The Google Pagerank Algorithm and How It Works

What is Pagerank?

- In short PageRank is a “vote”, by all the other pages on the Web, about how important a page is.
- A link to a page counts as a vote of support
- PR(A) = (1-d) + d(PR(T1)/C(T1) +…+PR(Tn)/C(Tn))

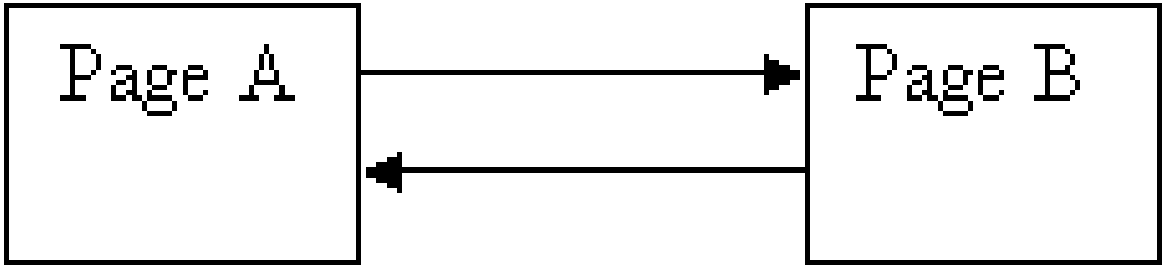
Breaking Down the Equation

- **PR(Tn)** - Each page has a notion of its own self-importance. That's "PR(T1)" for the first page in the web all the way up to "PR(Tn)" for the last page
- **C(Tn)** - Each page spreads its vote out evenly amongst all of it's outgoing links. The count, or number, of outgoing links for page 1 is "C(T1)", "C(Tn)" for page n, and so on for all pages.
- **PR(Tn)/C(Tn)** - so if our page (page A) has a backlink from page "n" the share of the vote page A will get is "PR(Tn)/C(Tn)"
- **d(…** - All these fractions of votes are added together but, to stop the other pages having too much influence, this total vote is "damped down" by multiplying it by 0.85 (the factor "d")
- **(1 - d)** - The (1 – d) bit at the beginning is a bit of probability math magic so the "*sum of all web pages\' PageRanks will be one*": it adds in the bit lost by the **d(…**. It also means that if a page has no links to it (no backlinks) even then it will still get a small PR of 0.15 (i.e. 1 – 0.85). (Aside: the Google paper says "the sum of all pages" but they mean the "the normalised sum" – otherwise known as "the average" to you and me.

How is it Calculated?

- The PR of each page depends on the PR of the pages pointing to it.
- But we won't know what PR those pages have until the pages pointing to them have their PR calculated and so on.
- So what we do is make a guess.

Simple Example



- Each page has one outgoing link. So that means C(A) = 1 and C(B) = 1.

We don’t know what their PR should be to begin with, so we will just guess 1 as a safe random number.

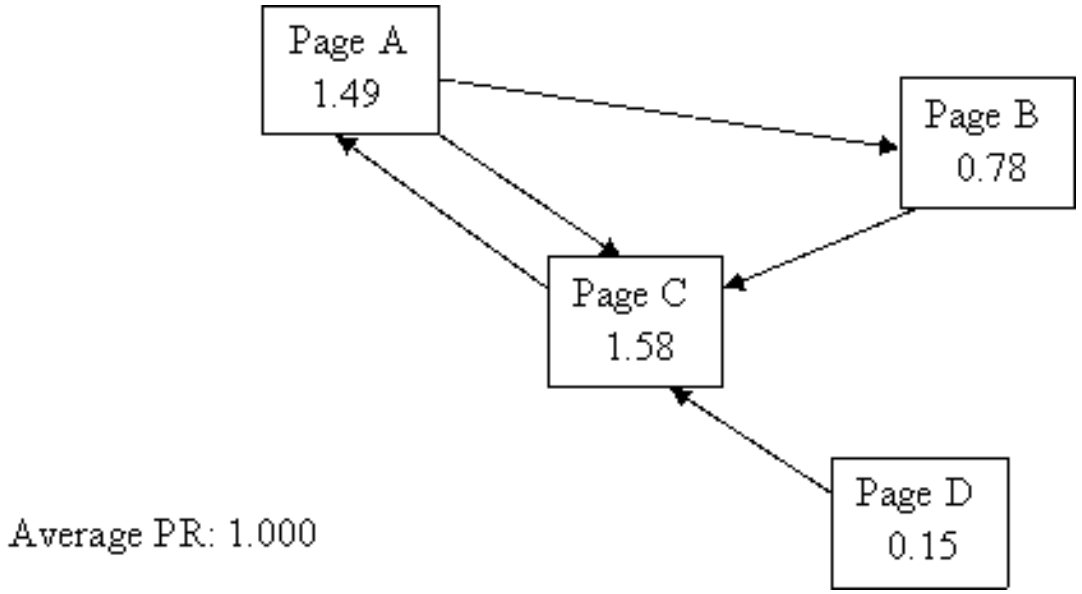
- d (damping factor) = 0.85
- PR(A)= (1 – d) + d(PR(B)/1)PR(B)= (1 – d) + d(PR(A)/1)

  i.e.

- PR(A)= 0.15 + 0.85 * 1
  = 1
- PR(B)= 0.15 + 0.85 * 1
  = 1

Principle

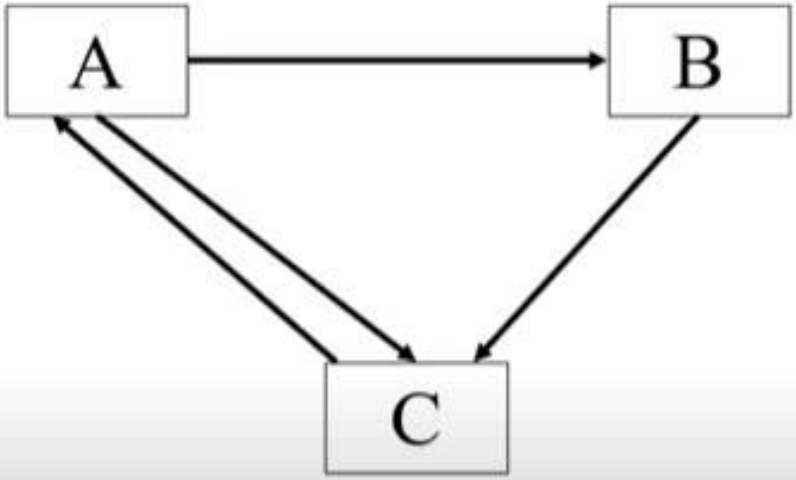
- It doesn’t matter where you start your guess, once the PageRank calculations have settled down, the “normalized probability distribution” (the average PageRank for all pages) will be 1.0



- Observation: every page has at least a PR of 0.15 to share out. But this may only be in theory – there are rumors that Google undergoes a post-spidering phase whereby any pages that have no incoming links at all are completely deleted from the index…

Page Rank Example

- Initially Page Rank (PR) for all the web pages = 1



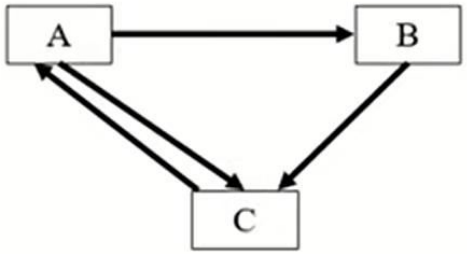
$$PR(A) = (1-d) + d(PR(Ti)/C(Ti) + \ldots + PR(Tn)/C(Tn))$$

Page Rank Example

- Initially Page Rank (PR) for all the web pages = 1

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

$$PR(A) = (1-d) + d\,[\,PR(C) / C(C)\,]$$
$$= (1-0.85) + 0.85\,[\,1/1\,]$$
$$= 0.15 + 0.85[\,1\,]$$
$$= 0.15 + 0.85$$
$$= 1$$

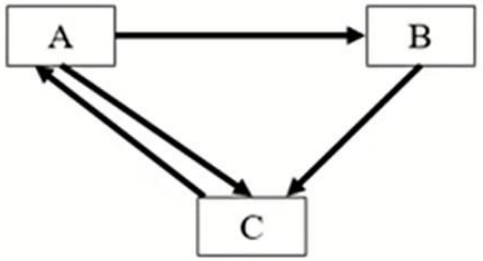


$$PR(B) = (1-d) + d\,[\,PR(A) / C(A)\,]$$
$$= (1-0.85) + 0.85\,[\,(1) / 2\,]$$
$$= 0.15 + 0.85\,[0.5]$$
$$= 0.15 + 0.425$$
$$= 0.575$$

Page Rank Example

- Initially Page Rank (PR) for all the web pages = 1

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

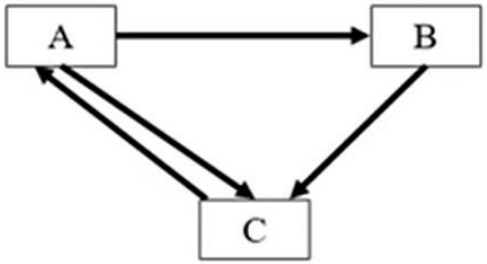


$$PR(C) = (1-d) + d\,[\,PR(A) / C(A) + PR(B) / C(B)\,]$$
$$= (1-0.85) + 0.85\,[\,(1/2) + (0.575 / 1)]$$
$$= 0.15 + 0.85[\,0.5 + 0.575]$$
$$= 0.15 + 0.85\,[\,1.075]$$
$$= 0.15 + 0.91375$$
$$= 1.06375$$

Page Rank Example

- Initially Page Rank (PR) for all the web pages = 1

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



Iteration	A	B	C
0	1	1	1
1	1	0.575	1.06375
2	1.0541875	0.5980296875	1.06354922