



# Εύρεση & Διαχείριση Πληροφορίας στον Παγκόσμιο Ιστό

Διδάσκων –  
Δημήτριος Κατσαρός, Ph.D.

@ Τμ. Μηχανικών Η/Υ, Τηλεπικοινωνιών & Δικτύων  
Πανεπιστήμιο Θεσσαλίας

Διάλεξη 8η: 18/04/2007

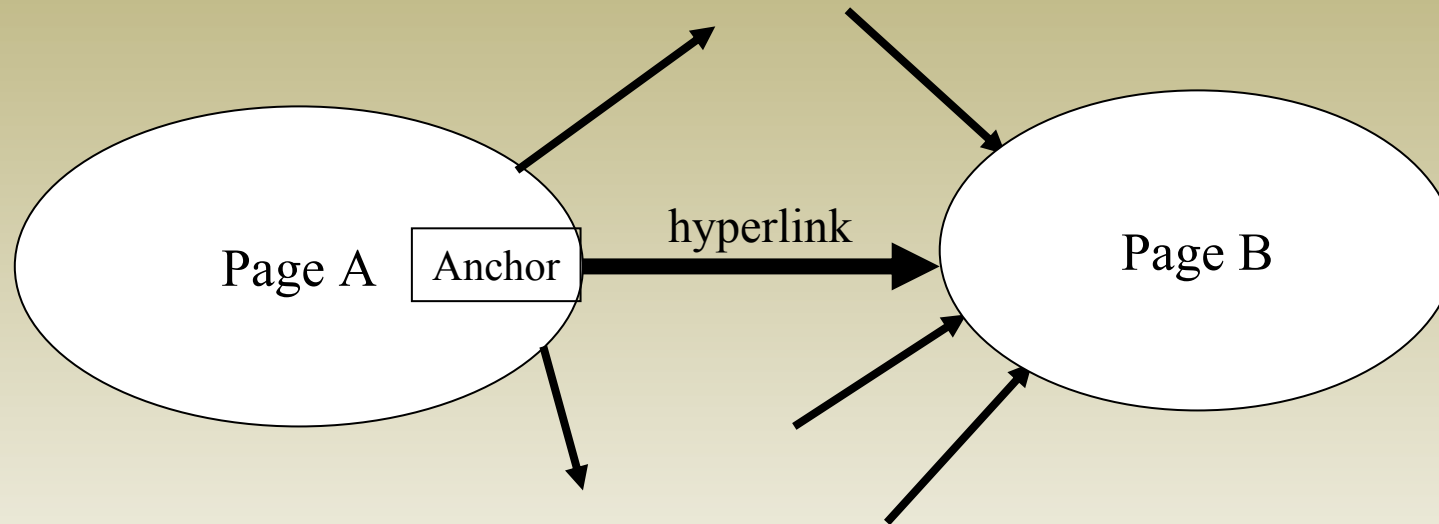


# Ανάλυση υπερσυνδέσμων

Πρακτικές έννοιες του PageRank



# The Web as a Directed Graph



**Assumption 1:** A hyperlink between pages denotes author perceived relevance (quality signal)

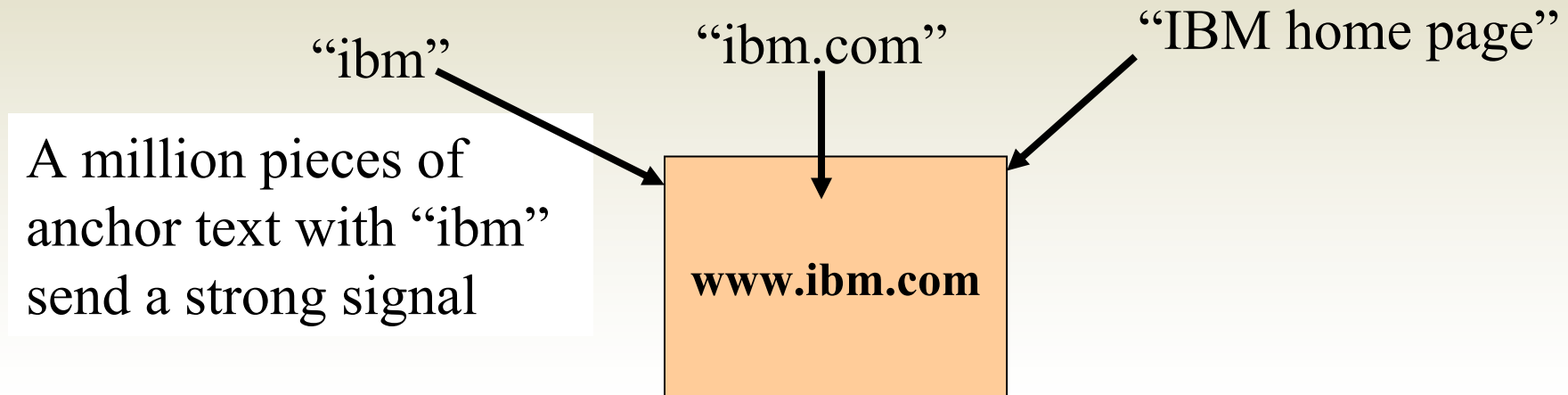
**Assumption 2:** The anchor of the hyperlink describes the target page (textual context)



# Anchor Text

*WWW Worm* - McBryan [Mcbr94]

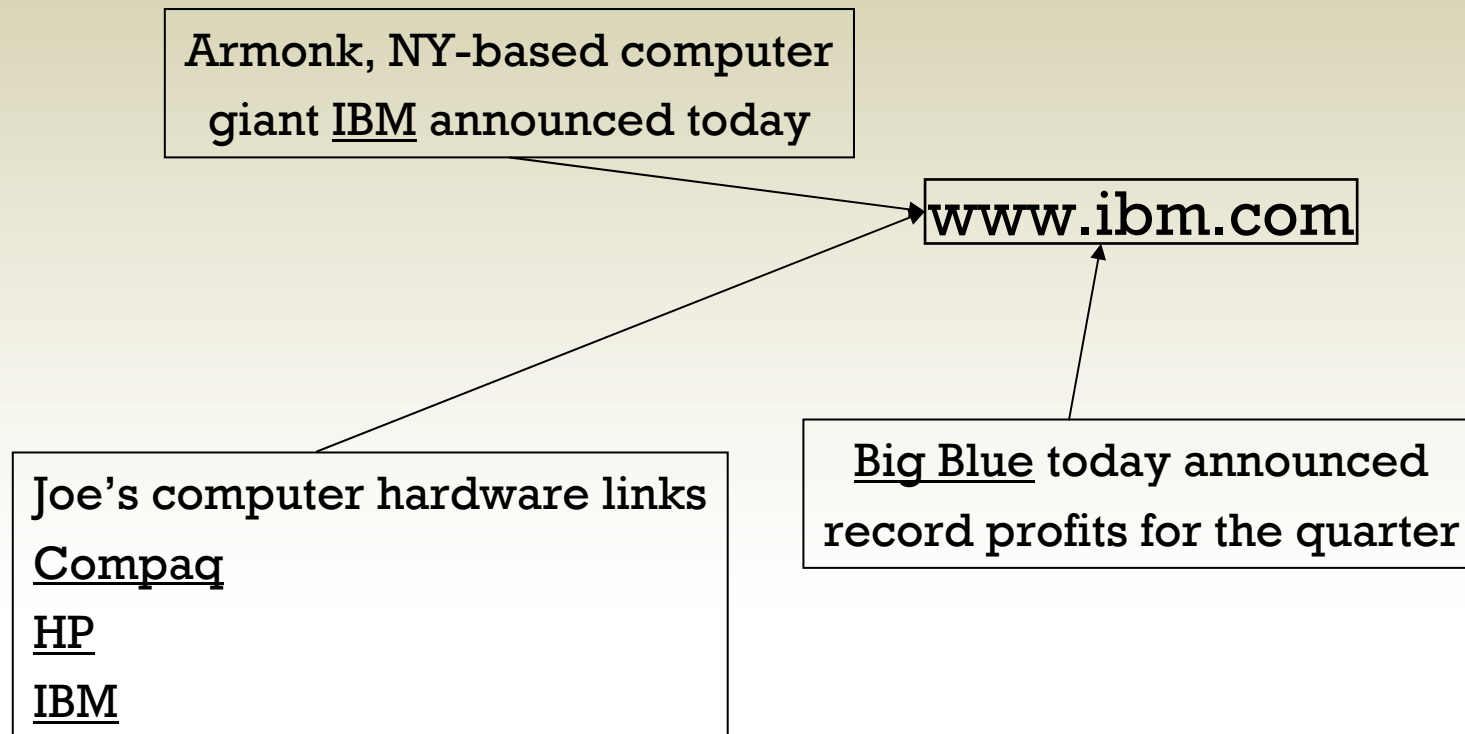
- For *ibm* how to distinguish between:
  - IBM's home page (mostly graphical)
  - IBM's copyright page (high term freq. for 'ibm')
  - Rival's spam page (arbitrarily high term freq.)





# Indexing anchor text

- When indexing a document  $D$ , include anchor text from links pointing to  $D$ .





## Indexing anchor text

- Can sometimes have unexpected side effects - *e.g., evil empire.*
- Can index anchor text with less weight.



# Anchor Text

- Other applications
  - Weighting/filtering links in the graph
    - HITS [Chak98], Hilltop [Bhar01]
  - Generating page descriptions from anchor text [Amit98, Amit00]



# Citation Analysis

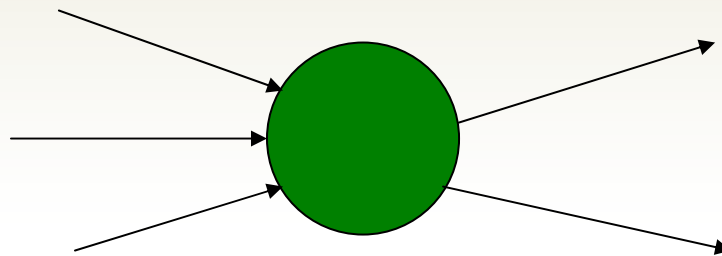
- Citation frequency
- Co-citation coupling frequency
  - Cocitations with a given author measures “impact”
  - Cocitation analysis [Mcca90]
- Bibliographic coupling frequency
  - Articles that co-cite the same articles are related
- Citation indexing
  - Who is author cited by? (Garfield [Garf72])
- Pagerank preview: Pinski and Narin '60s





# Query-independent ordering

- First generation: using link counts as simple measures of popularity.
- Two basic suggestions:
  - Undirected popularity:
    - Each page gets a score = the number of in-links plus the number of out-links ( $3+2=5$ ).
  - Directed popularity:
    - Score of a page = number of its in-links (3).





## Query processing

- First retrieve all pages meeting the text query (say *venture capital*).
- Order these by their link popularity (either variant on the previous page).

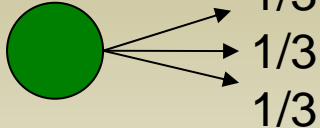


## Spamming simple popularity

- *Exercise:* How do you spam each of the following heuristics so your page gets a high score?
- Each page gets a score = the number of in-links plus the number of out-links.
- Score of a page = number of its in-links.



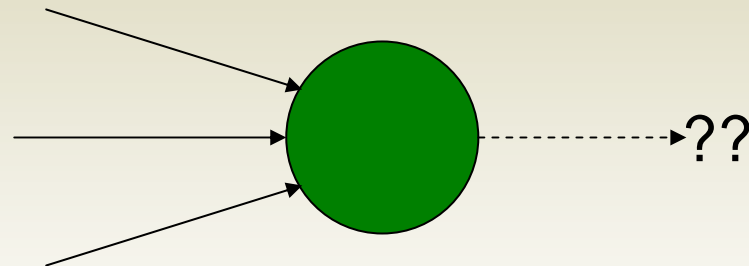
## Pagerank scoring

- Imagine a browser doing a random walk on web pages:
  - Start at a random page 
  - At each step, go out of the current page along one of the links on that page, equiprobably
- “In the steady state” each page has a long-term visit rate - use this as the page’s score.



## Not quite enough

- The web is full of dead-ends.
  - Random walk can get stuck in dead-ends.
  - Makes no sense to talk about long-term visit rates.





# Teleporting

- At a dead end, jump to a random web page.
- At any non-dead end, with probability 10%, jump to a random web page.
  - With remaining probability (90%), go out on a random link.
  - 10% - a parameter.



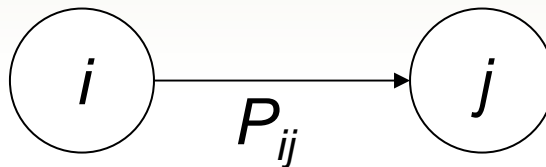
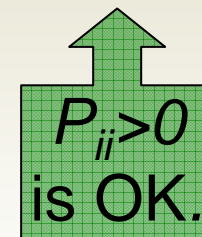
## Result of teleporting

- Now cannot get stuck locally.
- There is a long-term rate at which any page is visited (not obvious, will show this).
- How do we compute this visit rate?



# Markov chains

- A Markov chain consists of  $n$  states, plus an  $n \times n$  transition probability matrix  $\mathbf{P}$ .
- At each step, we are in exactly one of the states.
- For  $1 \leq i, j \leq n$ , the matrix entry  $P_{ij}$  tells us the probability of  $j$  being the next state, given we are currently in state  $i$ .

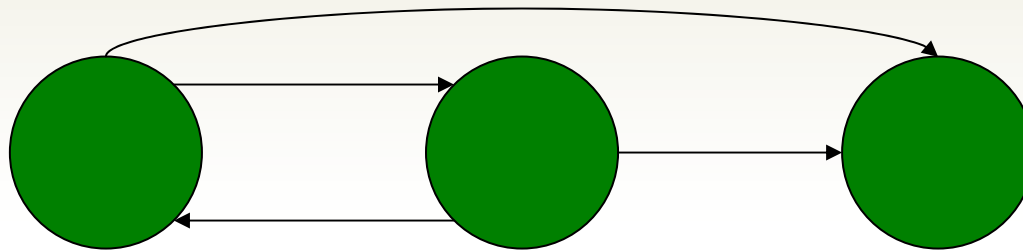






# Markov chains

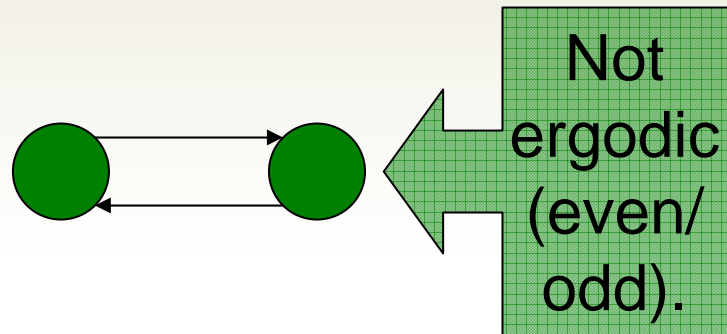
- Clearly, for all  $i$ ,  $\sum_{j=1}^n P_{ij} = 1$ .
- Markov chains are abstractions of random walks.
- *Exercise*: represent the teleporting random walk from 3 slides ago as a Markov chain, for this case:





# Ergodic Markov chains

- A Markov chain is ergodic if
  - you have a path from any state to any other
  - For any start state, after a finite transient time  $T_0$ , the probability of being in any state at a fixed time  $T > T_0$  is nonzero.





## Ergodic Markov chains

- For any ergodic Markov chain, there is a unique long-term visit rate for each state.
  - *Steady-state probability distribution.*
- Over a long time-period, we visit each state in proportion to this rate.
- It doesn't matter where we start.



## Probability vectors

- A probability (row) vector  $\mathbf{x} = (x_1, \dots, x_n)$  tells us where the walk is at any point.
- E.g., (000...1...000) means we're in state  $i$ .

$$1 \quad i \quad n$$

More generally, the vector  $\mathbf{x} = (x_1, \dots, x_n)$  means the walk is in state  $i$  with probability  $x_i$ .

$$\sum_{i=1}^n x_i = 1.$$



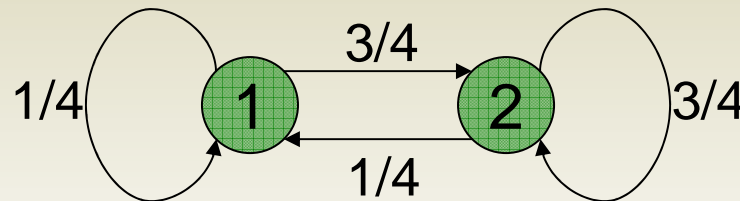
## Change in probability vector

- If the probability vector is  $\mathbf{x} = (x_1, \dots, x_n)$  at this step, what is it at the next step?
- Recall that row  $i$  of the transition prob. Matrix  $\mathbf{P}$  tells us where we go next from state  $i$ .
- So from  $\mathbf{x}$ , our next state is distributed as  $\mathbf{xP}$ .



## Steady state example

- The steady state looks like a vector of probabilities  $\mathbf{a} = (a_1, \dots, a_n)$ :
  - $a_i$  is the probability that we are in state  $i$ .



For this example,  $a_1=1/4$  and  $a_2=3/4$ .



## How do we compute this vector?

- Let  $\mathbf{a} = (a_1, \dots, a_n)$  denote the row vector of steady-state probabilities.
- If we our current position is described by  $\mathbf{a}$ , then the next step is distributed as  $\mathbf{aP}$ .
- But  $\mathbf{a}$  is the steady state, so  $\mathbf{a}=\mathbf{aP}$ .
- Solving this matrix equation gives us  $\mathbf{a}$ .
  - So  $\mathbf{a}$  is the (left) eigenvector for  $\mathbf{P}$ .
  - (Corresponds to the “principal” eigenvector of  $\mathbf{P}$  with the largest eigenvalue.)
  - Transition probability matrices always have largest eigenvalue 1.



## One way of computing $\mathbf{a}$

- Recall, regardless of where we start, we eventually reach the steady state  $\mathbf{a}$ .
- Start with any distribution (say  $\mathbf{x}=(10\dots0)$ ).
- After one step, we're at  $\mathbf{xP}$ ;
- after two steps at  $\mathbf{xP}^2$ , then  $\mathbf{xP}^3$  and so on.
- “Eventually” means for “large”  $k$ ,  $\mathbf{xP}^k = \mathbf{a}$ .
- Algorithm: multiply  $\mathbf{x}$  by increasing powers of  $\mathbf{P}$  until the product looks stable.





# Pagerank summary

- Preprocessing:
  - Given graph of links, build matrix  $\mathbf{P}$ .
  - From it compute  $\mathbf{a}$ .
  - The entry  $a_i$  is a number between 0 and 1: the pagerank of page  $i$ .
- Query processing:
  - Retrieve pages meeting query.
  - Rank them by their pagerank.
  - Order is *query-independent*.



## PageRank: Πώς χρησιμοποιείται (1/3)

- PageRank is one of the methods Google uses to determine a page's relevance or importance. It is only one part of the story when it comes to the Google listing
- PageRank is also displayed on the toolbar of your browser if you've installed the Google toolbar (<http://toolbar.google.com/>). But the Toolbar PageRank only goes from 0 – 10 and seems to be something like a logarithmic scale:

Toolbar PageRank (log base 10)	Real PageRank
0	0 - 10
1	100 - 1,000
2	1,000 - 10,000
3	10,000 - 100,000
4	and so on...



## PageRank: Πώς χρησιμοποιείται (2/3)

- We can't know the exact details of the scale because, as we'll see later, the maximum PR of all pages on the web changes every month when Google does its re-indexing
- Also the toolbar sometimes guesses! The toolbar often shows me a Toolbar PR for pages I've only just uploaded and cannot possibly be in the index yet!



## PageRank: Πώς χρησιμοποιείται (3/3)

- What seems to be happening is that the toolbar looks at the URL of the page the browser is displaying and strips off everything down the last “/” (i.e. it goes to the “parent” page in URL terms)
- If Google has a Toolbar PR for that parent then it subtracts 1 and shows that as the Toolbar PR for this page. If there’s no PR for the parent it goes to the parent’s parent’s page, but subtracting 2, and so on all the way up to the root of your site
- If it can’t find a Toolbar PR to display in this way, that is if it doesn’t find a page with a real calculated PR, then the bar is greyed out



## PageRank: Βασική εξίσωση (1/3)

- 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
- If there’s no link there’s no support (but it’s an abstention from voting rather than a vote against the page)

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

- Note that the PageRanks form a probability distribution over web pages, so the sum of all web pages' PageRanks will be 1



## PageRank: Βασική εξίσωση (2/3)

- **PR(T<sub>n</sub>)** - Each page has a notion of its own self-importance. That's “PR(T<sub>1</sub>)” for the first page in the web all the way up to “PR(T<sub>n</sub>)” for the last page
- **C(T<sub>n</sub>)** - Each page spreads its vote out evenly amongst all of its outgoing links. The count, or number, of outgoing links for page 1 is “C(T<sub>1</sub>)”, “C(T<sub>n</sub>)” for page n, and so on
- **PR(T<sub>n</sub>)/C(T<sub>n</sub>)** - so if our page (page A) has a backlink from page “n” the share of the vote page A will get is “PR(T<sub>n</sub>)/C(T<sub>n</sub>)”



## PageRank: Βασική εξίσωση (3/3)

- $d(\dots$  - 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  $d(\dots$ . 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.



## PageRank: Πώς υπολογίζεται; (1/5)

- 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...
- And when you consider that page links can form circles it seems impossible to do this calculation!
- But actually it's not that bad. Remember this bit of the Google paper:
  - PageRank or  $PR(A)$  can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the Web





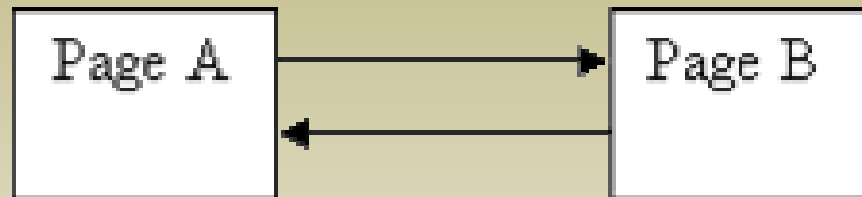
## PageRank: Πώς υπολογίζεται; (2/5)

- What that means to us is that we can just go ahead and calculate a page's PR without knowing the final value of the PR of the other pages
- That seems strange but, basically, each time we run the calculation we're getting a closer estimate of the final value
- So all we need to do is remember the each value we calculate and repeat the calculations lots of times until the numbers stop changing much.



## PageRank: Πώς υπολογίζεται; (3/5)

- Lets take the simplest example network: two pages, each pointing to the other:



- Each page has one outgoing link (the outgoing count is 1, i.e.  $C(A) = 1$  and  $C(B) = 1$ )
- **Guess 1:** We don't know what their PR should be to begin with, so let's take a guess at 1.0 and do some calculations with  $d = 0.85$
- $PR(A) = (1-d) + d(PR(B)/1)$      $PR(B) = (1-d) + d(PR(A)/1)$
- $PR(A) = 0.15 + 0.85 * 1 = 1$      $PR(B) = 0.15 + 0.85 * 1 = 1$
- The numbers aren't changing at all! So it looks like we started out with a lucky guess!!!



## PageRank: Πώς υπολογίζεται; (4/5)

- **Guess 2:** No, that's too easy, maybe I got it wrong (and it wouldn't be the first time). Ok, let's start the guess at 0 instead and re-calculate:
- $PR(A) = 0.15 + 0.85 * 0 = 0.15$
- $PR(B) = 0.15 + 0.85 * 0.15 = 0.2775$
- $PR(A) = 0.15 + 0.85 * 0.2775 = 0.385875$
- $PR(B) = 0.15 + 0.85 * 0.385875 = 0.47799375$
- $PR(A) = 0.15 + 0.85 * 0.47799375 = 0.5562946875$
- $PR(B) = 0.15 + 0.85 * 0.5562946875 = 0.622850484375$ 
  - and so on. The numbers just keep going up. But will the numbers stop increasing when they get to 1.0? What if a calculation over-shoots and goes above 1.0?



## PageRank: Πώς υπολογίζεται; (5/5)

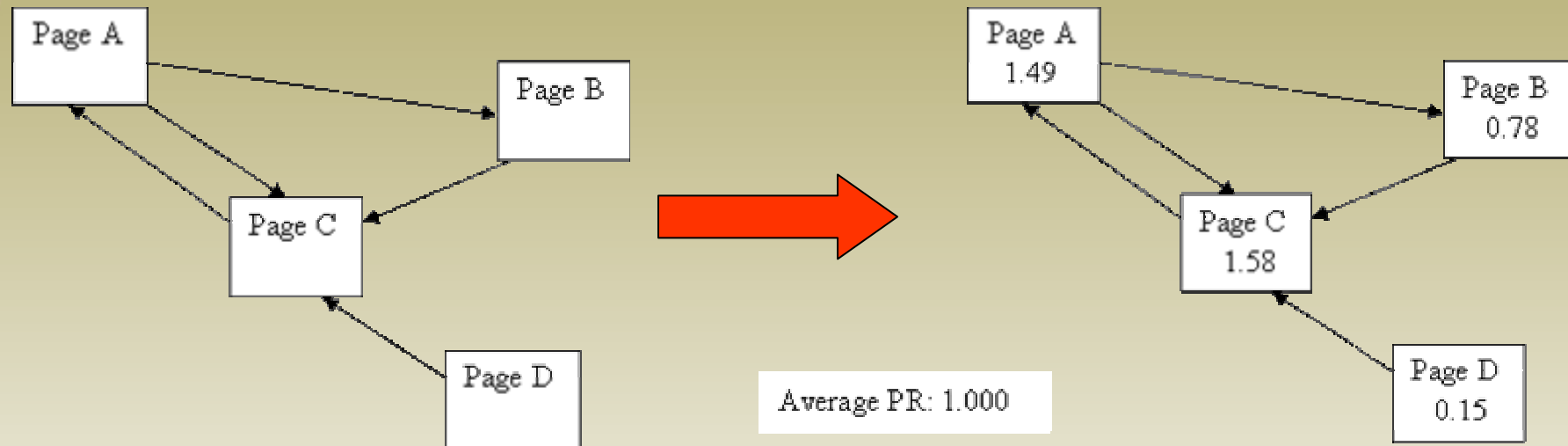
- **Guess 3:** Well let's see. Let's start the guess at 40 each and do a few cycles:
- $PR(A) = 40$
- $PR(B) = 40$
- First calculation
- $PR(A) = 0.15 + 0.85 * 40 = 34.25$
- $PR(B) = 0.15 + 0.85 * 0.385875 = 29.1775$
- $PR(A) = 0.15 + 0.85 * 29.1775 = 24.950875$
- $PR(B) = 0.15 + 0.85 * 24.950875 = 21.35824375$
- Those numbers are heading down alright! It sure looks the numbers will get to 1.0 and stop



## PageRank: Γρήγορος υπολογισμός

- How many times do we need to repeat the calculation for big networks? That's a difficult question; for a network as large as the World Wide Web it can be many millions of iterations!
- The “damping factor” is quite subtle. If it's too high then it takes ages for the numbers to settle, if it's too low then you get repeated over-shoot, both above and below the average - the numbers just swing about the average like a pendulum and never settle down.
- Also choosing the order of calculations can help. The answer will always come out the same no matter which order you choose, but some orders will get you there quicker than others

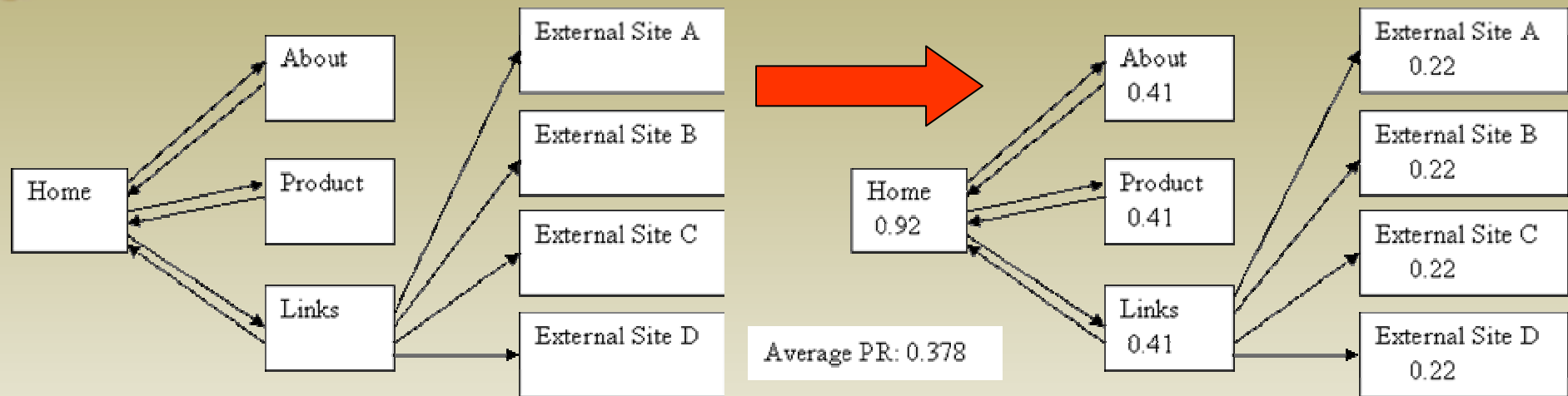
# PageRank: Παράδειγμα 1



- it took about 20 iterations before the network began to settle on these values
- Look at Page D - it has a PR of 0.15 even though no-one is voting for it. So, for Page D, no backlinks means the equation looks like this:  
$$PR(D) = (1-d) + d * (0) = 0.15$$
- **Observation:** every page has at least a PR of 0.15 to share out
  - But this may only be in theory - there are rumours that Google undergoes a post-spidering phase whereby any pages that have no incoming links at all are completely deleted from the index



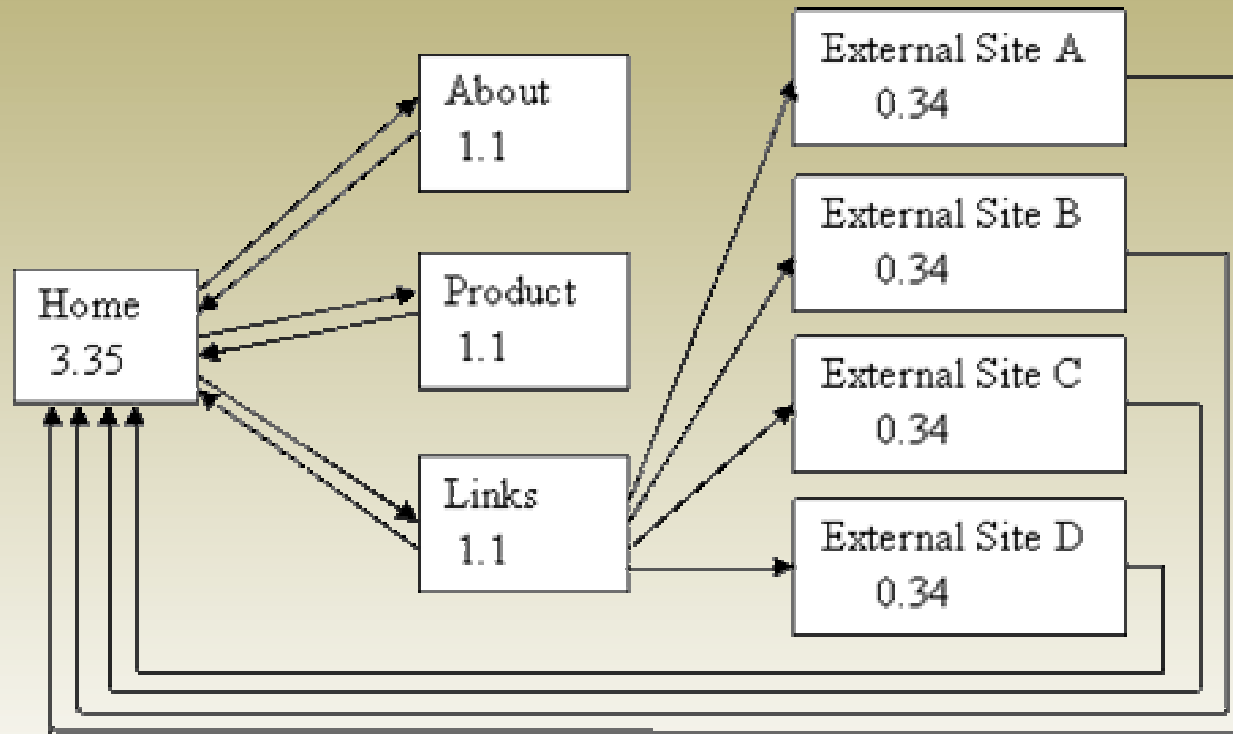
## PageRank: Παράδειγμα 2



- As you'd expect, the home page has the most PR –it has the most incoming links! But what's happened to the average? It's only 0.378!!! That doesn't tie up with what I said earlier so something is wrong somewhere!
- Well no, everything is fine. But take a look at the “external site” pages – what's happening to their PageRank? They're not passing it on, they're not voting for anyone, they're wasting their PR!!!



## PageRank: Παράδειγμα 3



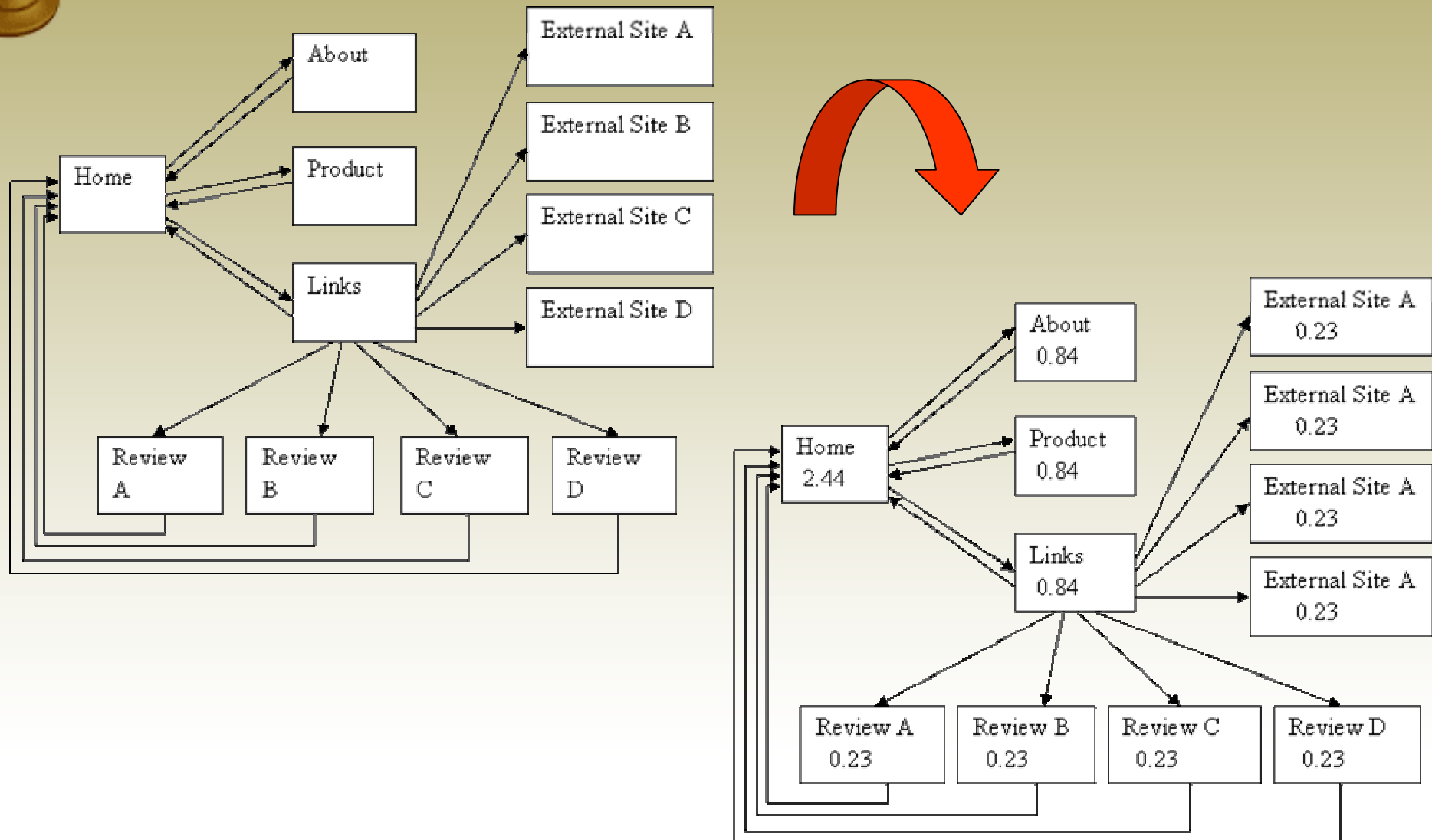
Average PR: 1.000

- That's better - it does work after all! And look at the PR of our home page! All those incoming links sure make a difference – we'll talk more about that later.



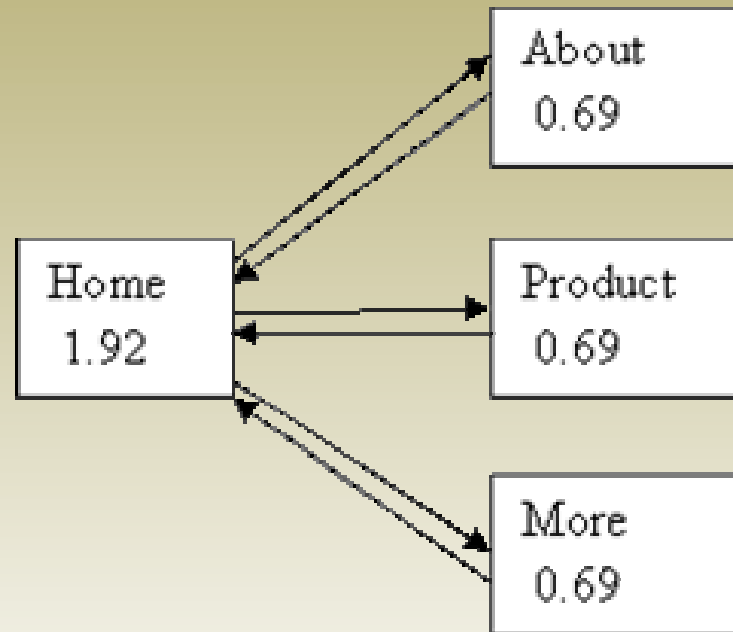


# PageRank: Παράδειγμα 4





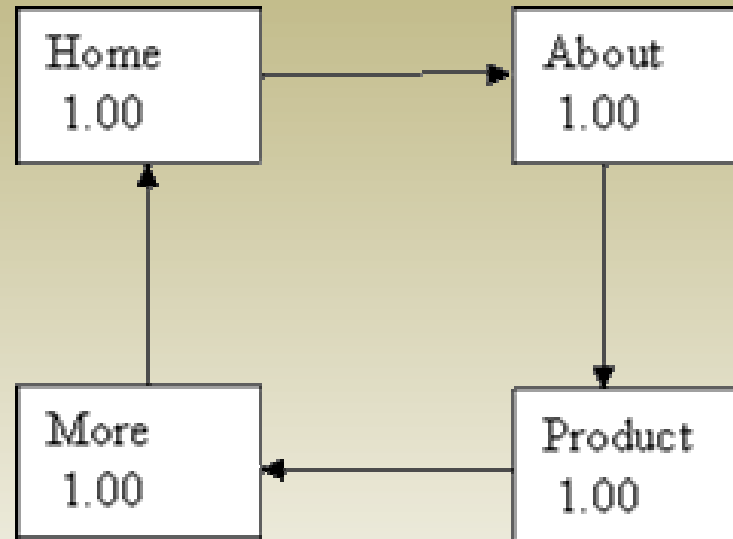
## PageRank: Παράδειγμα 5



- Our home page has 2 and a half times as much PR as the child pages! Excellent!
- **Observation:** a hierarchy concentrates votes and PR into one page



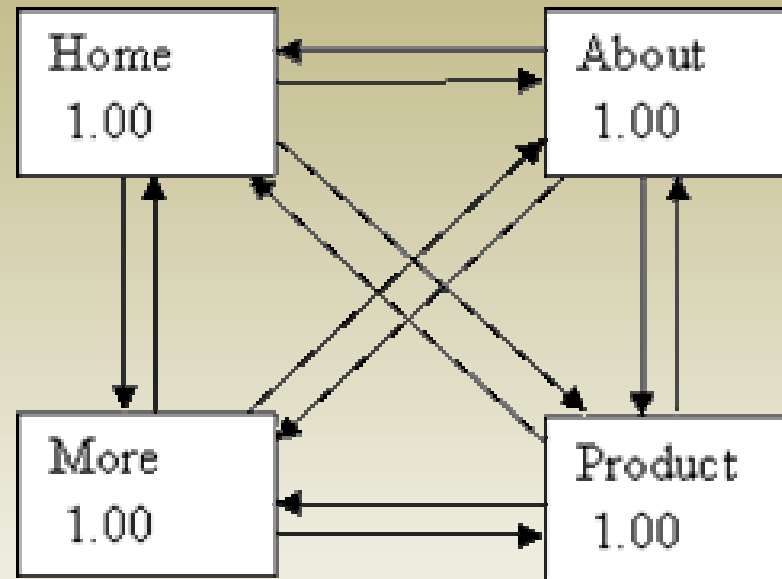
## PageRank: Παράδειγμα 6



- This is what we'd expect. All the pages have the same number of incoming links, all pages are of equal importance to each other, all pages get the same PR of 1.0 (i.e. the “average” probability).



## PageRank: Παράδειγμα 7



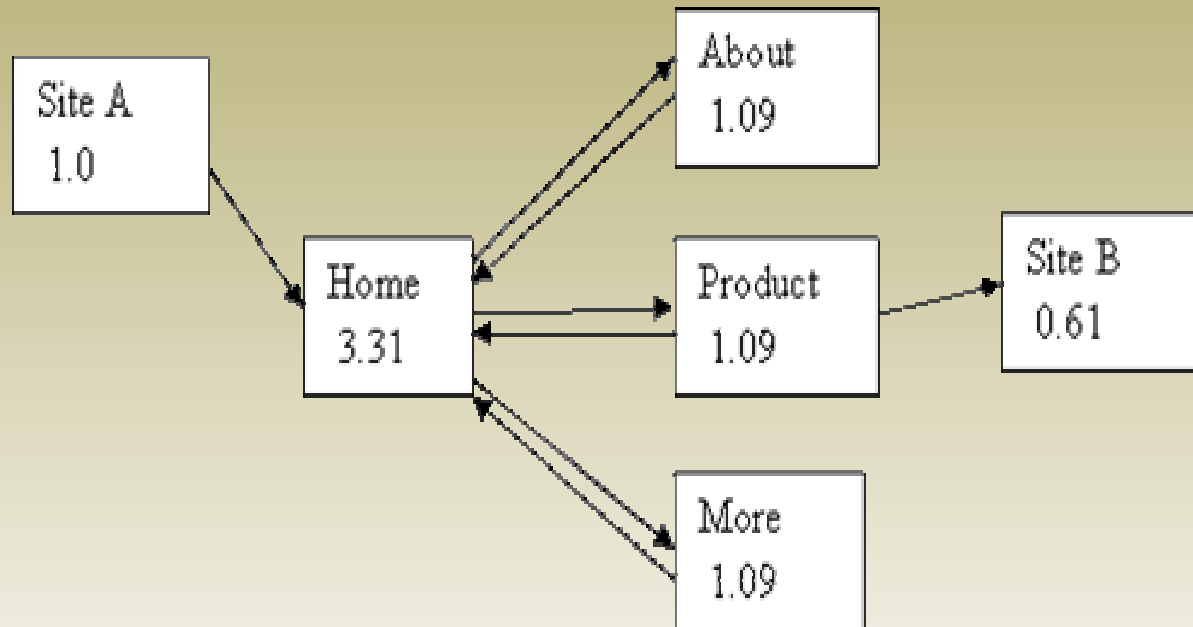
- Yes, the results are the same as the Looping example above and for the same reasons



## PageRank: Παράδειγμα 8

We'll assume there's an external site that has lots of pages and links with the result that one of the pages has the average PR of 1.0.

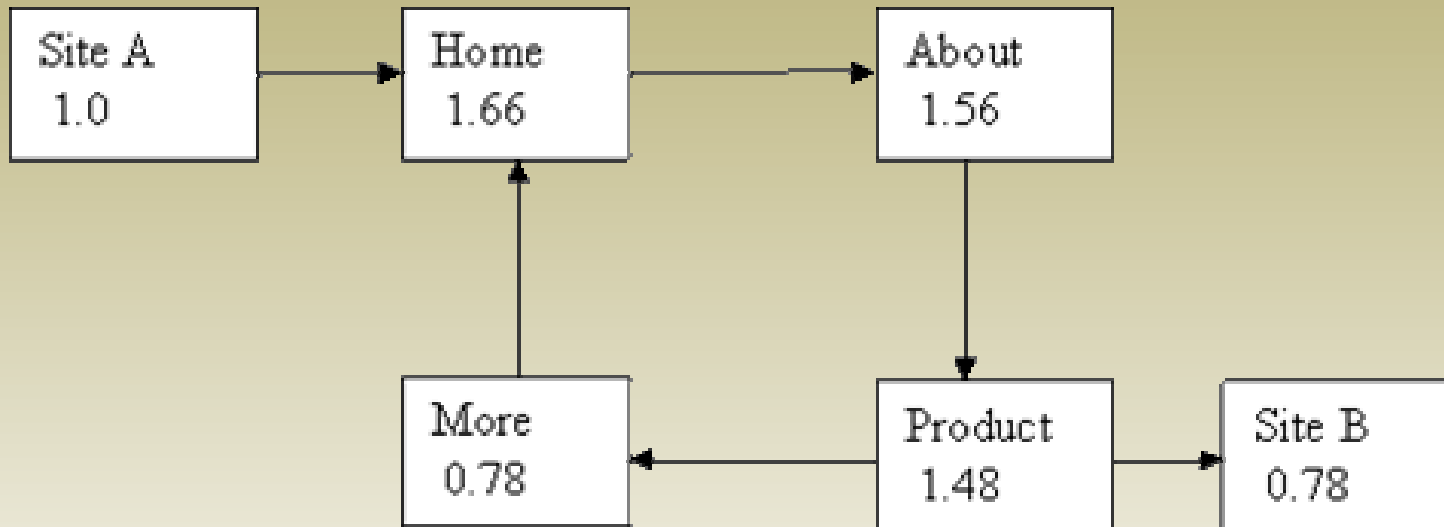
We'll also assume the webmaster really likes us – there's just one link from that page and it's pointing at our home page



- In example 5 the home page only had a PR of 1.92 but now it is 3.31! Excellent! Not only has site A contributed 0.85 PR to us, but the raised PR in the “About”, “Product” and “More” pages has had a lovely “feedback” effect, pushing up the home page’s PR even further!
- **Principle:** a well structured site will amplify the effect of any contributed PR



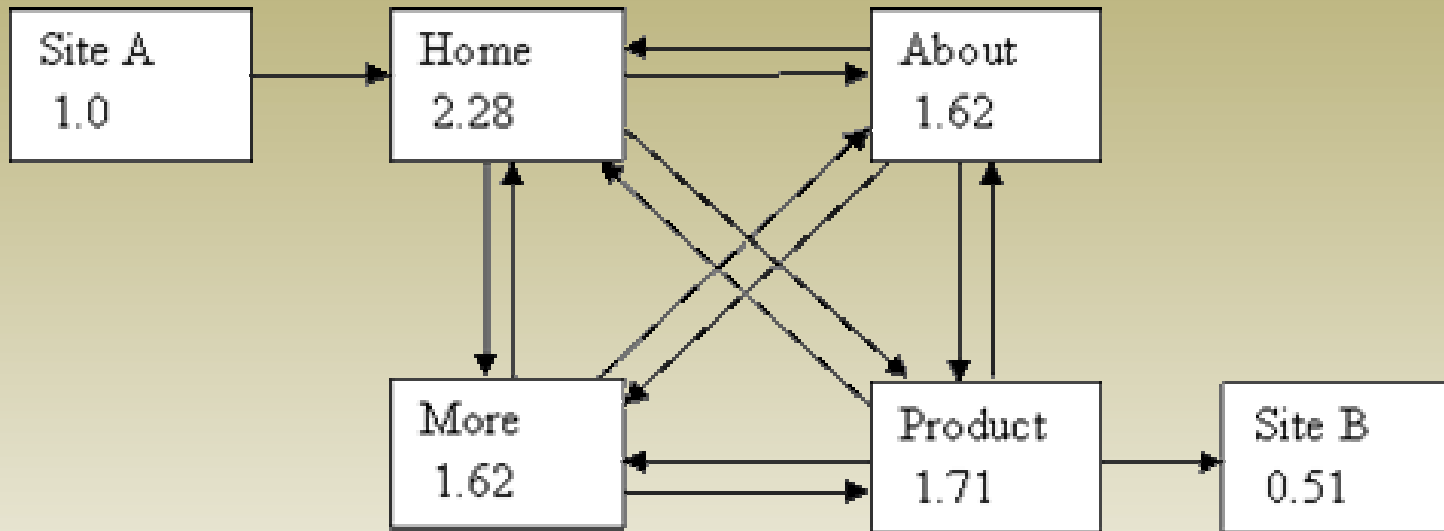
## PageRank: Παράδειγμα 9



- Well, the PR of our home page has gone up a little, but what's happened to the “More” page?
- The vote of the “Product” page has been split evenly between it and the external site. We now value the external Site B equally with our “More” page. The “More” page is getting only half the vote it had before – this is good for Site B but very bad for us



## PageRank: Παράδειγμα 10 (1/2)



- That's much better. The “More” page is still getting less share of the vote than in example 7 of course, but now the “Product” page has kept three quarters of its vote within our site - unlike example 9 where it was giving away fully half of it's vote to the external site!
- Keeping just this small extra fraction of the vote within our site has had a very nice effect on the Home Page too – PR of 2.28 compared with just 1.66 in example 9

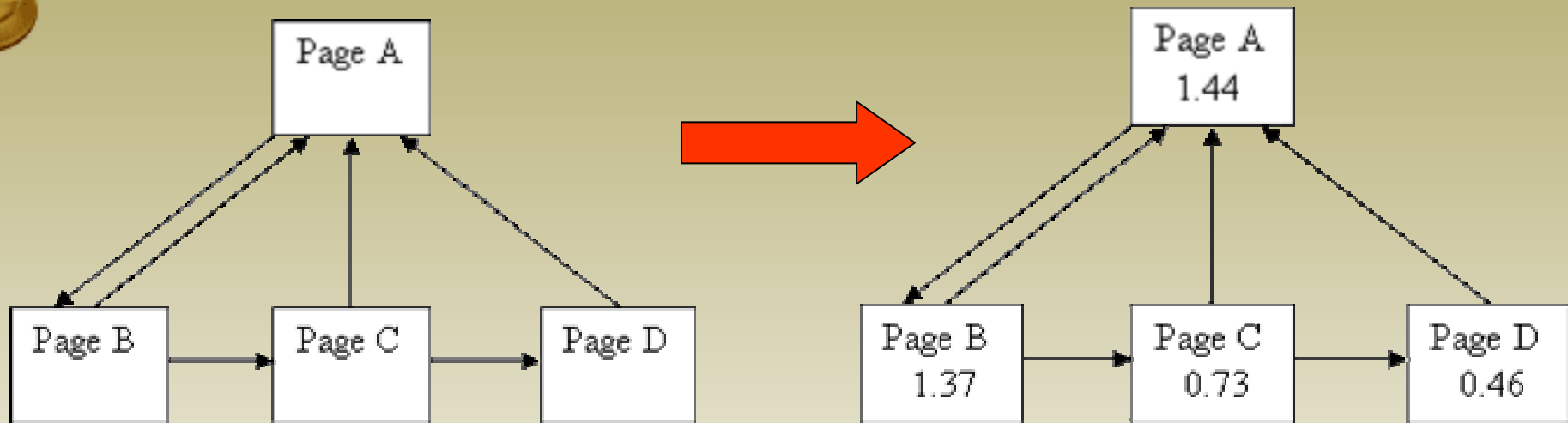


## PageRank: Παράδειγμα 10 (2/2)

- **Observation:** increasing the internal links in your site can minimize the damage to your PR when you give away votes by linking to external sites.
- **Principle:**
  - If a particular page is highly important – use a hierarchical structure with the important page at the “top”.
  - Where a group of pages may contain outward links – increase the number of internal links to retain as much PR as possible.
  - Where a group of pages do not contain outward links – the number of internal links in the site has no effect on the site’s average PR. You might as well use a link structure that gives the user the best navigational experience



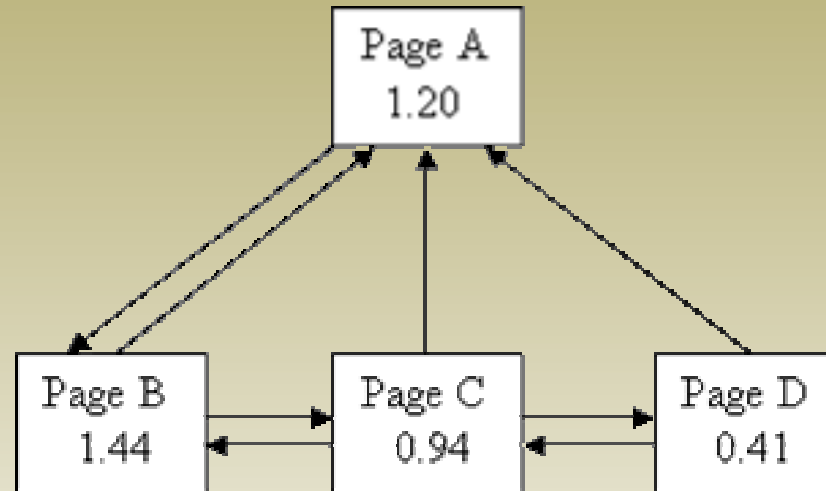
# PageRank: Παράδειγμα 11



- Lets try to fix our site to artificially concentrate the PR into the home page. That looks good, most of the links seem to be pointing up to page A so we should get a nice PR
- Oh— it's much worse than just an ordinary hierarchy! What's going on is that pages C and D have such weak incoming links that they're no help to page A at all!
- **Principle:** trying to abuse the PR calculation is harder than you think



## PageRank: Παράδειγμα 12 (1/2)



- A common web layout for long documentation is to split the document into many pages with a “Previous” and “Next” link on each plus a link back to the home page. The home page then only needs to point to the first page of the document
- In this simple example, where there’s only one document, the first page of the document has a higher PR than the Home Page! This is because page B is getting all the vote from page A, but page A is only getting fractions of pages B, C and D

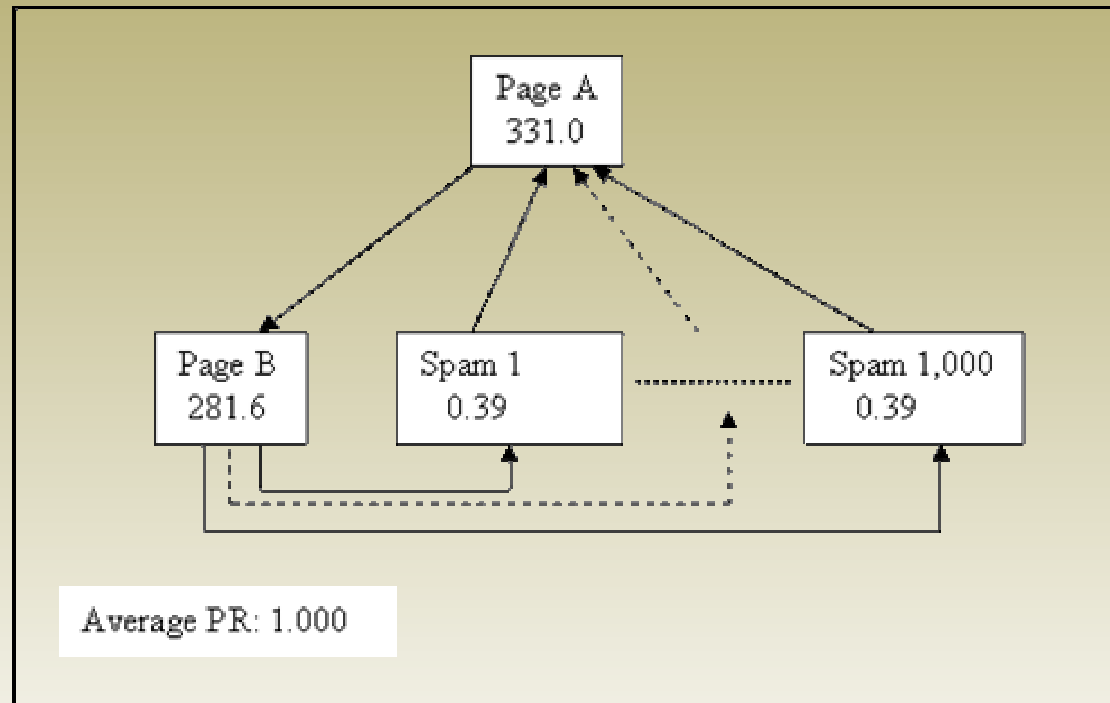


## PageRank: Παράδειγμα 12 (1/2)

- Principle: in order to give users of your site a good experience, you may have to take a hit against your PR. There's nothing you can do about this - and neither should you try to or worry about it! If your site is a pleasure to use lots of other webmasters will link to it and you'll get back much more PR than you lost.
- Can you also see the trend between this and the previous example? As you add more internal links to a site it gets closer to the Fully Meshed example where every page gets the average PR for the mesh.
- Observation: as you add more internal links in your site, the PR will be spread out more evenly between the pages



# PageRank: Παράδειγμα 13



- let's see if we can get 1,000 pages pointing to our home page, but only have one link leaving it
- Yup, those spam pages are pretty worthless but they sure add up!
- **Observation:** it doesn't matter how many pages you have in your site, your average PR will always be 1.0 at best. But a hierarchical layout can strongly concentrate votes, and therefore the PR, into the home page!



## Συμπεράσματα (1/2)

- From the Brin and Page paper, the average Actual PR of all pages in the index is 1.0!
- So if you add pages to a site you're building the total PR will go up by 1.0 for each page (but only if you link the pages together so the equation can work), but the average will remain the same.
- If you want to concentrate the PR into one, or a few, pages then hierarchical linking will do that. If you want to average out the PR amongst the pages then "fully meshing" the site (lots of evenly distributed links) will do that - examples 5, 6, and 7 above.



## Συμπεράσματα (2/2)

- Getting inbound links to your site is the only way to increase your site's average PR. How that PR is distributed amongst the pages on your site depends on the details of your internal linking and which of your pages are linked to.
- If you give outbound links to other sites then your site's average PR will decrease (you're not keeping your vote "in house" as it were). Again the details of the decrease will depend on the details of the linking.
- Given that the average of every page is 1.0 we can see that for every site that has an actual ranking in the millions (and there are some!) there must be lots and lots of sites who's Actual PR is below 1.0 (particularly because the absolute lowest Actual PR available is  $(1 - d)$ )