Ανάκληση Πληροφορίας

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 12: 07/04/2014

Τα μαθηματικά του PageRank

Η αρχική εξίσωση αθροίσματος

- Το PageRank μιας σελίδας είναι το άθροισμα του PageRank των σελίδων που δείχνουν σ' αυτή:

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$

- Το πρόβλημα με τη εξίσωση αυτή είναι ότι δεν ξέρουμε το PageRank των σελίδων που "δείχνουν" στη P_i
- Το πρόβλημα επιλύθηκε με επαναληπτική διαδικασία
  - Αρχικά κάθε σελίδα έχει το ίδιο PageRank, ίσο με 1/n
  - Ακολουθούμε την παραπάνω εξίσωση επαναληπτικά

Η επαναληπτική διαδικασία (1/2)

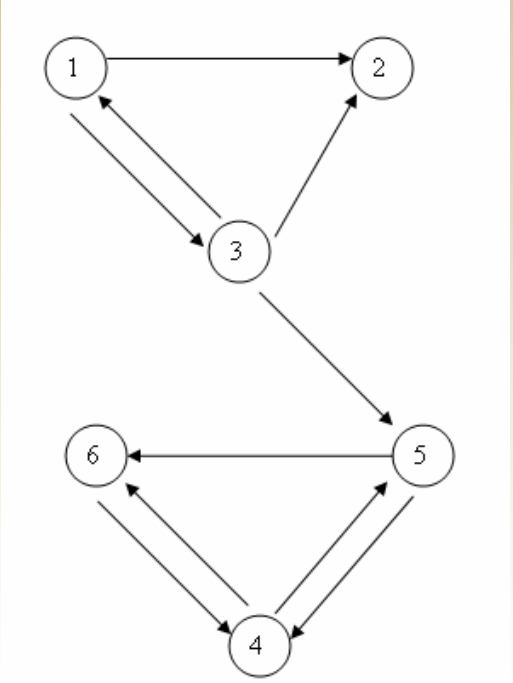
- Έστω ότι $r_{k+1}(P_i)$ είναι το PageRank της σελίδας P_i στην επανάληψη k+1:

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

- Η διαδικασία ξεκινά με $r_0(P_i)=1/n$ για κάθε σελίδα
- Συνεχίζεται με την ελπίδα ότι τελικά θα συγκλίνει

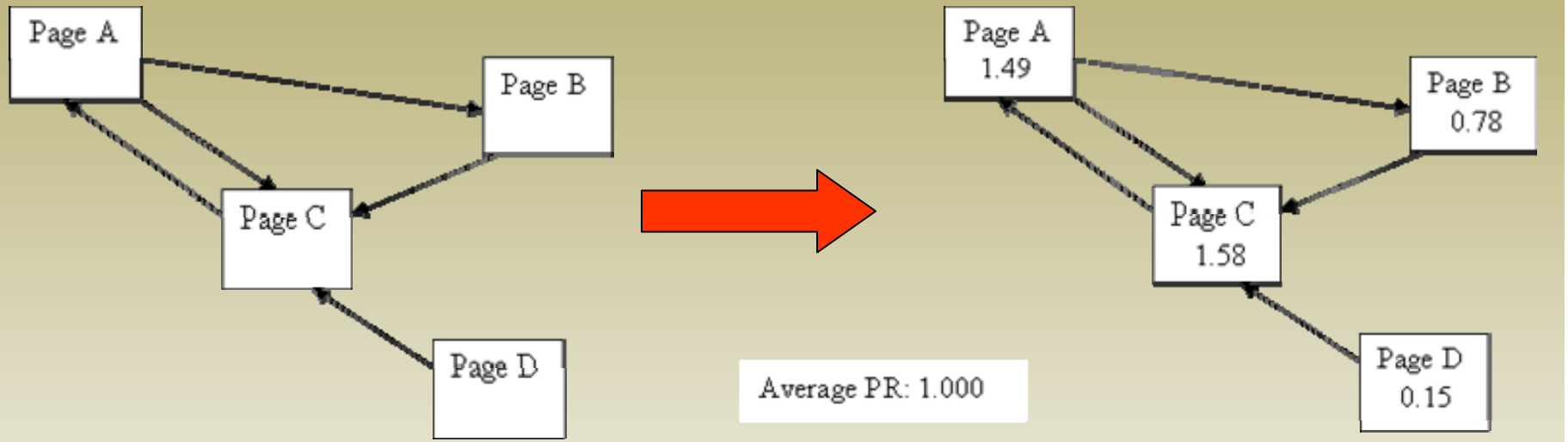
Η επαναληπτική διαδικασία (2/2)

- Εφαρμόζοντας την επαναληπτική διαδικασία στο μικρό γράφημα αριστερά, μετά από μερικές επαναλήψεις έχουμε τον πίνακα δεξιά:



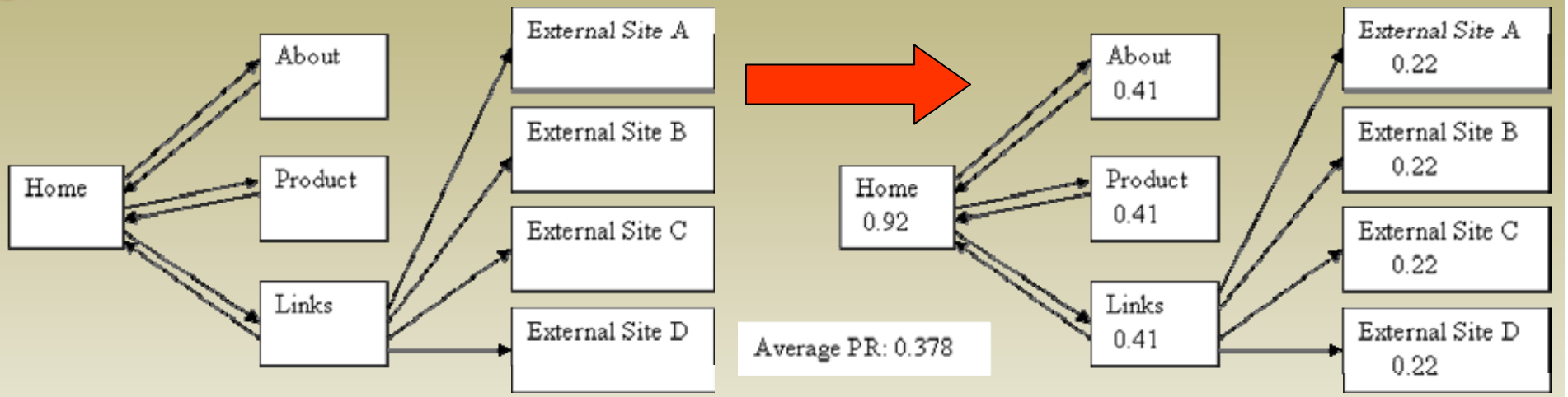
Iteration 0	Iteration 1	Iteration 2	Rank at Iter. 2
$r_0(P_1) = 1/6$	$r_1(P_1) = 1/18$	$r_2(P_1) = 1/36$	5
$r_0(P_2) = 1/6$	$r_1(P_2) = 5/36$	$r_2(P_2) = 1/18$	4
$r_0(P_3) = 1/6$	$r_1(P_3) = 1/12$	$r_2(P_3) = 1/36$	5
$r_0(P_4) = 1/6$	$r_1(P_4) = 1/4$	$r_2(P_4) = 17/72$	1
$r_0(P_5) = 1/6$	$r_1(P_5) = 5/36$	$r_2(P_5) = 11/72$	3
$r_0(P_6) = 1/6$	$r_1(P_6) = 1/6$	$r_2(P_6) = 14/72$	2

PageRank: Παράδειγμα 1



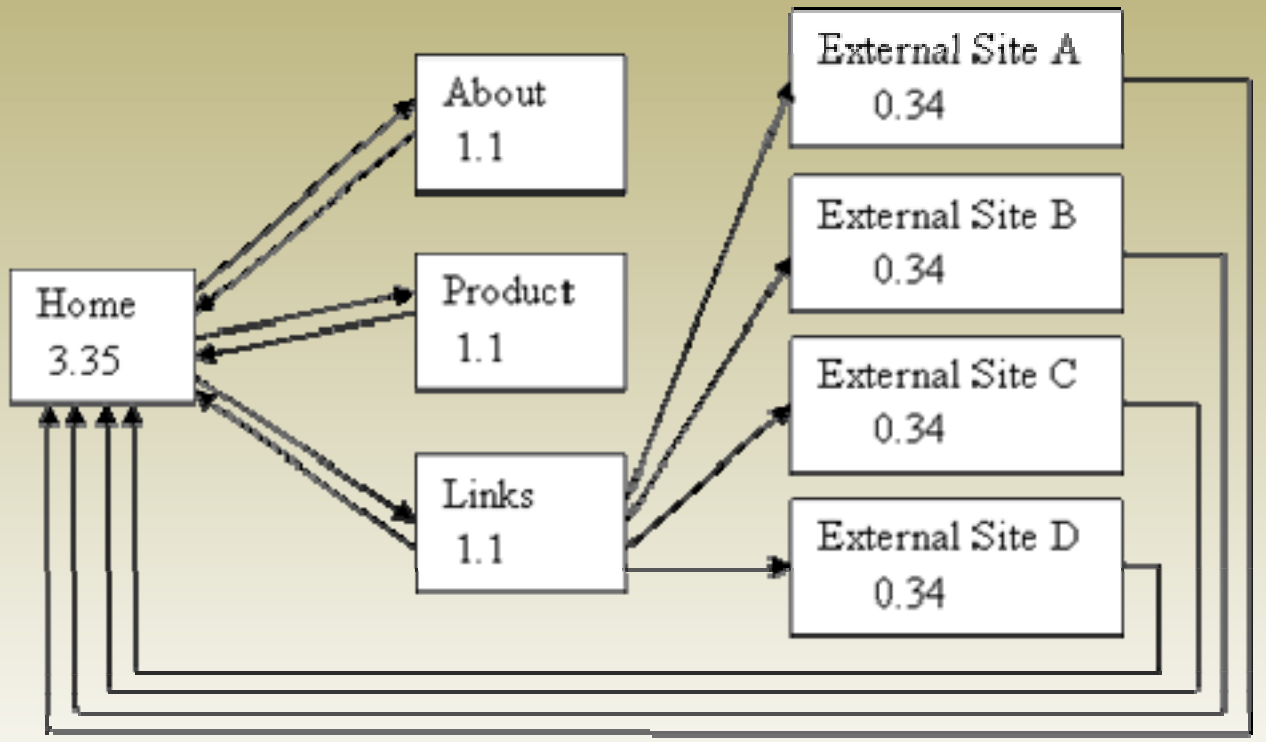
- it took about 20 iterations before the network began to settle on these values
- Look at Page D - it has a PR of 0.15 even though no-one is voting for it. So, for Page D, no backlinks means the equation looks like this: PR(A) = (1-d) + d * (0)= 0.15
- **Observation**: every page has at least a PR of 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!!!

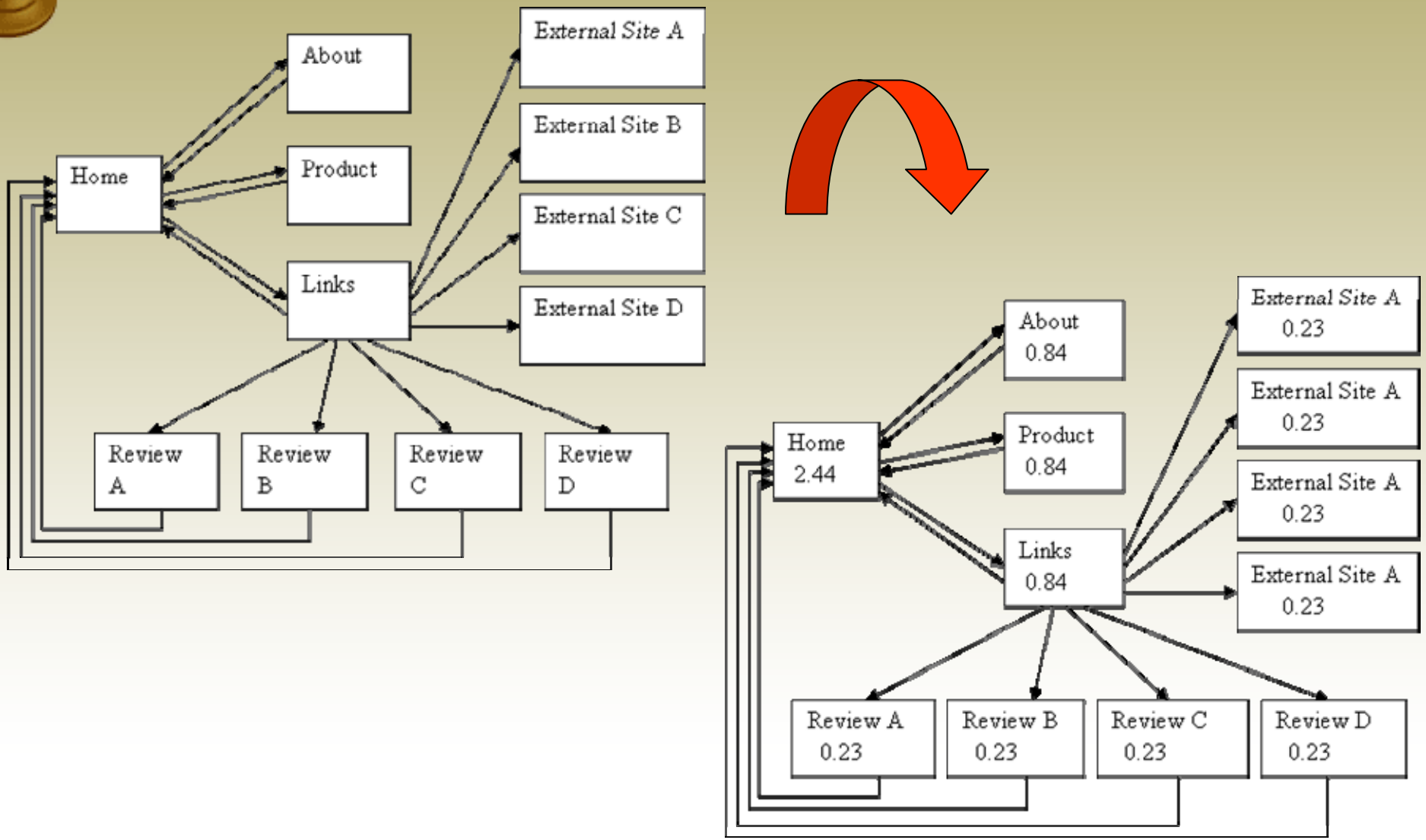
PageRank: Παράδειγμα 3



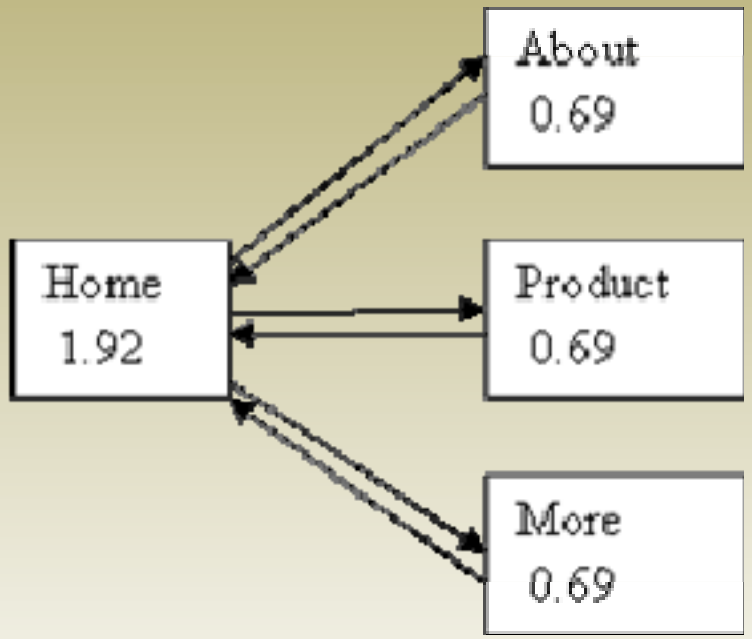
Average PR: 1.000

- 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

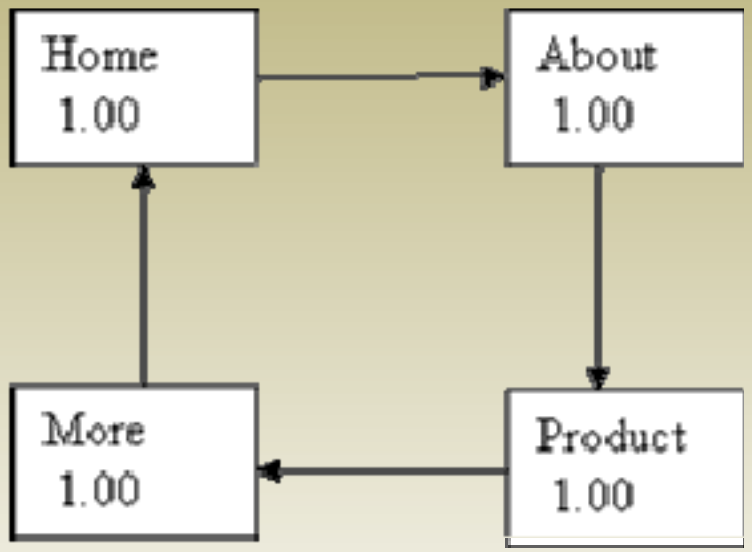


PageRank: Παράδειγμα 5



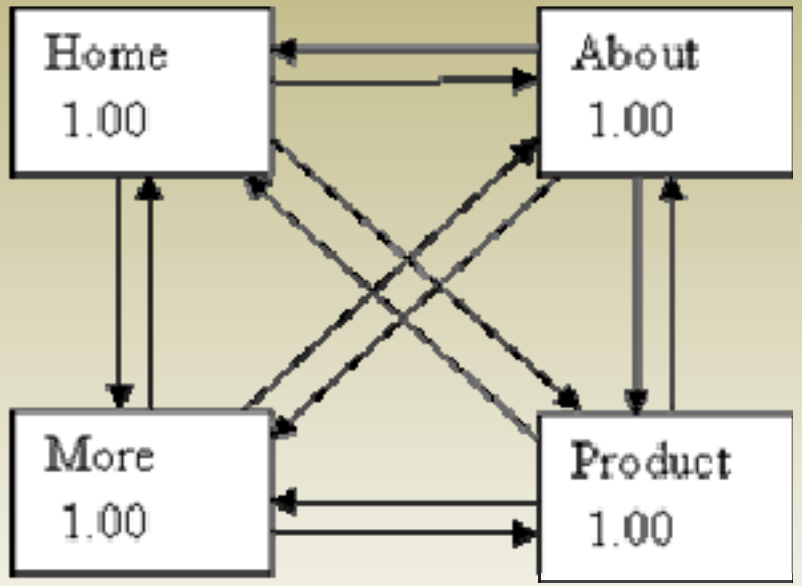
- Our home page has 2 and a half times as much PR as the child pages! Excellent!
- **Observation**: a hierarchy concentrates votes and PR into one page

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

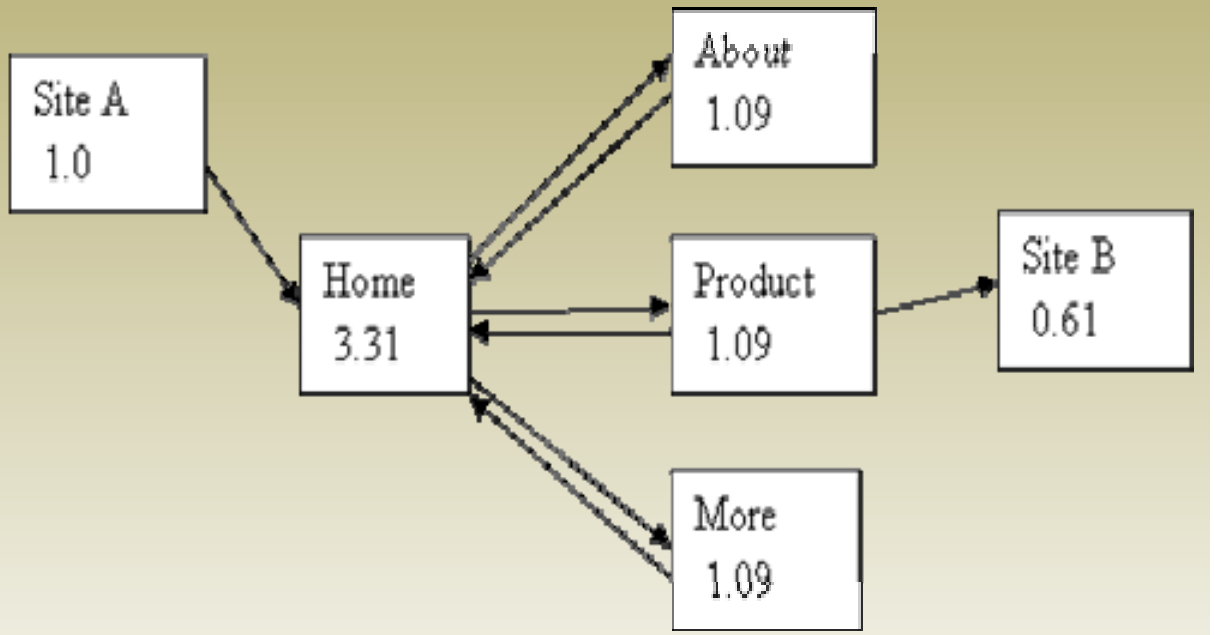


- Yes, the results are the same as the Looping example above and for the same reasons

PageRank: Παράδειγμα 8

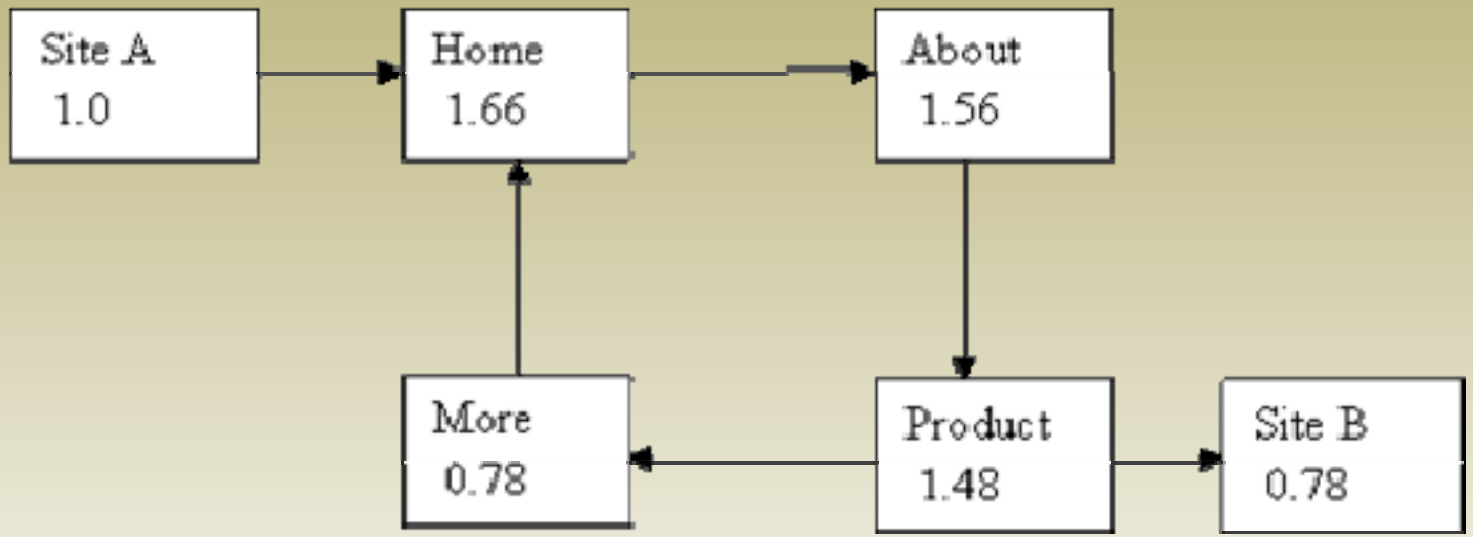
We'll assume there's an external site that has lots of pages and links with the result that one of the pages has the average PR of 1.0.

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



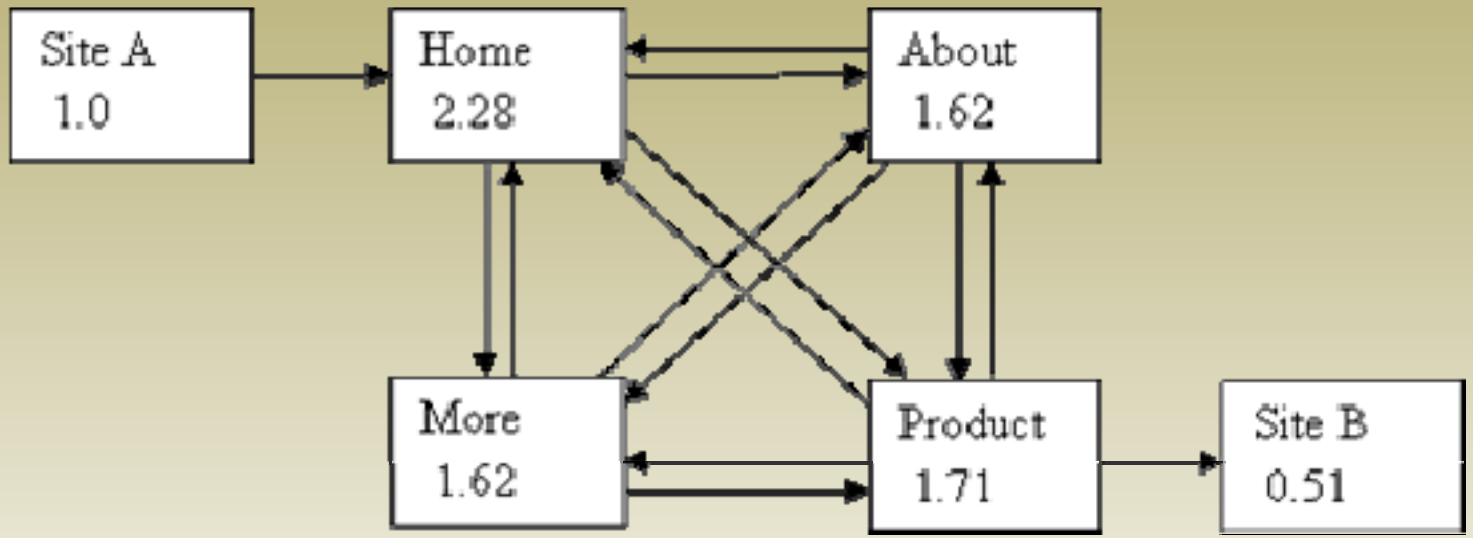
- In example 5 the home page only had a PR of 1.92 but now it is 3.31! Excellent! Not only has site A contributed 0.85 PR to us, but the raised PR in the "About", "Product" and "More" pages has had a lovely "feedback" effect, pushing up the home page's PR even further!
- **Priciple**: a well structured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



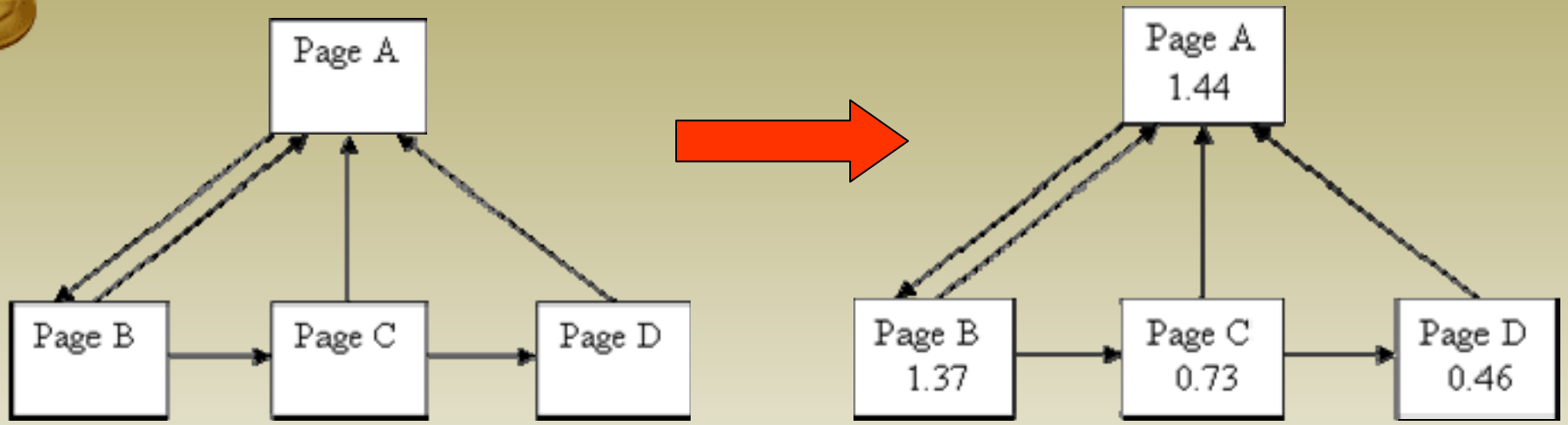
- That's much better. The "More" page is still getting less share of the vote than in example 7 of course, but now the "Product" page has kept three quarters of its vote within our site - unlike example 9 where it was giving away fully half of it's vote to the external site!

- 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)

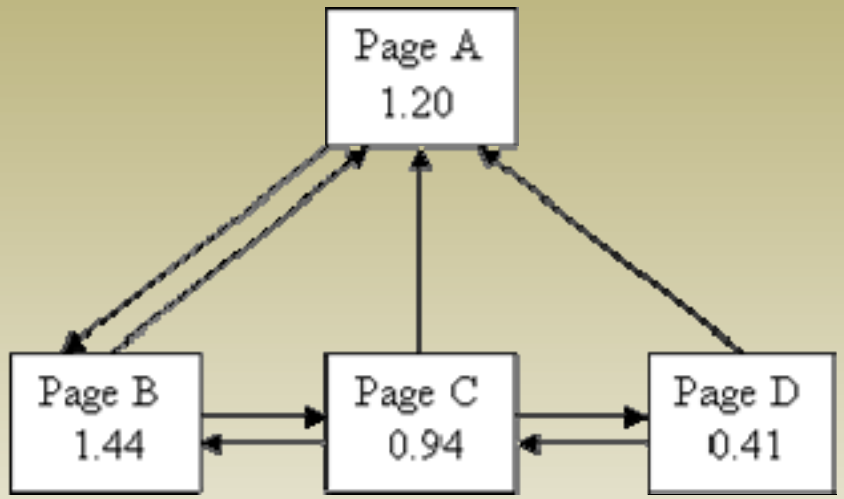
- **Observation**: increasing the internal links in your site can minimize the damage to your PR when you give away votes by linking to external sites.
- **Principle**:
  - If a particular page is highly important – use a hierarchical structure with the important page at the "top".
  - Where a group 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)

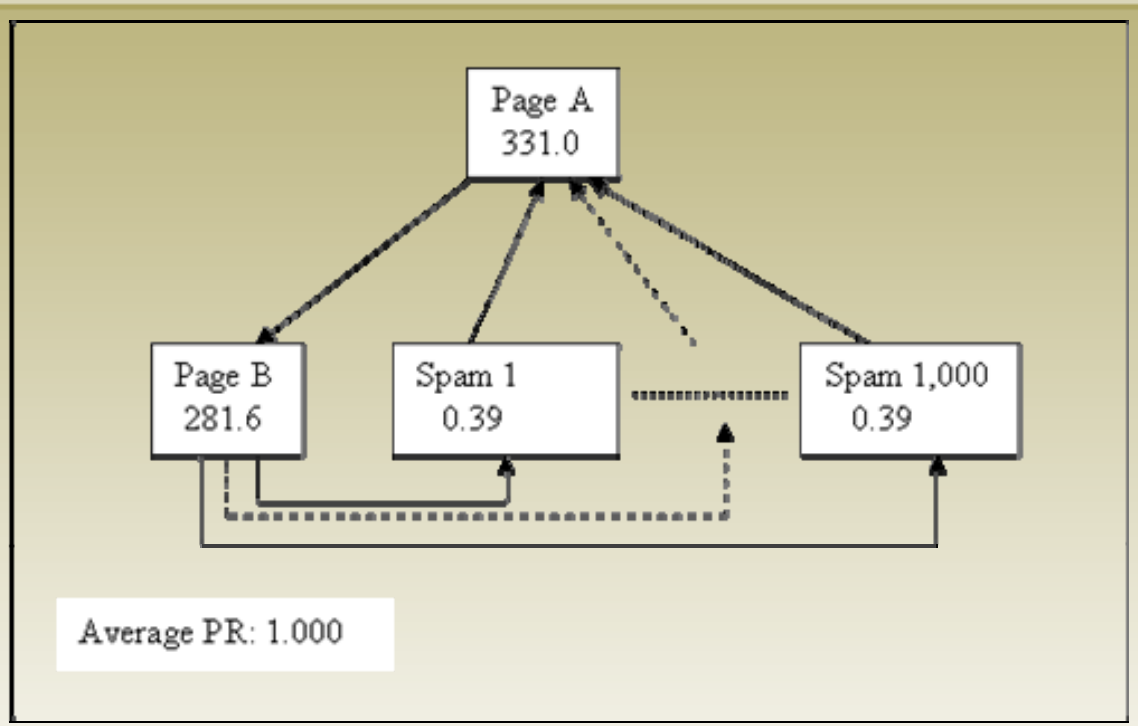


- A common web layout for long documentation is to split the document into many pages with a “Previous” and “Next” link on each plus a link back to the home page. The home page then only needs to point to the first page of the document
- In this simple example, where there’s only one document, the first 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))

Αναπαράσταση της επανάληψης με πίνακα

- Η προηγούμενες εξισώσεις υπολογίζουν το PageRank των σελίδων μια σελίδα κάθε φορά
- Με χρήση πινάκων αντικαθιστούμε το σύμβολο Σ
- Εισαγάγουμε
  - τον πίνακα H, και
  - το 1 x n διάνυσμα π^T
- Ο H είναι ένας row-normalized πίνακας υπερσυνδέσεων με $H_{ij}=1/|P_i|$, εάν υπάρχει σύνδεσμος από τον κόμβο i στον j, αλλιώς $H_{ij}=0$
- Παρόλο που ο H έχει την ίδια μη-μηδενική δομή με τον δυαδικό πίνακα γειτνιάσεων, τα μη μηδενικά στοιχεία του H είναι πιθανότητες

Παράδειγμα αναπαράστασης με πίνακα

$$\mathbf{H} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

- Τα μη-μηδενικά στοιχεία της γραμμής i αναπαριστούν τους εξερχόμενους συνδέσμους της σελίδας i
- Τα μη-μηδενικά στοιχεία της στήλης i αναπαριστούν τους εισερχόμενους συνδέσμους στη σελίδα i
- Η προηγούμενη εξίσωση γίνεται τώρα:

$$\pi^{(k+1)T} = \pi^{(k)T}\mathbf{H}$$

Επίδοση της αναπαράστασης με πίνακα

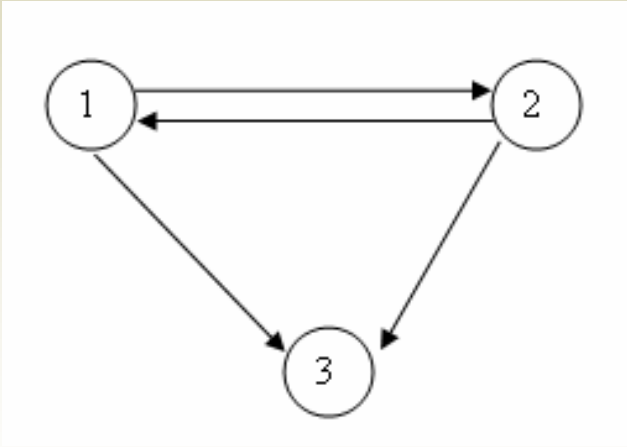
1. Κάθε επανάληψη της προηγούμενης εξίσωσης απαιτεί έναν πολλαπλασιασμό, άρα $O(n^2)$ πολυπλοκότητα
2. Ο H είναι γενικά πολύ αραιός (sparse), άρα
   - Απαιτεί μικρό αποθηκευτικό χώρο
   - Ο πολλαπλασιασμός είναι πιο οικονομικός σε σχέση με το $O(n^2)$
     - Απαιτεί μόνο O(nnz(H)), όπου nnz(H) είναι ο αριθμός των μη-μηδενικών
     - Μετρήσεις δείχνουν ότι το nnz(H) ~ 10n
     - Άρα υπολογιστικό κόστος της τάξης O(n)
3. Η επαναληπτική διαδικασία είναι απλά μια linear stationary process: είναι η κλασική power method πάνω στον H
4. Ο H μοιάζει με *στοχαστικό πίνακα πιθανοτήτων μετάβασης*, όμως είναι ***substochastic***, γιατί υπάρχουν **dangling nodes**, δηλ., χωρίς εξερχόμενους συνδέσμους

Προβλήματα της επαναληπτικής διαδικασίας

- Θα συγκλίνει;
- Κάτω από ποιες προϋποθέσεις ή ιδιότητες του Η θα συγκλίνει;
- Θα συγκλίνει σε κάτι που έχει “μαθηματικό”νόημα;
- Θα συγκλίνει σε ένα ή περισσότερα διανύσματα;
- Η σύγκλιση εξαρτάται από το αρχικό διάνυσμα $\pi^{(0)T}$;
- Πόσο γρήγορα θα συγκλίνει;

Προβλήματα της επαναληπτικής διαδικασίας

- Αρχικά, η επαναληπτική διαδικασία ξεκίνησε με $\pi^{(0)T}=1/n\mathbf{e}^T$ (όπου $\mathbf{e}^T$ είναι διάνυσμα-γραμμή με όλα 1)
- Προέκυψε το πρόβλημα της **καταβόθρας** (rank sinks)
  - σελίδες που αυξάνουν συνεχώς το PageRank τους
  - Στο παρακάτω παράδειγμα το κόμβος 3, ενώ στο προηγούμενο παράδειγμα η ομάδα των κόμβων 4, 5, και 6



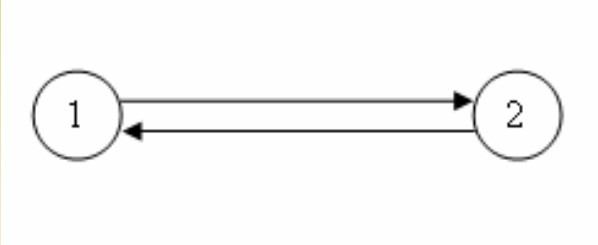
DK1

- Μετά από 13 επαναλήψεις, $\pi^{(13)T}=(0\ \ 0\ \ 0\ \ 2/3\ \ 1/3\ \ 1/5)$

DK1 Δεν αθροίζει στο 1. Γιατί?
Dimitrios Katsaros; 14/4/2011

Προβλήματα της επαναληπτικής διαδικασίας

- Επίσης, καθώς οι κόμβοι αυξάνουν συνεχώς το PageRank τους, μερικοί δεν έχουν καθόλου
  - Τότε, ποιο είναι το νόημα της ταξινόμησης με βάση το PageRank, όταν η πλειονότητα έχει PageRank ίσο με 0;
- Υπάρχει το πρόβλημα των κύκλων



- Εάν, ξεκινήσουμε με $\pi^{(0)T}=(1 \;\; 0)$, καταλήγουμε σε ατέρμονη διαδικασία
  - Στο διάνυσμα $\pi^{(k)T}=(1 \;\; 0)$ για άρτιο k
  - Στο διάνυσμα $\pi^{(k)T}=(0 \;\; 1)$ για περιττό k

Υπενθύμιση εννοιών Markov chains

- Με οποιοδήποτε διάνυσμα ξεκινήσουμε, όταν εφαρμοστεί η power method σε έναν Markov πίνακα P, συγκλίνει σε ένα μοναδικό θετικό διάνυσμα, το οποίο αποκαλείται *stationary vector*
- Προϋποθέσεις σύγκλισης
  - Ο P είναι stochastic: οι γραμμές αθροίζουν στο "1"
  - Ο P είναι irreducible: το υποκείμενο γράφημα είναι "strongly-connected"
  - Ο P είναι aperiodic: για οποιεσδήποτε σελίδες P_i και P_j υπάρχουν μονοπάτια από την P_i στην P_j (με οποιεσδήποτε επαναλήψεις) οποιουδήποτε μήκους, εκτός από ένα πεπερασμένο σύνολο μηκών
- Irreducible + aperiodic = primitive (πρωτογενής)
- Τα προβλήματα σύγκλισης του PageRank θα ξεπεραστούν εάν ο H τροποποιηθεί, ώστε να ικανοποιεί τις παραπάνω προϋποθέσεις

Πρώιμες προσαρμογές στο βασικό μοντέλο

- Οι Sergey Brin και Lawrence Page δεν χρησιμοποίησαν την έννοια της Markov chain, αλλά την έννοια του **random surfer**
- Μετά από "άπειρο χρόνο ταξιδιού", το ποσοστό του χρόνου που ο random surfer περνά σε μια σελίδα είναι ένα μέτρο της σημαντικότητας της σελίδας
- Δυστυχώς, υπάρχουν παγίδες για τον random surfer
  - pdf
  - image
  - data tables

Προσαρμογή στοχαστικότητας (1/2)

- Οι γραμμές $\mathbf{0}^T$ του H αντικαθίστανται με $1/n\mathbf{e}^T$
- Άρα ο random surfer, όταν συναντήσει έναν dangling node μπορεί από κει να μεταβεί σε οποιαδήποτε άλλη σελίδα
- Τον στοχαστικό πίνακα που προέκυψε από τον H τον συμβολίζουμε με S
- Για το γράφημα με τους 6 κόμβους είναι ο παρακάτω:

$$\mathbf{S} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Προσαρμογή στοχαστικότητας (2/2)

- Ο **S** παράγεται από μια *rank-one update* του **H**
- $\mathbf{S} = \mathbf{H} + \mathbf{a}(1/n\mathbf{e}^T)$
  - $a_i = 1$ εάν η σελίδα i είναι dangling node
  - $a_i = 0$ εάν η σελίδα i δεν είναι dangling node
- Ο **S** είναι συνδυασμός του αρχικού **H** με τον rank-one πίνακα $\mathbf{a}(1/n\mathbf{e}^T)$
- Η προσαρμογή αυτή εγγυάται ότι ο **S** είναι πίνακας μιας Markov chain
- Δεν εγγυάται όμως τη σύγκλιση

Προσαρμογή πρωτογένειας (1/2)

- Ο random surfer δεν ακολουθεί πάντα υπερσυνδέσμους
- Εγκαταλείπει την πλοήγηση και μεταβαίνει σε ένα "τυχαίο" URL
- "**Τηλεμεταφέρεται**" (**teleportation step**) και ξεκινά ξανά την πλοήγηση
- Προκύπτει ο πίνακας **G**, *Google matrix*

$$\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)1/n\mathbf{e}\mathbf{e}^T$$

- α (ελληνικό άλφα) έχει τιμή μεταξύ 0 και 1, και ελέγχει το ποσοστό του χρόνου που random surfer ακολουθεί υπερσυνδέσμους ή τηλεμεταφέρεται
- Η τελεμεταφορά είναι τυχαία, γιατί ο πίνακας τηλεμεταφοράς $\mathbf{E}=1/n\mathbf{e}\mathbf{e}^T$ είναι ομοιόμορφος

Συνέπειες της προσαρμογής πρωτογένειας

- Ο **G** είναι *stochastic*: κυρτός συνδυασμός δυο στοχαστικών πινάκων **S** και **E**
- Ο **G** είναι *irreducible*: κάθε σελίδα συνδέεται άμεσα με κάθε άλλη
- Ο **G** είναι *aperiodic*: οι βρόχοι ($\mathbf{G_{ii}} > 0$ για κάθε i) δημιουργούν aperiodicity
- Ο **G** είναι *primitive*: επειδή $\mathbf{G^k} > 0$ για κάποιο k (για k=1)
  - Υπάρχει ένα μοναδικό π^T και όταν εφαρμόσουμε την power method στον G, θα συγκλίνει σ' αυτό

Συνέπειες της προσαρμογής πρωτογένειας

- Ο **G** είναι πολύ πυκνός, ευτυχώς μπορεί να γραφεί ως rank-one update του πολύ αραιού πίνακα υπερσυνδέσμων **H**

$$\begin{aligned}\mathbf{G} &= \alpha\mathbf{S} + (1-\alpha)1/n\mathbf{e}\mathbf{e}^T \\ &= \alpha(\mathbf{H} + 1/n\mathbf{a}\mathbf{e}^T) + (1-\alpha)1/n\mathbf{e}\mathbf{e}^T \\ &= \alpha\mathbf{H} + (\alpha\mathbf{a} + (1-\alpha)\mathbf{e})1/n\mathbf{e}^T\end{aligned}$$

- Ο **G** είναι τεχνητός
  - Το stationary vector δεν υπάρχει για τον **H**
  - Αλλά υπάρχει για τον **G**

Σύμβολα

- **H**: πολύ αραιός, substochastic πίνακας υπερσυνδέσμων
- **S**: αραιός, στοχαστικός, πιθανώς reducible πίνακας
- **G**: τελείως πυκνός, στοχαστικός, πρωτογενής πίνακας
- **E**: τελείως πυκνός, rank-one πίνακας τηλεμεταφοράς
- n: αριθμός σελίδων στη μηχανή της Google
- α: παράμετρος μεταξύ 0 και 1
- π^T: stationary row vector, PageRank διάνυσμα
- a^T: δυαδικό διάνυσμα dangling nodes

Η μέθοδος του PageRank

$$\pi^{(k+1)T} = \pi^{(k)T}\,\mathbf{G}$$

που είναι απλά η power method εφαρμοζόμενη στον **G**

Το παράδειγμα γραφήματος με 6 κόμβους

$$\mathbf{G} = .9\mathbf{H} + (.9 \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + .1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}) 1/6 (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1)$$

$$\mathbf{G} = \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix}$$

$$\pi^T = (.03721 \quad .05369 \quad .04151 \quad .3751 \quad .206 \quad .2862)$$

Υπολογισμός του διανύσματος PageRank

- Το πρόβλημα μπορεί να περιγραφεί με δυο τρόπους
  - Επίλυση του παρακάτω προβλήματος ιδιοδιανυσμάτων του π^T

$$\pi^T = \pi^T \mathbf{G}$$
$$\pi^T \mathbf{e} = 1$$

  - Επίλυση του γραμμικού ομογενούς συστήματος για το π^T

$$\pi^T (\mathbf{I} - \mathbf{G}) = \mathbf{0}^T$$
$$\pi^T \mathbf{e} = 1$$

Υπολογισμός του διανύσματος PageRank

- Στο πρώτο σύστημα, ο στόχος είναι να βρεθεί το κανονικοποιημένο κυρίαρχο αριστερό ιδοδιάνυσμα που αντιστοιχεί στην κυρίαρχη ιδιοτιμή $\lambda_1=1$
- Στο δεύτερο σύστημα ο στόχος είναι να βρεθεί το κανονικοποιημένο αριστερό null vector του (**I-G**)
- Η εξίσωση κανονικοποίσης υπάρχει για να εγγυηθεί ότι το π^T είναι διάνυσμα πιθανοτήτων

Power method υπολογισμού του PageRank

- Είναι η παλιότερη και απλούστερη μέθοδος εύρεσης της κυρίαρχης (dominant) ιδιοτιμής και ιδιοδιανύσματος ενός πίνακα
- Άρα μπορεί να χρησιμοποιηθεί για εύρεση του stationary vector μιας Markov chain
  - Το stationary vector είναι απλά το κυρίαρχο αριστερό ιδιοδιάνυσμα
- Είναι εξαιρετικά αργή μέθοδος, μεταξύ των Gauss-Seidel, Jacobi, restarted GMRES
- Γιατί χρησιμοποιήθηκε;

Power method υπολογισμού του PageRank

- Είναι προγραμματιστικά απλή
- Εφαρμοζόμενη στον **G** μπορεί να γραφεί ως εφαρμογή στον πολύ αραιό **H**

$$\begin{aligned}\pi^{(k+1)T} &= \pi^{(k)T}\mathbf{G} \\ &= \alpha\pi^{(k)T}\mathbf{S} + \frac{1-\alpha}{n}\pi^{(k)T}\mathbf{e}\mathbf{e}^T \\ &= \alpha\pi^{(k)T}\mathbf{H} + (\alpha\pi^{(k)T}\mathbf{a} + 1 - \alpha)\mathbf{e}^T/n\end{aligned}$$

- Εκτελείται πάνω στον **H** και όχι πάνω στους **S** ή **G**
- Αποθηκεύονται μόνο οι **a**, **e**

Power method υπολογισμού του PageRank

- Οι άλλες μέθοδοι αναγκάζονται να προσπελάσουν τα στοιχεία του πίνακα, ενώ η power method μόνο διαμέσου του πολλαπλασιασμού διανύσματος-πίνακα
- Εκτός από την αποθήκευση του **H** και **a** απαιτεί μόνο την αποθήκευση του π^T και όχι πολλαπλά διανύσματα όπως οι άλλες μέθοδοι
- Απαιτεί πολύ λίγες επαναλήψεις για να επιτευχθεί η σύγκλιση
  - 50-100
- Το ερώτημα που προκύπτει είναι από ποιο/ποιους παράγοντες εξαρτάται/καθορίζεται η σύγκλιση

Ρυθμός σύγκλισης (1/2)

- Ο ασυμπτωτικός ρυθμός σύγκλισης της power method όταν εφαρμόζεται σε κάποιο Markov πίνακα εξαρτάται από το κλάσμα των δυο ιδιοτιμών που έχουν το μεγαλύτερο μέγεθος, λ_1, λ_2
- Για τους στοχαστικούς πίνακες, όπως ο G, ισχύει ότι $\lambda_1 = 1$
- Άρα η σύγκλιση εξαρτάται από την τιμή του λ_2
- Επειδή ο **G** είναι πρωτογενής, ισχύει ότι $|\lambda_2| < 1$
- Η εύρεση του είναι χρονοβόρα, οπότε δεν είναι φρόνιμο να σπαταλήσουμε πόρους για να έχουμε μια εκτίμηση του ρυθμού σύγκλισης

Ρυθμός σύγκλισης (2/2)

- Στις επόμενες διαφάνειες θα δείξουμε ότι εάν οι ιδιοτιμές του **S** είναι $\sigma(\mathbf{S})=\{1,\mu_2,\mu_3,\mu_n\}$ και του **G** είναι $\sigma(\mathbf{G})=\{1,\lambda_2,\lambda_3,\lambda_n\}$, τότε

$$\lambda_k = \alpha\mu_k \quad k=2,3,\ldots n$$

- Η δομή του Παγκοσμίου Ιστού είναι τέτοια που καθιστά πολύ πιθανό να ισχύει ότι $|\mu_2| =1$ (ή $|\mu_2| \approx 1$)
- Άρα $\lambda_2(\mathbf{G})=\alpha$ (ή $\lambda_2(\mathbf{G}) \approx \alpha$)
- Με $\alpha=.85$, σημαίνει ότι μετά από 50 επαναλήψεις $\alpha^{50}=.85^{50} \approx .000296$, δηλ., 2-3 θέσεις ακρίβειας που είναι αρκετά ικανοποιητικές όταν το ranking συνδυάζεται με το περιεχόμενο