Ανάκληση Πληροφορίας

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 11

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)

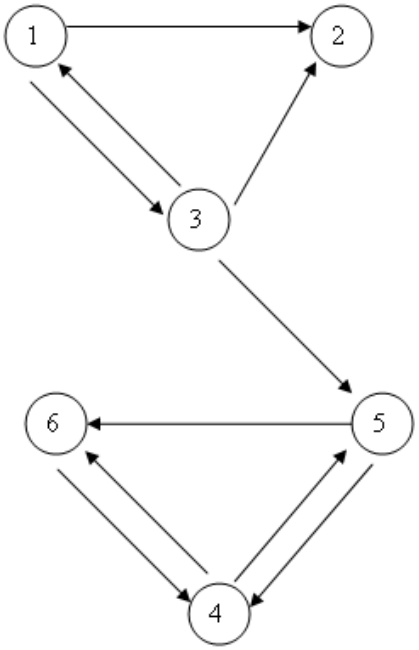
- Έστω ότι $r_{k+1}(P_i)$ είναι το PageRank της σελίδας P_i στην επανάληψη k+1:

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

- 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

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

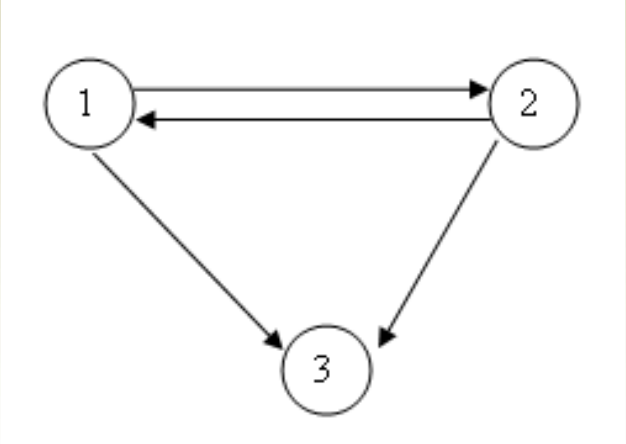
1. Κάθε επανάληψη της προηγούμενης εξίσωσης απαιτεί έναν πολλαπλασιασμό, άρα $O(n^2)$ πολυπλοκότητα
2. Ο H είναι γενικά πολύ αραιός (sparse), άρα
   - Απαιτεί μικρό αποθηκευτικό χώρο
   - Ο πολλαπλασιασμός είναι πιο οικονομικός σε σχέση με το $O(n^2)$
     - Απαιτεί μόνο O(nnz(H)), όπου nnz(H) είναι 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

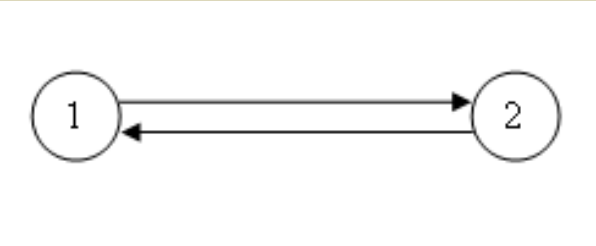
- Αρχικά, η επαναληπτική διαδικασία ξεκίνησε με $\pi^{(0)T}=1/n\mathbf{e}^T$ (όπου $\mathbf{e}^T$ είναι διάνυσμα-γραμμή με όλα 1)
- Προέκυψε το πρόβλημα της **καταβόθρας** (rank sinks)
  - σελίδες που αυξάνουν συνεχώς το PageRank τους
  - Στο παρακάτω παράδειγμα το κόμβος 3, ενώ στο προηγούμενο παράδειγμα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



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

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 $E=1/n\mathbf{e}\mathbf{e}^T$ 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 ιδιοδιάνυσμα 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)

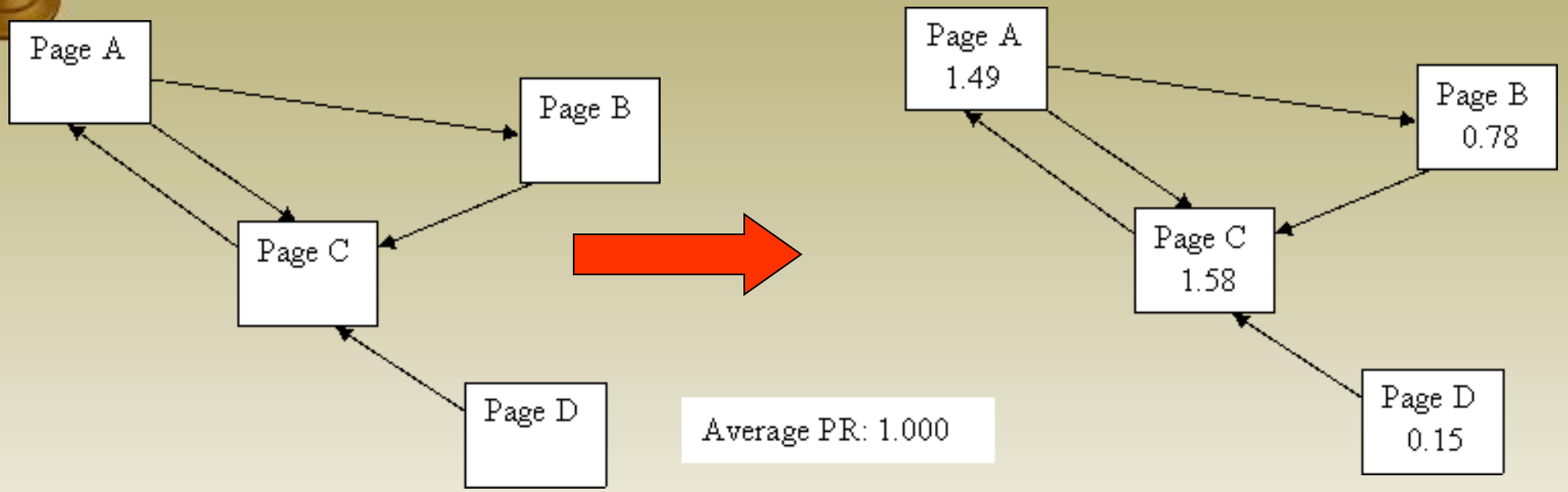
- Στις επόμενες διαφάνειες θα δείξουμε ότι εάν οι ιδιοτιμές του **S** είναι $\sigma(\mathbf{S})=\{1,\mu_2,\mu_3,\mu_n\}$ και του **G** είναι $\sigma(\mathbf{G})=\{1,\lambda_2,\lambda_3,\lambda_n\}$, τότε

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

- 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 α=.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

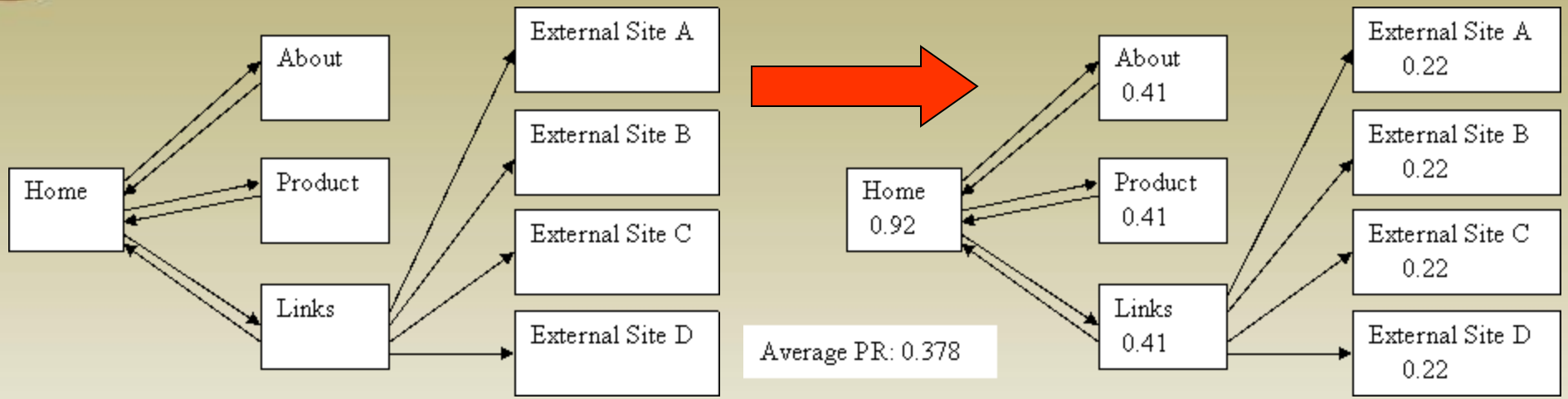
Ο PageRank στην σχεδίαση Web sites

PageRank: Παράδειγμα 1



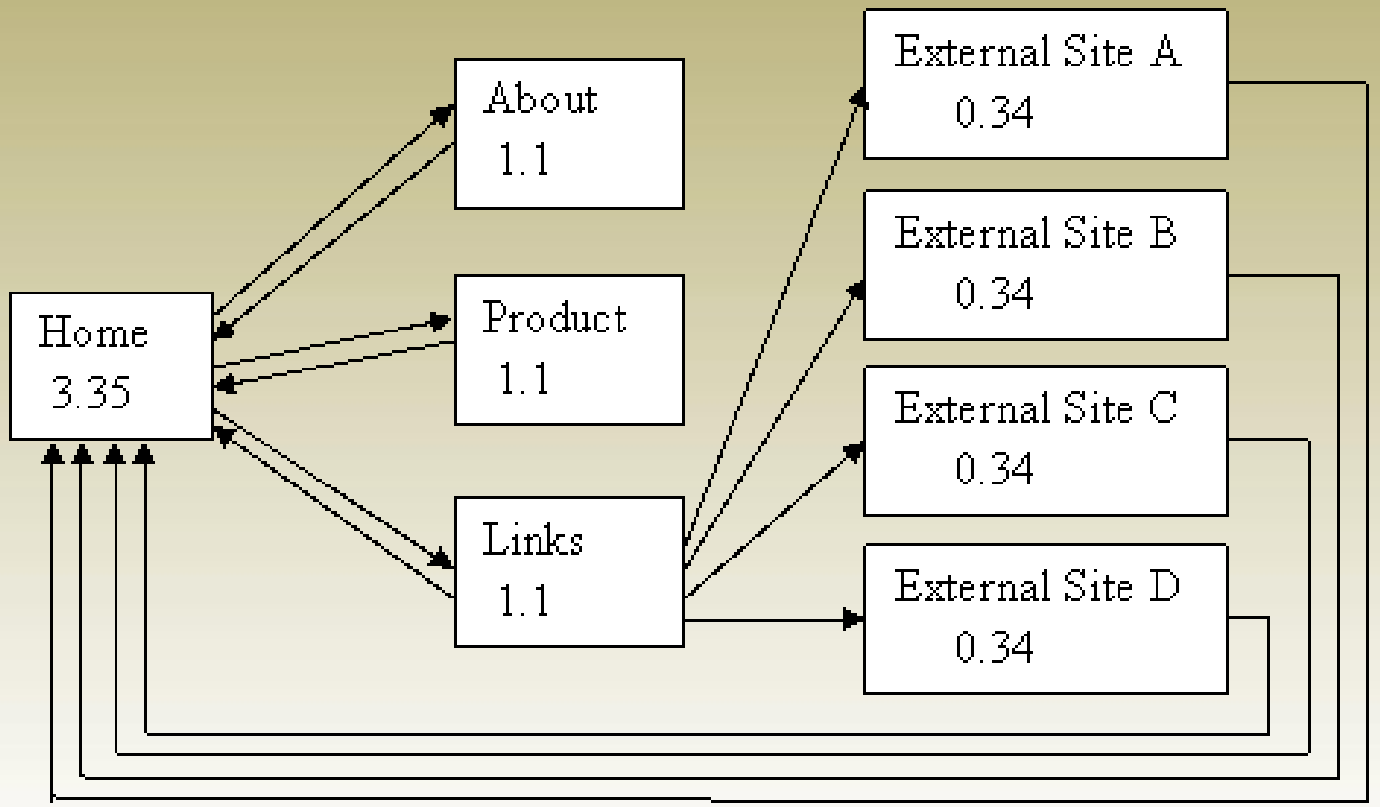
- Χρειάστηκε περίπου 20 επαναλήψεις μέχρι να σταθεροποιηθούν αυτές οι τιμές
- Δείτε την Page D – έχει PR ίσο με of 0.15 παρόλο που δεν έχει εισερχόμενους συνδέσμους. Άρα, για την Page D, απουσία εισερχομένων συνδέσμων σημαίνει ότι η εξίσωση έχει ως εξής: PR(D)=(1-d)+d*(0)= 0.15
- **Παρατήρηση**: Κάθε σελίδα έχει PR τουλάχιστον 0.15 (damping factor)
  - Αλλά αυτό είναι θεωρητικό – υπάρχουν φήμες ότι η Google διαγράφει από τον index της όσες σελίδες δεν έχουν εισερχόμενους συνδέσμους

PageRank: Παράδειγμα 2



- Ως αναμενόταν, η home page έχει το υψηλότερο PR – έχει τους περισσότερους εισερχόμενους συνδέσμους! Αλλά τι συνέβη με το average? Είναι μόνο 0.378!!! Γιατί;
- Οφείλεται στις "external site" pages – τι συμβαίνει με το δικό τους PageRank? Δεν το στέλνουν πουθενά, και έτσι το σπαταλούν!!!

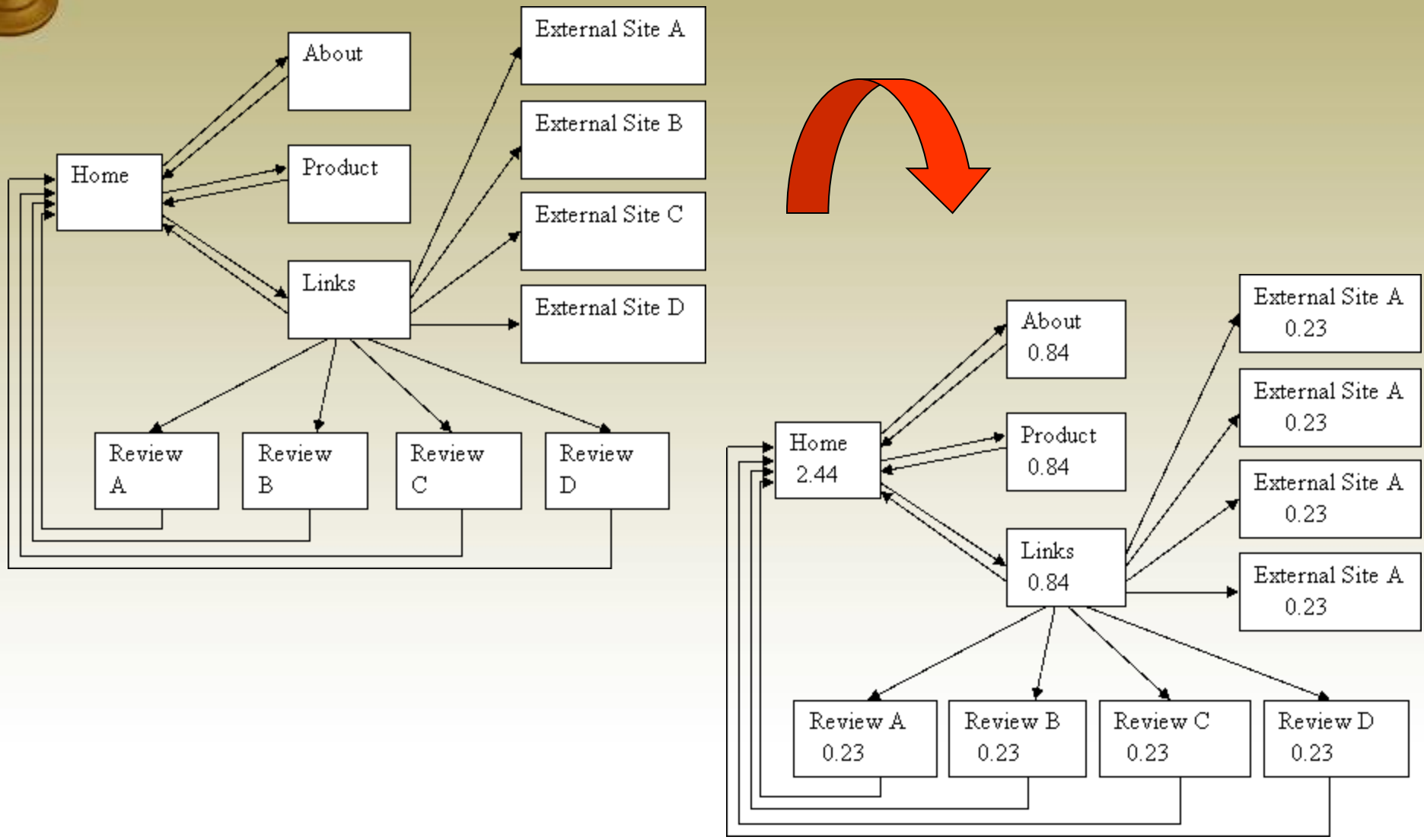
PageRank: Παράδειγμα 3



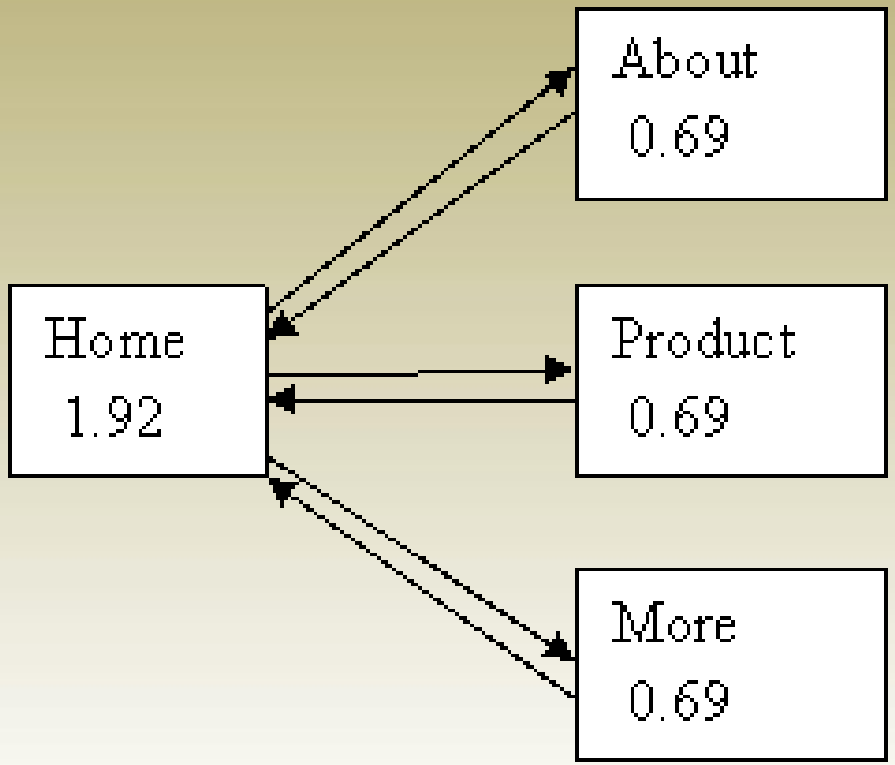
Average PR: 1.000

- Καλύτερη συνδεσμολογία! Δείτε το PR της home page! Όλοι αυτοί οι εισερχόμενοι σύνδεσμοι κάνουν σημαντική διαφορά στο PR της

PageRank: Παράδειγμα 4

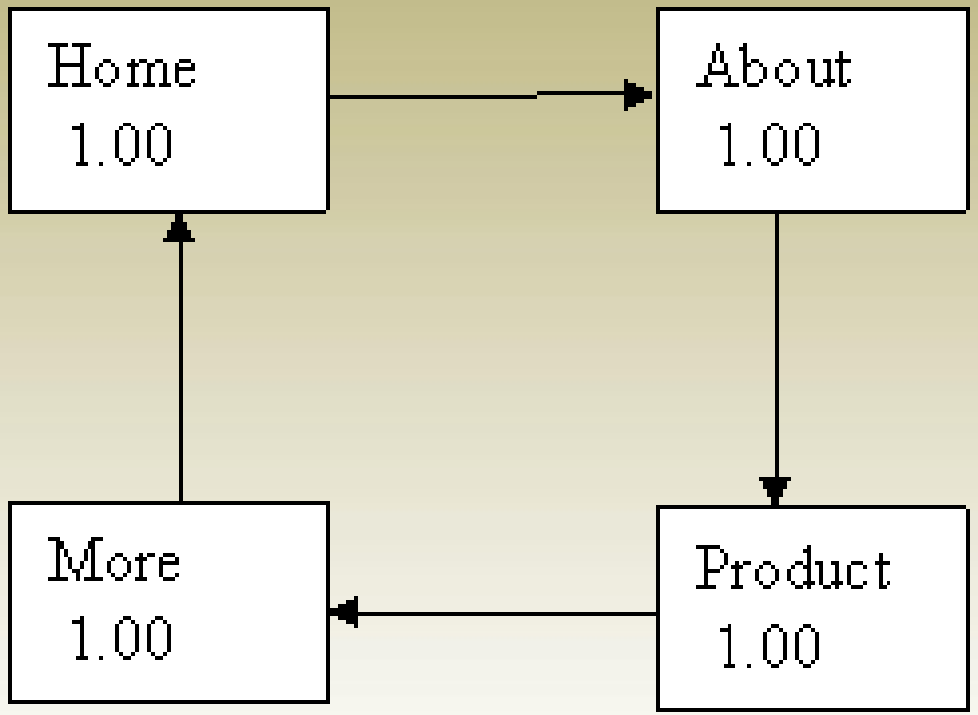


PageRank: Παράδειγμα 5



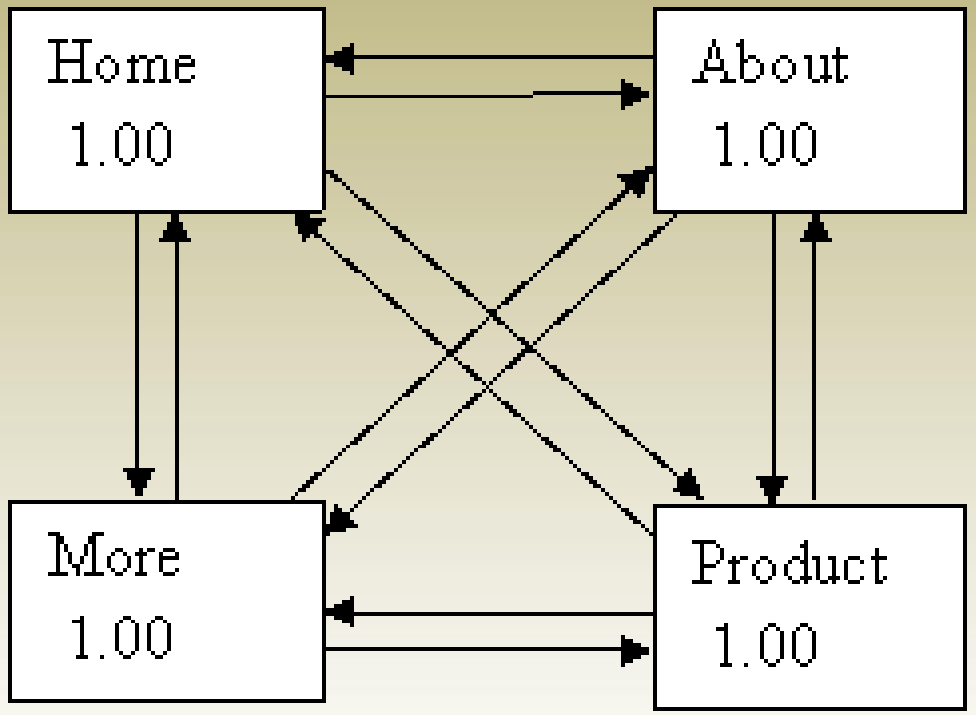
- Η home page έχει περίπου 2.5 φορές το PR των παιδιών της!
- **Παρατήρηση**: Μια ιεραρχία συγκεντρώνει το PR σε μια page

PageRank: Παράδειγμα 6



- Αναμενόμενο. Όλες οι σελίδες έχουν το ίδιο αριθμό εισερχομένων συνδέσμων, και όλες οι pages έχουν την ίδια συνδεσμολογία η μια σε σχέση με την άλλη. Άρα όλες έχουν το ίδιο PR ίσο με 1.0

PageRank: Παράδειγμα 7

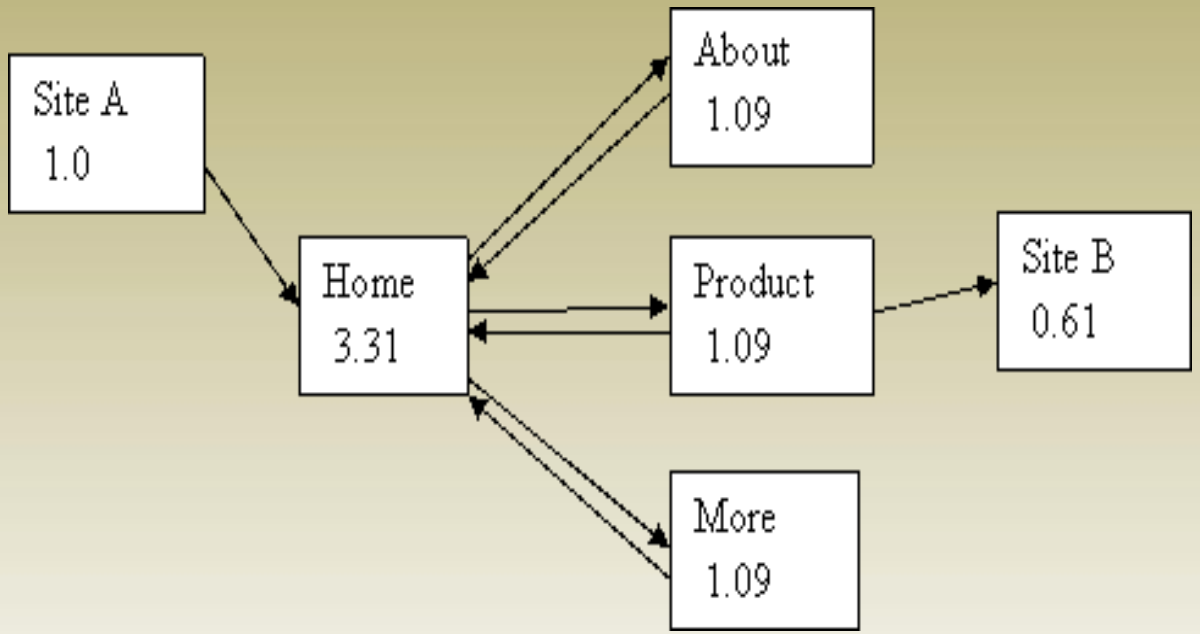


- Ανάλογη κατάσταση με αυτή που περιγράφηκε στην προηγούμενη διαφάνεια

PageRank: Παράδειγμα 8

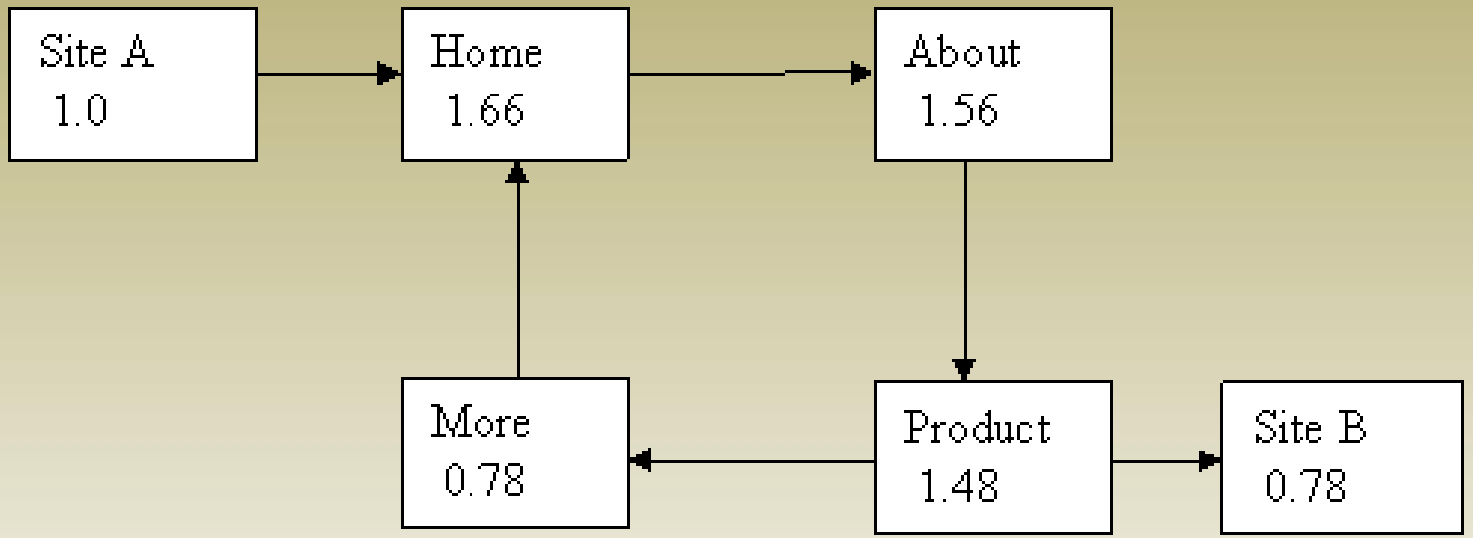
Υποθέτουμε την ύπαρξη ενός external site με πολλές σελίδες και συνδέσμους και ότι μια από τις pages έχει PR ίσο με 1.0

Υποθέτουμε επίσης ότι ο Webmaster μας εκτιμά – υπάρχει μόνο ένας σύνδεσμος από αυτή την page, και αυτός δείχνει στην home page μας



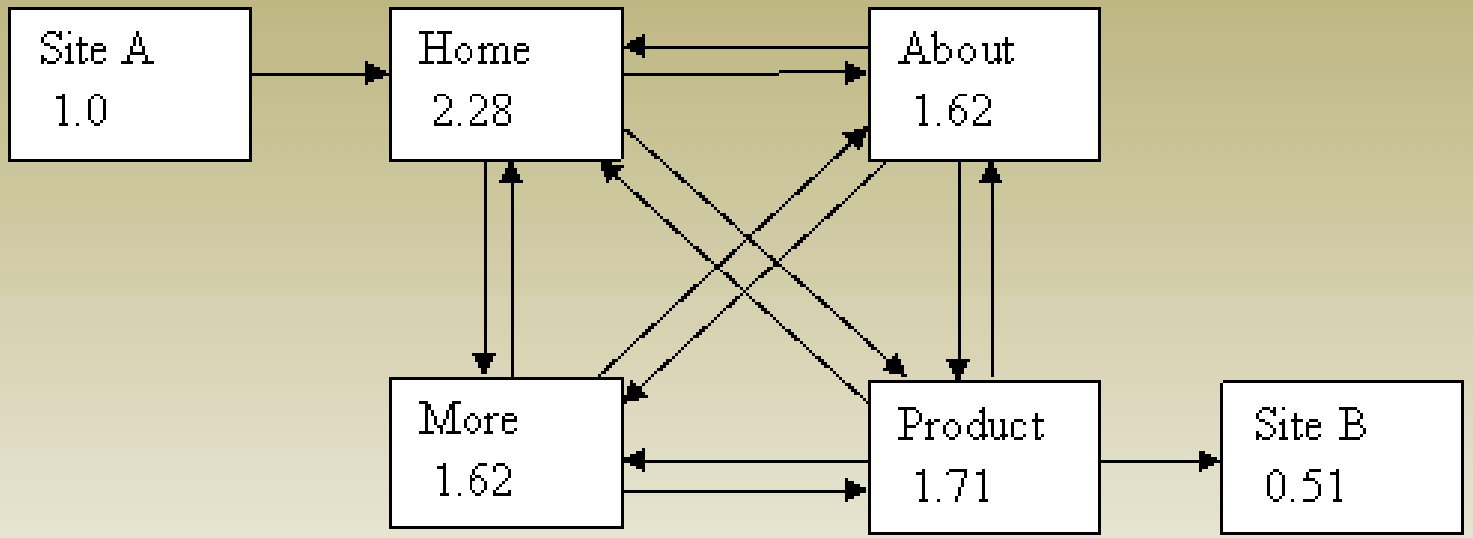
- Στο Παράδειγμα 5 η home page είχε PR ίσο με 1.92, αλλά τώρα είναι 3.31! Γιατί; Όχι μόνο το site A συνεισέφερε 0.85 PR σε εμάς, αλλά η ανάδραση των “About”, “Product” και “More” pages είχε ένα επιπλέον θετικό αποτέλεσμα, αυξάνοντας περαιτέρω το PR της home page!
- **Αρχή**: Ένα καλώς δομημένο site θα ενισχύσει το αποτέλεσμα ενός συνεισφερομένου PR

PageRank: Παράδειγμα 9



- Το PR της home page έχει αυξηθεί κάπως, αλλά τι συνέβη με την "More" page;
- Η "Product" page έχει δυο εξερχόμενους συνδέσμους, έναν προς την "More" και έναν προς κάποιο external site. Ως να θεωρούμε το external Site B ισο-σημαντικό με την δική μας "More" page. Η "More" page παίρνει μόνο το μισο PR από αυτό που είχε πριν – αυτό είναι καλό για το Site B, αλλά κάκκιστο για εμάς

PageRank: Παράδειγμα 10 (1/2)



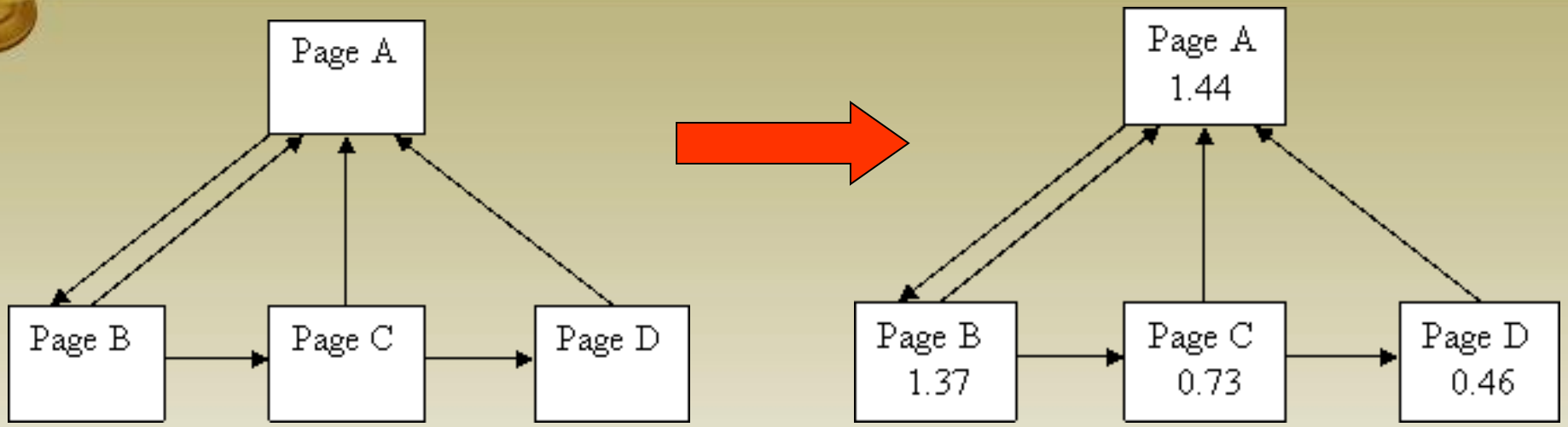
- Καλύτερα τώρα. Η "More" page εξακολουθεί να παίρνει λιγότερο αναλογικά PR, αλλά τώρα η "Product" page έχει κρατήσει τα ¾ του PR εντός του δικού μας site – αντίθετα από ότι στο Παράδειγμα 9 όπου "έδιωχνε" το μισό της PR προς ένα external site!

- Κρατώντας αυτό το μικρό κλάσμα του PR εντός του site μας έχει πολύ καλή επίδραση στην Home Page επίσης – έχει PR ίσο με 2.28, συγκρινόμενο με μόνο 1.66 στο Παράδειγμα 9

PageRank: Παράδειγμα 10 (2/2)

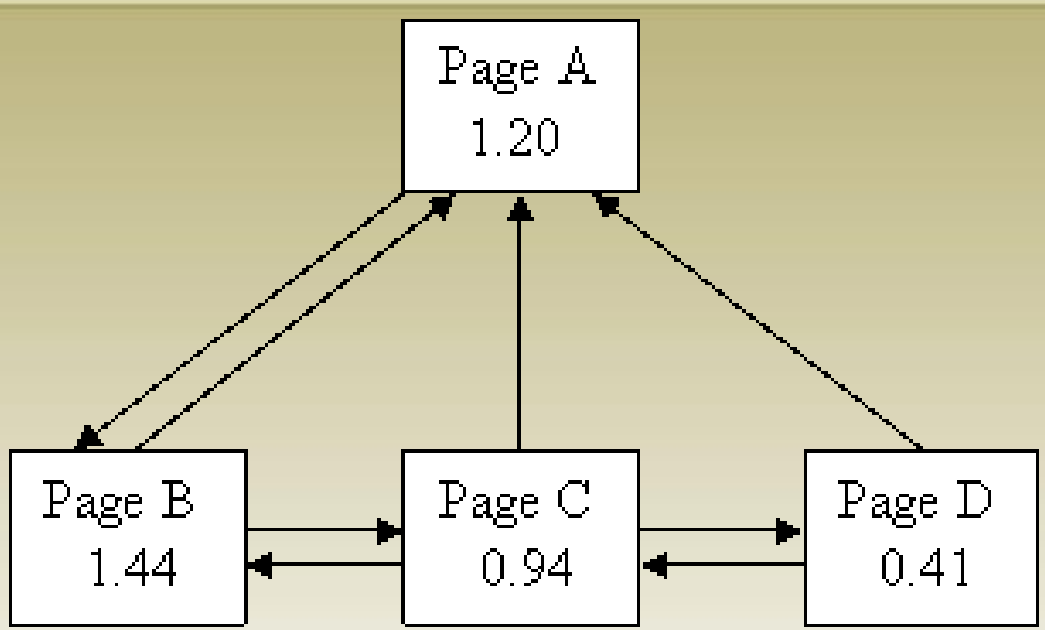
- **Παρατήρηση**: Αυξάνοντας τους εσωτερικούς συνδέσμους στο site μας, μπορούμε να ελαχιστοποιήσουμε την ζημιά στο δικό μας PR, όταν είμαστε αναγκασμένοι να έχουμε συνδέσμους προς external sites
- **Αρχές**:
  - Εάν μια συγκεκριμένη σελίδα είναι πολύ σημαντική: χρησιμοποιήστε μια ιεραρχική δόμηση με την σημαντική σελίδα στην "κορυφή"
  - Όταν μια ομάδα σελίδων μπορεί/πρέπει να περιέχει εξωτερικούς (προς το site μας) συνδέσμους: αυξήστε τον αριθμό των εσωτερικών συνδέσμων ώστε να διατηρήσετε όσο περισσότερο PR γίνεται
  - Όταν μια ομάδα σελίδων δεν περιέχει εξωτερικούς (προς το site μας) συνδέσμους : ο αριθμός των εσωτερικών συνδέσμων δεν έχει κανένα απολύτως αποτέλεσμα στο μέσο PR του site μας. Απλά, ρυθμίστε την συνδεσμολογία ώστε να διευκολύνετε την εμπειρία περιήγησης των επισκεπτών του site

PageRank: Παράδειγμα 11



- Ας προσπαθήσουμε να φτιάξουμε το site μας έτσι ώστε να συγκεντρώσουμε το PR στην home page. Η παραπάνω σχεδίαση φαίνεται καλή: Οι περισσότεροι σύνδεσμοι δείχνουν στην page A, και συνεπώς αναμένουμε να πάρει υψηλό PR
- Όμως είναι πολύ χειρότερο από ότι μια απλή ιεραρχία! Αυτό που συμβαίνει είναι ότι οι pages C και D έχουν ασθενείς εισερχομένους συνδέσμους και έτσι δεν βοηθούν καθόλου την page A!
- **Αρχή**: Η προσπάθεια "λαθροχειρίας" του PR είναι κάπως δύσκολη (δείτε μελλοντική διάλεξη για link spamming), αλλά όχι ακατόρθωτη

PageRank: Παράδειγμα 12 (1/2)

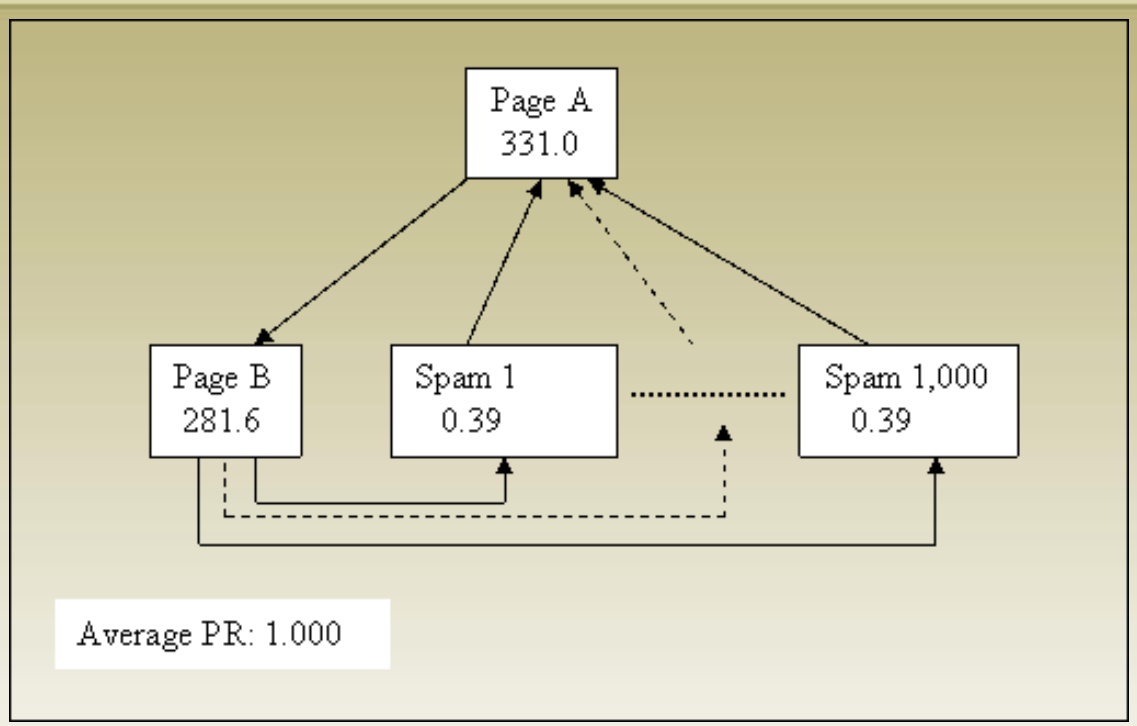


- Ένας κοινός τρόπος σχεδίασης των long documentation είναι να διαιρούμε το "έγγραφο" σε πολλές pages με ένα "Previous" και "Next" σύνδεσμο σε κάθε σελίδα, συν έναν σύνδεσμο πίσω στην home page. Τότε η home page αρκεί να δείχνει στην πρώτη σελίδα του "εγγράφου"
- Σε αυτό το απλό παράδειγμα, όπου υπάρχει μόνο ένα "έγγραφο", η πρώτη σελίδα του "εγγράφου" έχει υψηλότερο PR από αυτό της Home Page! Αυτό συμβαίνει γιατί η page B παίρνει όλο το PR της page A, αλλά η page A παίρνει μόνο ένα ποσοστό του PR των pages B, C και D

PageRank: Παράδειγμα 12 (2/2)

- **Αρχή:** για να δώσουμε στους χρήστες του site μας ευχάριστη εμπειρία περιήγησης, ίσως χρειαστεί να κακοποιήσουμε το PR. Αυτό δεν είναι απαραιτήτως κακό. Εάν το site μας είναι **χρήσιμο**, τότε πολλοί Webmasters θα βάλουν συνδέσμους από τα sites τους προς το δικό μας, και έτσι θ' αναπληρώσουμε το χαμένο PR
- Μπορείτε να διακρίνετε την τάση μεταξύ αυτού και του προηγουμένου παραδείγματος; Καθώς προσθέτετε περισσότερα "εσωτερικούς" συνδέσμους σε ένα site, αυτό βαίνει προς μια Fully Meshed τοπολογία όπου κάθε σελίδα αποκτά το ιδιο average PR
- **Παρατήρηση:** Καθώς προστίθενται περισσότεροι "εωτερικοί" σύνδεσμοι, το PR θα κατανεμηθεί πιο ομοιόμορφα μεταξύ των σελίδων

PageRank: Παράδειγμα 13



- Ας δούμε εάν μπορούμε να βάλουμε 1,000 pages να δείχνουν στην home page μας, και να έχουμε μόνο έναν σύνδεσμο να φεύγει από αυτήν
- Αυτές οι spam pages είναι "άχρηστες" αλλά έχουν έντονο προσθετικό ενδιαφέρον για την Α!
- **Παρατήρηση**: δεν έχει σημασία πόσες pages απαρτίζουν το site σας, το μέσο PR θα είνια πάντα 1.0 στην καλύτερη περίπτωση. Αλλά, μια τοπολογία ιεραρχίας θα συγκεντρώσει το PR στην home page!

Συμπεράσματα (1/2)

- Από το άρθρο[1] των Brin και Page (συνιδρυτές της Google), το μέσο πραγματικό PR όλων των pages στον inverted index είναι 1.0!
- Συνεπώς, εάν προσθέτετε pages στο site που χτίζετε, το συνολικό PR θα ανεβαίνει κατά 1.0 για κάθε page (φυσικά μόνο εάν αλληλοσυνδέετε τις pages για να δουλέψει η εξίσωση), αλλά το average θα παραμένει το ίδιο
- Εάν επιθυμείτε να συγκεντρώσετε το PR σε μια (ή σε λίγες σελίδες), τότε μια τοπολογία ιεραρχίας θα το επιτύχει. Εάν επιθυμείτε να ισο-διανείμετε το PR μεταξύ των σελίδων, τότε το "fully meshing" θα το επιτύχει

[1] http://www.sciencedirect.com/science/article/pii/S016975529800110X

Συμπεράσματα (2/2)

- Η προσέλκυση εισερχομένων “εξωτερικών” συνδέσμων στο site σας είναι ο μόνος τρόπος ν’ αυξήσετε το average PR. Το πώς κατανέμεται αυτό μεταξύ των σελίδων του site σας, εξαρτάται από την εσωτερική συνδεσμολογία του site σας, και από το ποιες σελίδες έχουν προσελκύσει τους “εξωτερικούς” συνδέσμους
- Εάν εσείς έχετε εξωτερικούς συνδέσμους προς άλλα sites, τότε το average PR του site σας θα ελαττωθεί. Το πόσο θα ελαττωθεί και από ποιες σελίδες, εξαρτάται πάλι από την συνδεσμολογία