement in these areas.**
  The algorithm is run over a graph which contains intersections connected by roads, where the PageRank score reflects the tendency of people to park, or end their journey, on each street.
  This is described in more detail in "Self-organized Natural Roads for Predicting Traffic Flow: A Sensitivity Study".

- There are many more use cases, which you can read about in David Gleich's "PageRank beyond the web".

Page Ranking

Constraints - when not to use the PageRank algorithm

- There are some things to be aware of when using the PageRank algorithm:
  1. If there are no links from within a group of pages to outside of the group, then the group is considered a spider trap.
  2. Rank sink can occur when a network of pages form an infinite cycle.
  3. Dead-ends occur when pages have no out-links. If a page contains a link to another page which has no out-links, the link would be known as a dangling link.

- If you see unexpected results from running the algorithm, it is worth doing some exploratory analysis of the graph to see if any of these problems are the cause.

Page Ranking

PageRank in social networks

➢PageRank serves three purposes in a social network.

- **First**, it can help **solve link prediction problems to find individuals that will become friends soon**.

- **Second**, it serves a classic role in evaluating the centrality of the people involved **to estimate their social status and power**.

- **Third**, it helps evaluate the potential influence of a node on the opinions of the network

Page Ranking

- **PageRank Computation demo**

```
>>> import networkx as nx
>>> G=nx.barabasi_albert_graph(60,41)
>>> pr=nx.pagerank(G,0.4)
>>> pr
```

Page Ranking - Assignment

PES UNIVERSITY ONLINE

Additional reading

- The Anatomy of a Large-Scale Hypertextual Web Search Engine - Sergey Brin and Lawrence Page.
- David Gleich's "PageRank beyond the web".
- The Implementation of PageRank in the Google Search Engine.

Page Ranking

Tools:

- https://checkpagerank.net/check-page-rank.php
  - www.pes.edu
  - www.google.com
- https://www.prchecker.info/check_page_rank.php

References

- Social Network Analysis: **Lada Adamic,** University of Michigan.
- Wikipedia – Current Literature
- https://neo4j.com/docs/graph-algorithms/current/algorithms/page-rank/
- http://pr.efactory.de/

THANK YOU

Prakash C O

Department of Computer Science and Engineering

coprakasha@pes.edu

+91 98 8059 1946