Ανάκληση Πληροφορίας

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 15η

Παράμετροι του μοντέλου PageRank

Η παράμετρος α (1/2)

- Η παράμετρος αυτή ελέγχει στην ουσία την προτεραιότητα που δίνεται στη δομή των υπερσυνδέσμων ή στην τηλεμεταφορά
- Είδαμε στην προηγούμενη διαφάνεια ότι οι Brin & Page πρότειναν τιμή .85 για την παράμετρο αυτή
- Γιατί αυτήν την τιμή;
- Ποια είναι η επίδραση του α στο πρόβλημα του PageRank;
- Με α=.5, τότε η επαναληπτική μέθοδος χρειάζεται μόνο 34 επαναλήψεις για να συγκλίνει σε μια ακρίβεια 10^{-10} !!
- Όμως αυτό σημαίνει ότι η τεχνητά εισαχθείσα έννοια της τηλεμεταφοράς θα είναι ίσης σημαντικότητας με τη δομή των υπερσυνδέσμων !?

Η παράμετρος α (2/2)

- Για α=1.0, οι αριθμός των επαναλήψεων για σύγκλιση γίνεται απαγορευτικός
- Ακόμα και για α=.85 απαιτούνται μερικές ημέρες για να επιτευχθεί η σύγκλιση όταν οι πίνακες είναι του μεγέθους του Παγκοσμίου Ιστού
- Απλώς το α=.85 επιτυγχάνει ένα αποδεκτό tradeoff
- Πέρα από αυτό όμως, η παράμετρος ελέγχει και την ευαισθησία του διανύσματος PageRank
- Για τιμές του α κοντά σε 1, τότε ακόμα και μικρές αλλαγές στη δομή του Web επηρεάζουν σημαντικά τις τιμές PageRank των σελίδων

α	num. of iterations
.5	34
.75	81
.8	104
.85	142
.9	219
.95	449
.99	2292
.999	23015

Ο πίνακας υπερσυνδέσμων **H**

- Διάφορες προσαρμογές μπορεί να γίνουν πάνω στον H
- Στην βασική υλοποίηση, κάθε εξερχόμενος σύνδεσμος έχει το ίδιο "βάρος/σημαντικότητα"
- Παρόλο που η τακτική αυτή είναι δημοκρατική, εύκολη στην υλοποίηση, εντούτοις δεν είναι η κατάλληλη για τα rankings
- Στην πραγματικότητα, ο random surfer δεν διαλέγει τυχαία με την ίδια πιθανότητα ποιον σύνδεσμο θα ακολουθήσει, αλλά λαμβάνει υπόψη του το πλούσιο περιεχόμενο των σελίδων όπου θα πάει, αλλά και το κείμενο πάνω στους υπερσυνδέσμους
- Έτσι, αντί για την υπόθεση του random surfer, έχουμε τον **intelligent surfer**

Παράδειγμα προσαρμοσμένου πίνακα **H**

- Πώς αποφασίζουμε με ποιο τρόπο θα αναθέσουμε διαφορετικά βάρη στους εξερχόμενους υπερσυνδέσμους;
- Από τα access logs!
- Παράδειγμα: Από την P_1 είναι δυο φορές πιο πιθανό να πάμε στην P_2 παρά στην P_3
- Προφανώς όλες οι παρόμοιες μέθοδοι θα είναι ευρεστικές
- Για παράδειγμα, τα στοιχεία $\mathbf{H_{45}}$ και $\mathbf{H_{46}}$ μπορούν να προσδιοριστούν με βάση την ομοιότητα (cosine similarity) μεταξύ των σελίδων P_4 με την P_5 και P_6
- Για το γράφημα με τους 6 κόμβους ο νέος πίνακας H θα μετατραπεί στον ακόλουθο:

Παράδειγμα προσαρμοσμένου πίνακα **H**

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

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

Ο πίνακας τηλεμεταφοράς **E** (1/3)

- Μια από τις πρώτες προσαρμογές ήταν ότι αντί για τη χρήση του $1/n\mathbf{e}\mathbf{e}^T$ προτιμήθηκε ο πίνακας $\mathbf{e}\mathbf{v}^T$
- Το $\mathbf{v}^T$ με $\mathbf{v}^T > 0$, είναι ένα διάνυσμα πιθανοτήτων που ονομάζεται **personalization** ή **teleportation διάνυσμα**
- Αφού το $\mathbf{v}^T$ είναι διάνυσμα πιθανοτήτων με θετικά στοιχεία, κάθε κόμβος είναι συνδεδεμένος με κάθε άλλο κόμβο, άρα ο **G** είναι πρωτογενής
- Χρησιμοποιώντας το $\mathbf{v}^T$ αντί για το $1/n\mathbf{e}^T$ σημαίνει ότι οι πιθανότητες τηλεμεταφοράς δεν είναι πλέον ομοιόμορφες

Ο πίνακας τηλεμεταφοράς **E** (2/3)

- Άρα για κάθε τηλεμεταφορά, ο surfer δεν επιλέγει ομοιόμορφα σε ποια σελίδα θα πάει, αλλά καθοδηγείται από το διάνυσμα $\mathbf{v}^T$
- Αυτή η μετατροπή ευτυχώς δεν καταστρέφει τα πλεονεκτήματα της power method
- Όταν $\mathbf{G}=\alpha\mathbf{S}+(1-\alpha)\mathbf{e}\mathbf{v}^T$, τότε η power method γίνεται:

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

Ο πίνακας τηλεμεταφοράς **E** (3/3)

- Αυτή η αλλαγή δεν έχει καμία επίδραση πάνω
  - στο ρυθμό σύγκλισης
  - στον πολλαπλασιασμό διανύσματος με αραιό πίνακα
  - στις μικρές αποθηκευτικές απαιτήσεις
- Όμως, αλλάζει το ίδιο το διάνυσμα PageRank!!
- Αυτό δεν είναι μειονέκτημα !?
- Δεν είναι απαραίτητο ότι σε όλους μας "ταιριάζει"το ίδιο ranking
- Άλλωστε, παρέχει μια ευελιξία ώστε ανάλογα τις ανάγκες μας να προσαρμόζουμε απλά το $\mathbf{v}^T$

Προσωποποίηση του PageRank

- Η προσωποποίηση αλλάζει το διάνυσμα PageRank, από query-independet και user-independent σε user-dependent και πιο δύσκολο στον υπολογισμό
- Στην θεωρία είναι ωραία η προσωποποίηση, αλλά στην πράξη είναι δύσκολα εφαρμόσιμη
  - Κάθε $\mathbf{\pi}^T$ απαιτεί μερικές ημέρες για τον υπολογισμό του
- Οπότε, αφού επικρατεί η άποψη ότι η προσωποποιημένη αναζήτηση είναι η μελλοντική τάση στις μηχανές αναζήτησης, αρκετοί δημιούργησαν ψευδο-προσωποποιημένα διανύσματα PageRank
  - Δεν απευθύνονται σε κάθε χρήστη, αλλά σε ομάδες χρηστών

Topic-sensitive PageRank (1/3)

- Δημιουργία ενός πεπερασμένου αριθμού PageRank διανυσμάτων $\mathbf{\pi}^T(\mathbf{v}_i^T)$, κάθε ένα από αυτά πολωμένο ως προς κάποια συγκεκριμένο θέμα
- Ποια θέματα επιλέχθηκαν;
- Ο Taher Haveliwala επέλεξε τα 16 πρώτα από το Open Directory Project (ODP)
- Τα 16 πολωμένα διανύσματα προϋπολογίζονται
- Το ζήτημα είναι να τα συνδυάσουμε αποτελεσματικά κατά την ερώτηση του χρήστη

Topic-sensitive PageRank (2/3)

- Ο Taher Haveliwala έφτιαξε έναν κυρτό συνδυασμό αυτών ως εξής

$$\mathbf{\pi}^T = \beta_1 \mathbf{\pi}^T(\mathbf{v}_1^T) + \beta_2 \mathbf{\pi}^T(\mathbf{v}_2^T) + \ldots + \beta_{16} \mathbf{\pi}^T(\mathbf{v}_{16}^T)$$

όπου $\Sigma\beta_i = 1$

- Για παράδειγμα, η ερώτηση *science project ideas* εμπίπτει μεταξύ των εξής κατηγοριών του ODP:
  - Κατηγορία 7: Kids και Teens
  - Κατηγορία 10: Reference
  - Κατηγορία 12: Science
- Προφανώς τα αντίστοιχα διανύσματα αυτών των κατηγοριών πρέπει να πάρουν μεγαλύτερο βάρος ή ίσως και όλο το βάρος

Topic-sensitive PageRank (3/3)

- Για τον υπολογισμό των βαρών χρησιμοποιήθηκε ένας classifier Bayes
- Όταν υπολογιστεί το topic-sensitive score, συνδυάζεται με το αντίστοιχο content score

- Ο Jeh Glen, Taher Haveliwala & Serendap Kamvar δημιούργησαν το καλοκαίρι του 2003 την εταιρεία Kaltix για να προωθήσουν την ιδέα του personalized PageRank, και τελικά η εταιρεία τους αγοράστηκε το Σεπτέμβριο του 2003 από την Google
- Τον Μάρτιο του 2004, η Google προώθησε την προσωποποίηση http://labs.google.com/personalized

Το φάσμα του personalized πίνακα **G** (1/4)

- <u>ΘΕΩΡΗΜΑ</u>: Εάν το φάσμα (ιδιοτιμές) του στοχαστικού πίνακα **S** είναι $\{1,\lambda_2,\lambda_3,\ldots,\lambda_n\}$, τότε το φάσμα του personalized πίνακα Google $\mathbf{G}=\alpha\mathbf{S}+(1-\alpha)\mathbf{e}\mathbf{v}^T$ είναι $\{1,\alpha\lambda_2,\alpha\lambda_3,\ldots,\alpha\lambda_n\}$, όπου το $\mathbf{v}^T$ είναι ένα διάνυσμα πιθανοτήτων

Το φάσμα του personalized πίνακα **G** (2/4)

- Αφού ο **S** είναι στοχαστικός, τότε το (1,**e**) είναι ένα ζεύγος του **S**
- Έστω ότι **Q** = (**e** **X**) είναι μη ιδιόμορφος (non-singular) πίνακας που έχει το ιδιοδιάνυσμα **e** ως πρώτη στήλη του
- Έστω ότι $$\mathbf{Q}^{-1} = \begin{pmatrix} \mathbf{y}^T \\ \mathbf{Y}^T \end{pmatrix}$$
- Τότε $$\mathbf{Q}^{-1}\mathbf{Q} = \begin{pmatrix} \mathbf{y}^T\mathbf{e} & \mathbf{y}^T\mathbf{X} \\ \mathbf{Y}^T\mathbf{e} & \mathbf{Y}^T\mathbf{X} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \mathbf{0} & \mathbf{I} \end{pmatrix}$$
- Απ' εδώ παίρνουμε δυο χρήσιμες ταυτότητες
  - $\mathbf{y}^T\mathbf{e}=1$
  - $\mathbf{Y}^T\mathbf{e}=\mathbf{0}$

Το φάσμα του personalized πίνακα **G** (3/4)

- Ως συνέπεια, ο μετασχηματισμός ομοιότητας

$$\mathbf{Q}^{-1}\mathbf{S}\mathbf{Q} = \begin{pmatrix} \mathbf{y}^T\mathbf{e} & \mathbf{y}^T\mathbf{S}\mathbf{X} \\ \mathbf{Y}^T\mathbf{e} & \mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{y}^T\mathbf{S}\mathbf{X} \\ 0 & \mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix}$$

φανερώνει ότι ο $\mathbf{Y}^T\mathbf{S}\mathbf{X}$ περιέχει τις υπόλοιπες ιδιοτιμές του **S**, $\lambda_2,\lambda_3,\ldots,\lambda_n$

Το φάσμα του personalized πίνακα **G** (4/4)

- Εφαρμόζοντας τον μετασχηματισμό ομοιότητας στον $\mathbf{G}=\alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$

$$\begin{aligned}
\mathbf{Q}^{-1}(\alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T)\mathbf{Q} &= \alpha\mathbf{Q}^{-1}\mathbf{S}\mathbf{Q} + (1-\alpha)\mathbf{Q}^{-1}\mathbf{e}\mathbf{v}^T\mathbf{Q} \\
&= \begin{pmatrix} \alpha & \alpha\mathbf{y}^T\mathbf{S}\mathbf{X} \\ 0 & \alpha\mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix} + (1-\alpha)\begin{pmatrix} \mathbf{y}^T\mathbf{e} \\ \mathbf{Y}^T\mathbf{e} \end{pmatrix}\begin{pmatrix} \mathbf{v}^T\mathbf{e} & \mathbf{v}^T\mathbf{X} \end{pmatrix} \\
&= \begin{pmatrix} \alpha & \alpha\mathbf{y}^T\mathbf{S}\mathbf{X} \\ 0 & \alpha\mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix} + \begin{pmatrix} (1-\alpha) & (1-\alpha)\mathbf{v}^T\mathbf{X} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \\
&= \begin{pmatrix} 1 & \alpha\mathbf{y}^T\mathbf{S}\mathbf{X} + (1-\alpha)\mathbf{v}^T\mathbf{X} \\ \mathbf{0} & \alpha\mathbf{Y}^T\mathbf{S}\mathbf{X} \end{pmatrix}
\end{aligned}$$

- Επομένως, οι ιδιοτιμές του $\mathbf{G}=\alpha\mathbf{S} + (1-\alpha)\mathbf{e}\mathbf{v}^T$ είναι οι $\{1,\alpha\lambda_2,\alpha\lambda_3,\ldots,\alpha\lambda_n\}$

Ευαισθησία του PageRank

Ευαισθησία του PageRank: Εισαγωγικά

- Η ευαισθησία του PageRank μπορεί να αναλυθεί εξετάζοντας ξεχωριστά κάθε παράμετρο του πίνακα Google
- Στην προηγούμενη διάλεξη δώσαμε έμφαση στις τρεις παραμέτρους που επηρεάζουν τον πίνακα **G**
  - Την παράμετρο α
  - Τον πίνακα υπερσυνδέσμων **H**
  - Το διάνυσμα προσωποποίησης $\mathbf{v}^T$
- Στην παρούσα διάλεξη θα μελετήσουμε την εξάρτηση του PageRank σε σχέση με κάθε μια από αυτές τις παραμέτρους

Ευαισθησία του PageRank σε σχέση με το α

- Θα χρησιμοποιήσουμε την έννοια της παραγώγου για να μελετήσουμε το αποτέλεσμα των αλλαγών του α πάνω στο $\mathbf{\pi}^T$
- Η παράγωγος του $\mathbf{\pi}^T$ σε σχέση με το α, δηλ., $d\mathbf{\pi}^T(\alpha)/d\alpha$, μας λέει πόσο μεταβάλλονται τα στοιχεία του διανύσματος PageRank $\mathbf{\pi}^T$ όταν το α μεταβάλλεται ελαφρά
- Εάν το στοιχείο j του $d\mathbf{\pi}^T(\alpha)/d\alpha$, που το συμβολίζουμε με $d\mathbf{\pi}_j(\alpha)/d\alpha$, είναι μεγάλο σε τιμή, τότε μπορούμε να συμπεράνουμε ότι καθώς το α μεταβάλλεται ελαφρά, το π είναι πολύ ευαίσθητο σε μικρές αλλαγές του α

Ευαισθησία του PageRank σε σχέση με το α

- Το πρόσημα των παραγώγων δίνουν επίσης σημαντική πληροφορία: εάν $d\boldsymbol{\pi}_j(\alpha)/d\alpha > 0$, τότε μικρές αλλαγές στην τιμή του α, θα σημαίνουν ότι η PageRank τιμή της σελίδας P_j αυξάνουν
- Είναι σημαντικό να έχουμε υπόψη μας ότι το $d\boldsymbol{\pi}^T(\alpha)/d\alpha$ είναι μόνο μια *προσέγγιση* τού πώς μεταβάλλονται τα στοιχεία του $\boldsymbol{\pi}^T$ όταν αλλάζει το α και ΔΕΝ περιγράφουν επακριβώς το πώς μεταβάλλονται
- Παρόλο που στο α δίνεται συνήθως η τιμή 0.85, θεωρητικά μπορεί να πάρει τιμή στο $(0 < \alpha < 1)$
- Φυσικά, ο **G** εξαρτάται από το α, και συνεπώς $\mathbf{G}(\alpha)=\alpha\mathbf{S}+(1-\alpha)\mathbf{e}\mathbf{v}^T$

Ευαισθησία του PageRank σε σχέση με το α

- Συνεπώς, με τη παράγωγο μπορούμε να μελετήσουμε το ρυθμό μεταβολής του $\boldsymbol{\pi}^T$ σε σχέση με μικρές μεταβολές του α
- Πρώτα όμως πρέπει να είμαστε βέβαιοι ότι η παράγωγος είναι καλά ορισμένη
- Είδαμε ότι η κατανομή του $\boldsymbol{\pi}^T(\alpha)$ είναι το αριστερό ιδιοδιάνυσμα του $\mathbf{G}(\alpha)$, αλλά τα ιδιοδιανύσματα δεν είναι κατ' ανάγκη παραγωγίσιμα ούτε κατ' ανάγκη συνεχείς συναρτήσεις των στοιχείων του $\mathbf{G}(\alpha)$
- Το επόμενο θεώρημα μας εφοδιάζει με το απαραίτητο υπόβαθρο σε σχέση με την προϋπόθεση ύπαρξης της παραγώγου

Ύπαρξη παραγώγου διανύσματος PageRank

- **<u>ΘΕΩΡΗΜΑ</u>**. Το διάνυσμα PageRank δίνεται από το

$$\pi^T(\alpha) = \frac{1}{\sum_{i=1}^{n} D_i(\alpha)} (D_1(\alpha), D_2(\alpha), \ldots, D_n(\alpha))$$

όπου το $D_i(\alpha)$ είναι η i-οστή κύρια μικρή ορίζουσα τάξης n-1 του $\mathbf{I}$-$\mathbf{G}(\alpha)$.

Επειδή κάθε κύρια μικρή (principal minor) $D_i(\alpha) > 0$ είναι απλά ένα άθροισμα γινομένων αριθμών του $\mathbf{I}$-$\mathbf{G}(\alpha)$, προκύπτει ότι κάθε συνιστώσα του $\boldsymbol{\pi}^T(\alpha)$ είναι παραγωγίσιμη συνάρτηση του α στο διάστημα (0,1)

Ύπαρξη παραγώγου διανύσματος PageRank

- **Απόδειξη**. Έστω ότι $\mathbf{G}=\mathbf{G}(\alpha)$, $\boldsymbol{\pi}^T(\alpha)=\boldsymbol{\pi}^T$, $D_i=D_i(\alpha)$, και θέτουμε $\mathbf{A}=\mathbf{I}-\mathbf{G}$
- Εάν με adj(**A**) συμβολίσουμε τον ανάστροφο του πίνακα των συμπαραγόντων (cofactors), που συχνά αποκαλείται *adjugate* ή *adjoint*, τότε

$$\mathbf{A}[\text{adj}(\mathbf{A})] = \mathbf{0} = [\text{adj}(\mathbf{A})]\mathbf{A}$$

- Από το θεώρημα των Perron-Frobenius προκύπτει ότι *rank*(**A**)=n-1, και ως αποτέλεσμα ότι *rank*(adj(A))=1
- Επιπλέον, το ίδιο θεώρημα εγγυάται ότι κάθε στήλη του [adj(**A**)] είναι πολλαπλάσιο του **e**, και συνεπώς [adj(**A**)] $=\mathbf{e}\mathbf{w}^T$, για κάποιο διάνυσμα w

Ύπαρξη παραγώγου διανύσματος PageRank

- **Απόδειξη** (συνέχεια)
- Αλλά, $[\text{adj}(\mathbf{A})]_{ii} = D_i$, και έτσι $\mathbf{w}^T=(D_1, D_2, .., D_n)$
- Όμοια, η σχέση $[\text{adj}(\mathbf{A})]\mathbf{A} = \mathbf{0}$ εγγυάται ότι κάθε γραμμή του του $[\text{adj}(\mathbf{A})]$ είναι πολλαπλάσιο του $\boldsymbol{\pi}^T$, και επομένως $\mathbf{w}^T=\alpha\boldsymbol{\pi}^T$ για κάποιο α
- Το α αυτό δεν μπορεί να είναι μηδέν, γιατί διαφορετικά $[\text{adj}(\mathbf{A})] = \mathbf{0}$, το οποίο είναι αδύνατο
- Επομένως, $\mathbf{w}^T\mathbf{e}= \alpha \neq 0$, και $\mathbf{w}^T/(\mathbf{w}^T\mathbf{e}) = \mathbf{w}^T/\alpha = \boldsymbol{\pi}^T$ ╣

Άνω όριο των συνιστωσών του PageRank

- **ΘΕΩΡΗΜΑ**. Εάν $\boldsymbol{\pi}^T(\alpha) = (\pi_1(\alpha), \pi_2(\alpha), \ldots, \pi_n(\alpha))$ είναι το διάνυσμα PageRank, τότε:

$$\left|\frac{d\pi_j(\alpha)}{d\alpha}\right| \leq \frac{1}{1-\alpha}, \qquad \forall j = 1, 2, \ldots, n$$

και το άνω όριο του αθροίσματος των συνιστωσών, δηλ., η *1-norm*, είναι:

$$\left|\left|\frac{d\pi^T(\alpha)}{d\alpha}\right|\right|_1 \leq \frac{2}{1-\alpha}$$

Σχόλια για το προηγούμενο θεώρημα

- Η χρησιμότητα του προηγούμενου θεωρήματος περιορίζεται στις μικρές τιμές του α
- Δηλαδή, για μικρές τιμές του α, η τιμές PageRank των αντίστοιχων ιστοσελίδων δεν είναι εξαιρετικά ευαίσθητες ως συνάρτηση του α
- Καθώς όμως το α πλησιάζει στο 1, το άνω όριο του 1/(1-α) τείνει στο άπειρο. Αυτό το όριο γίνεται σταδιακά άχρηστο, γιατί δεν υπάρχει καμία εγγύηση ότι είναι εφικτό
- Όμως οι μεγαλύτερες τιμές του α είναι αυτές που έχουν σημασία, γιατί δίνουν προτεραιότητα στην πραγματική δομή των υπερσυνδέσμων του Web
- Συνεπώς απαιτείται μεγαλύτερη ανάλυση για να αντιληφθούμε το βαθμό ευαισθησίας του PageRank στις μεγαλύτερες τιμές του α

Ευαισθησία του PageRank σε "μεγάλα" α

- **ΘΕΩΡΗΜΑ**. Εάν $\boldsymbol{\pi}^T(\alpha)$ είναι το PageRank διάνυσμα του πίνακα Google $\mathbf{G}=\alpha\mathbf{S}+(1-\alpha)\mathbf{e}\mathbf{v}^T$, τότε:

$$\frac{d\pi^T(\alpha)}{d\alpha} = -\mathbf{v}^T(\mathbf{I}-\mathbf{S})(\mathbf{I}-\alpha\mathbf{S})^{-2}$$

Ειδικότερα, οι τιμές των παραγώγων στα όρια 0 και 1

$$\lim_{a\to 0}\frac{d\pi^T(\alpha)}{d\alpha} = -\mathbf{v}^T(\mathbf{I}-\mathbf{S})$$

$$\lim_{a\to 1}\frac{d\pi^T(\alpha)}{d\alpha} = -\mathbf{v}^T(\mathbf{I}-\mathbf{S})^{\#}$$

όπου με $(*)^{\#}$ συμβολίζουμε τον group inverse

Σχόλια για το προηγούμενο θεώρημα (1/3)

- Η κυρίαρχη ιδιοτιμή (dominant eigenvalue) $\lambda_1=1$ όλων των στοχαστικών πινάκων είναι *semisimple*, έτσι όταν ο **S** μετατρέπεται σε *μορφή Jordan* με έναν μετασχηματισμό ομοιότητας, το αποτέλεσμα είναι:

$$\mathbf{J} = \mathbf{X}^{-1}\mathbf{S}\mathbf{X} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix}, \; 1 \notin \sigma(\mathbf{C}), \Longrightarrow (\mathbf{I}-\mathbf{S}) = \mathbf{X}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}-\mathbf{C} \end{pmatrix}\mathbf{X}^{-1}$$

και

$$(\mathbf{I}-\mathbf{S})^{\#} = \mathbf{X}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\mathbf{I}-\mathbf{C})^{-1} \end{pmatrix}\mathbf{X}^{-1}$$

- Ο πίνακας **C** αποτελείται από *Jordan μπλοκ* $\mathbf{J}_*$, που συσχετίζονται με τις ιδιοτιμές $\lambda_k \neq 1$, και τα αντίστοιχα μπλοκ στον $(\mathbf{I}-\mathbf{C})^{-1}$ είναι $(\mathbf{I}-\mathbf{J}_*)^{-1}$

Σχόλια για το προηγούμενο θεώρημα (2/3)

- Συνδυάζοντας αυτά με το προηγούμενο θεώρημα, συνάγουμε ότι η ευαισθησία του $\boldsymbol{\pi}^T(\alpha)$ καθώς το α τείνει στο 1 καθορίζεται από το μέγεθος των στοιχείων του $(\mathbf{I-S})^{\#}$
- $||(\mathbf{I-S})^{\#}|| \leq \kappa(\mathbf{X})\ ||(\mathbf{I-C})^{-1}||$, όπου $\kappa(\mathbf{X})$ είναι ο *condition number* του **X**
- Επομένως, η ευαισθησία του $\boldsymbol{\pi}^T(\alpha)$ καθώς το α τείνει στο 1 καθορίζεται κυρίως από το μέγεθος του $||(\mathbf{I-C})^{-1}||$, το οποίο καθορίζεται από το μέγεθος του $||1-\lambda_2||^{-1}$, όπου $\lambda_2 \neq 1$ είναι η ιδιοτιμή του S που είναι πλησιέστερα στη λ_1
- Με άλλα λόγια, <u>όσο πιο κοντά είναι η τιμή του λ_2 στο $\lambda_1=1$, τόσο πιο ευαίσθητο είναι το $\boldsymbol{\pi}^T(\alpha)$, όταν το α είναι κοντά στο 1</u>

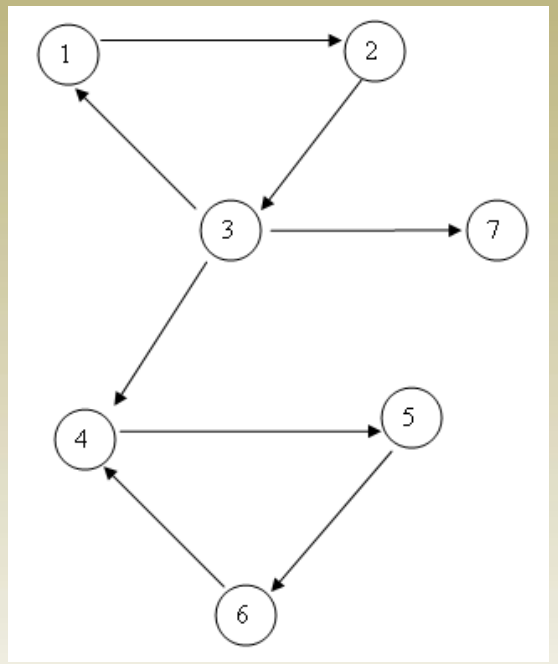
Σχόλια για το προηγούμενο θεώρημα (3/3)

- Μιλώντας γενικά, οι στοχαστικοί πίνακες των οποίων η υποκυρίαρχη ιδιοτιμή (subdominant eigenvalue) είναι κοντά στο 1, αντιπροσωπεύουν *nearly uncoupled chains* (ή *nearly completely decomposable chains*)
- Αυτές είναι οι αλυσίδες των οποίων οι καταστάσεις σχηματίζουν ομάδες (clusters), τέτοιες ώστε οι καταστάσεις μέσα στις ομάδες έχουν ισχυρή σύνδεση μεταξύ τους, ενώ οι ομάδες είναι χαλαρά συνδεδεμένες μεταξύ τους
  - οι καταστάσεις μπορούν να διαταχτούν έτσι ώστε ο πίνακας πιθανοτήτων μεταβάσεων να αποκτήσει τη μορφή $\mathbf{S}=\mathbf{D}+\varepsilon\mathbf{E}$, όπου ο $\mathbf{D}$ είναι διαγώνιος με μπλοκ (block diagonal), $||\mathbf{E}|| \leq 1$, και $0 \leq \varepsilon < 1$ είναι μικρό σχετικά με το 1
- Η αλυσίδα που ορίζεται από το Web είναι σχεδόν βέβαιο ότι είναι nearly uncoupled, οπότε το λ_2 είναι πολύ κοντά στο 1

Συμπεράσματα για την ευαισθησία του $\boldsymbol{\pi}^T$

- Για μικρό α, το διάνυσμα PageRank δεν επηρεάζεται από μικρές αλλαγές στο α
- Καθώς το α μεγαλώνει, η ευαισθησία του διανύσματος PageRank αυξάνει σε μικρές αλλαγές του α
- Όταν το α είναι κοντά στο 1, το διάνυσμα PageRank είναι πάρα πολύ ευαίσθητο σε μικρές αλλαγές του α
  - Ο βαθμός ευαισθησίας ελέγχεται από το βαθμό στον οποίο ο **S** είναι nearly uncoupled

Παράδειγμα 1 (democratic surfer) (1/4)



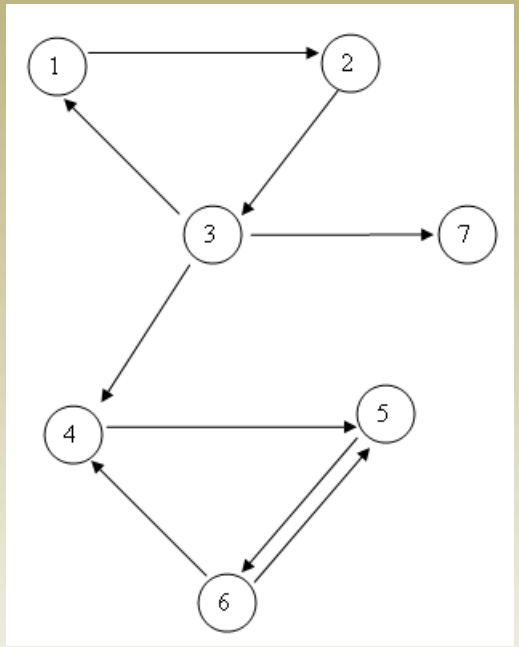
λ_2

		$\alpha = 0.8$			$\alpha = 0.9$			$\alpha = 0.99$		
$\sigma(\mathbf{H})$	$\sigma(\mathbf{S})$	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank
1	1	1	0.0641	6	1	0.0404	6	1	0.0054	6
-.50+0.87i	-0.50+0.87i	-0.40+0.69i	0.0871	5	-0.45+0.78i	0.0558	5	-0.50+0.86i	0.0075	5
-.50-0.87i	-0.50-0.87i	-0.40-0.69i	0.1056	4	-0.45-0.78i	0.0697	4	-0.50-0.86i	0.0096	4
-.35+0.60i	0.7991	0.6393	0.2372	1	0.7192	0.2720	1	-0.7911	0.3253	1
-.35-0.60i	-0.33+0.61i	-0.26+0.49i	0.2256	2	-0.30+0.55i	0.2643	2	-0.33+0.60i	0.3240	2
-.6934	-0.33-0.61i	-0.26-0.49i	0.2164	3	-0.30-0.55i	0.2573	3	-0.33-0.60i	0.3231	3
0	0	0	0.0641	6	0	0.0404	6	0	0.0054	6

Παράδειγμα 1 (democratic surfer) (2/4)

- Οι σελίδες είναι διατεταγμένες από την πιο δημοφιλή προς τη λιγότερο δημοφιλή (4 5 6 3 2 1 7)
- $|\lambda_2(\mathbf{G})| = \alpha$
- Καθώς το α τείνει στο 1, το PageRank αλλάζει σημαντικά
- Όμως, η διάταξη (ranking) δεν αλλάζει!
  - Σε μεγαλύτερα γραφήματα και η διάταξη είναι δυνατόν να αλλάξει
- Η δεύτερη μεγαλύτερη σε τιμή ιδιοτιμή του **S** είναι 0.7991 (Επισημάναμε ήδη ότι αυτή η τιμή, που μετρά επίσης το βαθμό σύζευξης (coupling) μιας Markov αλυσίδας, ελέγχει την ευαισθησία του διανύσματος PageRank)
- Αφού το 0.7991 δεν είναι κοντά στο 1, αναμένουμε ότι αυτή η αλυσίδα δεν θα είναι πολύ ευαίσθητη σε μικρές αλλαγές του α
- Ας ελέγξουμε αυτή την υπόθεση προσθέτοντας έναν υπερσύνδεσμο από τη σελίδα 6 στην 5 (δείτε τον επόμενο πίνακα)

Παράδειγμα 1 (democratic surfer) (3/4)



λ_2

		$\alpha = 0.8$			$\alpha = 0.9$			$\alpha = 0.99$		
$\sigma(\mathbf{H})$	$\sigma(\mathbf{S})$	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank
1	1	1	0.0641	6	1	0.0404	6	1	0.0054	6
-.50+0.50i	0.7991	0.6393	0.0871	5	0.7192	0.0558	5	0.7911	0.0075	5
-.50-0.50i	-0.50+0.50i	-0.40+0.40i	0.1056	4	-0.45+0.45i	0.0697	4	-0.50+0.50i	0.0096	4
-0.6934	-0.50-0.50i	-0.40-0.40i	0.1637	3	-0.45-0.45i	0.1765	3	-0.50-0.50i	0.1968	3
-.35+0.60i	-0.33+0.61i	-0.26+0.49i	0.2664	1	-0.30+0.55i	0.3145	1	-0.33+0.60i	0.3885	1
-.35-0.60i	-0.33-0.61i	-0.26-0.49i	0.2491	2	-0.30-0.55i	0.3025	2	-0.33-0.60i	0.3848	2
0	0	0	0.0641	6	0	0.0404	6	0	0.0054	6

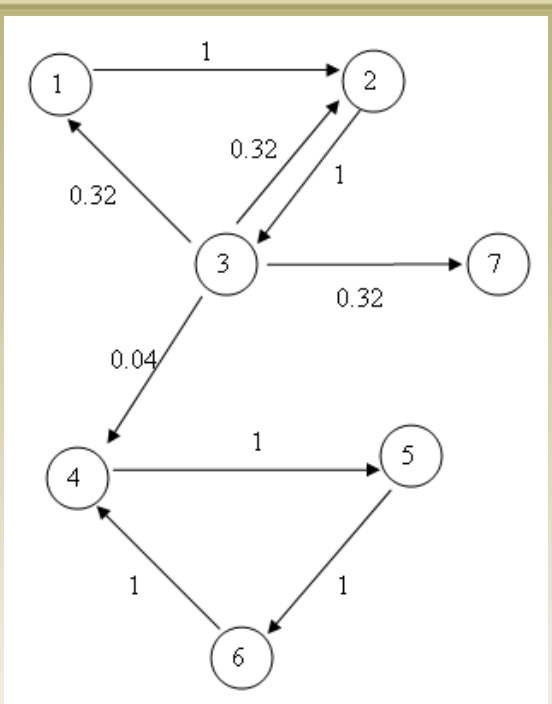
Παράδειγμα 1 (democratic surfer) (4/4)

- Μετά την προσθήκη ενός μόνο υπερσυνδέσμου οι ιστοσελίδες τώρα διατάσσονται από την πιο σημαντική προς τη λιγότερο σημαντική ως εξής: (5 6 4 3 2 1 7). Πριν ήταν: (4 5 6 3 2 1 7)
- Η σελίδα 4 "έπεσε" από την 1η θέση στην 3η θέση!
- Παρατηρούμε ότι μόνο οι PageRank τιμές των ιστοσελίδων 4, 5 και 6 έχουν αλλάξει, ως συνέπεια της reducibility της αλυσίδας [Δεν υπάρχουν links από το κάτω προς το πάνω cluster του γραφήματος.]
- Στο επόμενο παράδειγμα εξετάζουμε μια αλυσίδα, της οποίας η δεύτερη σε μέγεθος ιδιοτιμή του **S** είναι πιο κοντά στο 1

Παράδειγμα 2 (intelligent surfer) (1/4)

Πολύ "πιο ασύζευκτη":

- Εάν ο surfer μπει στο κάτω cluster, τότε "παγιδεύεται" εκεί
- Εάν είναι στο κόμβο 3, έχει 8 φορές (0.32=8 * 0.04) μεγαλύτερη πιθανότητα να ξαναπάει σε κόμβο του πάνω cluster, παρά να πάει στο κάτω cluster



- Ο **S** είναι *πολύ πιο ασύζευκτος* (uncoupled)
- $\lambda_2(S)=0.9193$

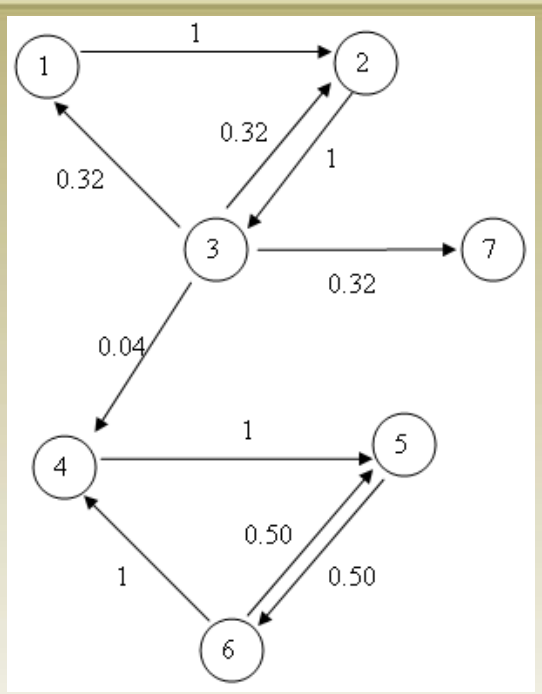
λ_2

$\sigma(\mathbf{H})$	$\sigma(\mathbf{S})$	$\alpha = 0.8$			$\alpha = 0.9$			$\alpha = 0.99$		
		$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank
1	1	1	0.0736	6	1	0.0538	6	1	0.0099	6
-.50+0.87i	-0.50+0.87i	-0.40+0.70i	0.1324	5	-0.45+0.78i	0.1022	5	-0.50+0.86i	0.0197	5
-.50-0.87i	-0.50-0.87i	-0.40-0.70i	0.1429	4	-0.45-0.78i	0.1132	4	-0.50-0.86i	0.0224	4
-0.8378	0.9193	0.7354	0.1943	1	0.8274	0.2271	1	0.9101	0.3130	1
-.42+0.43i	-0.39+0.44i	-0.31+0.36i	0.1924	2	-0.35+0.40i	0.2256	2	-0.38+0.44i	0.3127	2
-.42-0.43i	-0.39-0.44i	-0.31-0.36i	0.1909	3	-0.35-0.40i	0.2242	3	-0.38-0.44i	0.3124	3
0	0	0	0.0736	6	0	0.0538	6	0	0.0099	6

Παράδειγμα 2 (intelligent surfer) (2/4)

- Η διάταξη πλέον των ιστοσελίδων από την πιο σημαντική προς τη λιγότερο σημαντική είναι η εξής (4 5 6 3 2 1 7)
- Ας κάνουμε και σ' αυτό το παράδειγμα την ίδια αλλαγή στο γράφημα που κάναμε προηγουμένως, και ας προσθέσουμε έναν σύνδεσμο από την ιστοσελίδα 6 προς την 5

Παράδειγμα 2 (intelligent surfer) (3/4)



λ_2

		$\alpha = 0.8$			$\alpha = 0.9$			$\alpha = 0.99$		
$\sigma(\mathbf{H})$	$\sigma(\mathbf{S})$	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank
1	1	1	0.0736	6	1	0.0538	6	1	0.0099	6
0.8378	0.9193	0.7354	0.1324	4	0.8274	0.1022	5	0.9101	0.0197	5
-0.50+0.50i	-0.50+0.50i	-0.40+0.40i	0.1429	3	-0.45+0.45i	0.1132	4	-0.50+0.50i	0.0224	4
-0.50-0.50i	-0.50-0.50i	-0.40-0.40i	0.1294	5	-0.45-0.45i	0.1439	3	-0.50-0.50i	0.1889	3
-.42+0.45i	-0.39+0.44i	-0.31+0.36i	0.2284	1	-0.35+0.40i	0.2694	1	-0.38+0.44i	0.3750	1
-.42-0.45i	-0.39-0.44i	-0.31-0.36i	0.2197	2	-0.35-0.40i	0.2636	2	-0.38-0.44i	0.3741	2
0	0	0	0.0736	6	0	0.0538	6	0	0.0099	6

Παράδειγμα 2 (intelligent surfer) (4/4)

- Μετά την αλλαγή αυτή, η διάταξη των σελίδων πλέον γίνεται (5 6 3 2 4 1 7). Πριν ήταν: (4 5 6 3 2 1 7)
- [Πώς συγκρίνουμε ranked lists? **Kendal τ**. Συνάρτηση correl στο xls]
- Η ιστοσελίδα 4 "πέφτει" ακόμα περισσότερο στο ranking
- Τόσο η διάταξη (ranking) όσο και οι πραγματικές τιμές του PageRank των ιστοσελίδων είναι πολύ πιο ευαίσθητες στο Παράδειγμα 2 απ' ότι ήταν στο Παράδειγμα 1
- Συνεπώς, βλέπουμε καθαρά την επίδραση του $\lambda_2(\mathbf{S})$ στην ευαισθησία του διανύσματος PageRank
- Οι Boldi, Santini και Vigna έχουν μελετήσει παραγώγους ανώτερης τάξης και έφτασαν σε πιο πλούσια αποτελέσματα για την ευαισθησία του PageRank

Ευαισθησία σε σχέση με τον πίνακα **H** (1/2)

- Παλιότερα γνωστά αποτελέσματα ευαισθησίας για Markov αλυσίδες δίνουν ότι:

  $\boldsymbol{\pi}^T$ είναι ευαίσθητο σε μεταβολές στο $\mathbf{P} \Leftrightarrow |\lambda_2(\mathbf{P})| \approx 1$
- Γνωρίζουμε ήδη ότι $|\lambda_2(\mathbf{G})| \leq \alpha$, και επιπλέον, όταν ο S είναι reducible ισχύει ότι $|\lambda_2(\mathbf{G})| = \alpha$
- Επομένως, καθώς το α τείνει στο 1, γίνεται όλο και πιο ευαίσθητο σε μικρές μεταβολές στο **G**
- Όμως το **G** εξαρτάται από τα α, **H** και $\mathbf{v}^T$ και επιθυμούμε να απομονώσουμε την εξάρτησή του από το **H**
- Ας υπολογίσουμε μια άλλη παράγωγο:

$$\frac{d\pi^T(h_{ij})}{dh_{ij}} = \alpha\pi_i(\mathbf{e}_j^T - \mathbf{v}^T)(\mathbf{I} - \alpha\mathbf{S})^{-1}$$

Ευαισθησία σε σχέση με τον πίνακα **H** (2/2)

- Η επίδραση του α είναι προφανής
  - Καθώς το α τείνει στο 1, τα στοιχεία του $(\mathbf{I}-\alpha\mathbf{S})^{-1}$ απειρίζονται και το PageRank διάνυσμα γίνεται πολύ ευαίσθητο σε μικρές αλλαγές της συνδεσμολογίας
  - Η προσθήκη ενός υπερσυνδέσμου ή η αύξηση του βάρους ενός υπερσυνδέσμου από μια σημαντική ιστοσελίδα (το π_i είναι υψηλό) έχει μεγαλύτερη επίδραση στην ευαισθησία του διανύσματος PageRank, παρά η αλλαγή ενός υπερσυνδέσμου από μια μη σημαντική σελίδα

Ευαισθησία σε σχέση με το $\mathbf{v}^T$

- Ας υπολογίσουμε την παράγωγο του $\boldsymbol{\pi}^T$ σε σχέση με το διάνυσμα $\mathbf{v}^T$:

$$\frac{d\pi^T(\mathbf{v}^T)}{d\mathbf{v}^T} = (1 - \alpha + \alpha \sum_{i \in D} \pi_i)(\mathbf{I} - \alpha \mathbf{S})^{-1}$$

  όπου το *D* είναι το σύνολο των dangling κόμβων

- Υπάρχει εξάρτηση από το α

  Fundamental matrix

  - Καθώς το α τείνει στο 1, τα στοιχεία του $(\mathbf{I}\text{-}\alpha\mathbf{S})^{-1}$ απειρίζονται, δηλ., καθώς το α τείνει στο 1, το $\boldsymbol{\pi}^T$ γίνεται όλο και πιο ευαίσθητο
- Εάν οι dangling κόμβοι συνδυάζονται για να αποκτήσουν ένα μεγάλο ποσοστό του PageRank, τότε το διάνυσμα $\boldsymbol{\pi}^T$ είναι πολύ ευαίσθητο σε αλλαγές στο διάνυσμα $\mathbf{v}^T$
- Αυτό συμφωνεί με την κοινή λογική
  - ο τυχαίος surfer περνάει αρκετό χρόνο στους dangling κόμβους, και έτσι πιο συχνά ακολουθεί τις teleportation πιθανότητες, δηλ., το $\mathbf{v}^T$

Νόρμα Google G υπό αλλαγές

- **ΘΕΩΡΗΜΑ**. Έστω ότι $\mathbf{G}=\alpha\mathbf{S}+(1-\alpha)\mathbf{e}\mathbf{v}^T$ είναι ο Google πίνακας με διάνυσμα PageRank $\boldsymbol{\pi}^T$ και $\hat{\mathbf{G}}=\alpha\hat{\mathbf{S}}+(1-\alpha)\mathbf{e}\mathbf{v}^T$ είναι ο ενημερωμένος πίνακας Google (ίδιου μεγέθους) με αντίστοιχο διάνυσμα $\tilde{\mathbf{n}}^T$. Τότε:

$$||\pi^T - \hat{\pi}^T||_1 \leq \frac{2\alpha}{1-\alpha} \sum_{i \in U} \pi_i$$

όπου U είναι το σύνολο όλων των ιστοσελίδων που έχουν ενημερωθεί

Άλλες προσεγγίσεις ευαισθησίας (1/3)

- Το προηγούμενο θεώρημα υπονοεί ότι όσο το α δεν είναι κοντά στο 1, και οι ενημερωμένες ιστοσελίδες δεν έχουν υψηλή τιμή PageRank, τότε οι νέες τιμές PageRank δεν αλλάζουν πολύ
- Ας εξετάσουμε τους δυο παράγοντες του ορίου
  - $2\alpha/(1-\alpha)$
  - $\Sigma\pi_\iota$
- Έστω ότι α=0.8 και ότι το άθροισμα των παλιών τιμών του PageRank των ενημερωμένων σελίδων είναι 10^{-6}
- Τότε η πολλαπλασιαστική σταθερά $2\alpha/(1-\alpha) = 8$, το οποίο σημαίνει ότι η 1-νόρμα της διαφοράς του παλιού με το νέο διάνυσμα PageRank είναι το πολύ $8x10^{-6}$
- Άρα οι τιμές PageRank δεν είναι επιρρεπείς στις αλλαγές

Άλλες προσεγγίσεις ευαισθησίας (2/3)

- Καθώς το α τείνει στο 1, το προηγούμενο όριο γίνεται σταδιακά λιγότερο χρήσιμο
- Η χρησιμότητα του ορίου ελέγχεται από το βαθμό στον οποίο το άθροισμα $\Sigma\pi_i$ μπορεί να εξισορροπήσει την αύξηση του κλάσματος $2\alpha/(1-\alpha)$
- Δυο πράγματα επηρεάζουν το μέγεθος του $\Sigma\pi_i$:
  - Ο αριθμός των ενημερωμένων σελίδων
  - Η τιμή του PageRank των σελίδων αυτών
- Το προηγούμενο όριο έχει ένα ακόμα μειονέκτημα: Δεν μας λέει κάτι για το τι συμβαίνει στο PageRank όταν ενημερώνονται οι σελίδες που έχουν μεγάλη τιμή PageRank

Άλλες προσεγγίσεις ευαισθησίας (3/3)

- Όλες οι προηγούμενες προσπάθειες μελέτης της ευαισθησίας του PageRank αφορούσαν την σταθερότητα των τιμών του PageRank
- Οι Lempel & Moran μελέτησαν τη σταθερότητα της διάταξης (ranking)
- Έδειξαν ότι η σταθερότητα των τιμών (PageRank value stability) δεν υπονοεί σταθερότητα διάταξης (rank stability)

Σταθερότητα ranking του PageRank (1/6)

- **ΟΡΙΣΜΟΣ**. Έστω ότι τα v_1 v_2 είναι N-διάστατα διανύσματα με πραγματικές συντεταγμένες. Η ranking distance d_r μεταξύ των v_1 και v_2 ορίζεται ως εξής (μια τυπική εκδοχή):

$$d_r(v_1, v_2) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbf{I}_{v_1,v_2}(i,j)$$

$$\mathbf{I}_{v_1,v_2}(i,j) = \begin{cases} 1, & v_1(i) < v_1(j) \text{ and } v_2(i) > v_2(j) \\ 0, & \text{otherwise} \end{cases}$$

Η d_r είναι μια κανονικοποιημένη έκδοση της Kendal T απόστασης

Π.χ., εάν v_1=(2,4,6,8) και v_2=(2,9,5,3), τότε $d_r(v_1, v_2)$=3/16, εξαιτίας των ζευγών (i,j) ∈ {(2,3),(2,4),(3,4)}

Σταθερότητα ranking του PageRank (2/6)

- Έστω ότι $\mathcal{G}$ είναι ένα σύνολο κατευθυνόμενων γραφημάτων, και $\mathcal{G}_N$ είναι εκείνο το υποσύνολο των γραφημάτων του $\mathcal{G}$ με N κόμβους. Έστω ότι A_1 και A_2 είναι δυο αλγόριθμοι link ranking που αναθέτουν |V|-διάστατα διανύσματα βάρους $A_1(G)$ και $A_2(G)$ στους κόμβους του γραφήματος $G \in \mathcal{G}_N$.
- **ΟΡΙΣΜΟΣ**. Δυο αλγόριθμοι ranking A_1 και A_2 θα λέμε ότι είναι rank-similar στο $\mathcal{G}$, εάν ισχύει ότι:

$$\lim_{N\to\infty} \max_{G\in\mathcal{G}_N} d_r(A(G_1), A(G_2)) \to 0$$

- **ΟΡΙΣΜΟΣ**. Ένας αλγόριθμος A θα λέμε ότι είναι rank-stable στο $\mathcal{G}$, εάν για κάθε σταθερό k, έχουμε ότι:

$$\lim_{N\to\infty} \max_{\{G_1,G_2\in\mathcal{G}_N \,|\, d_e(G_1,G_2)\le k\}} d_r(A(G_1), A(G_2)) \to 0$$

όπου $d_e(G_1,G_2) \equiv |(E_1 \cup E_2) \setminus (E_1 \cap E_2)|$

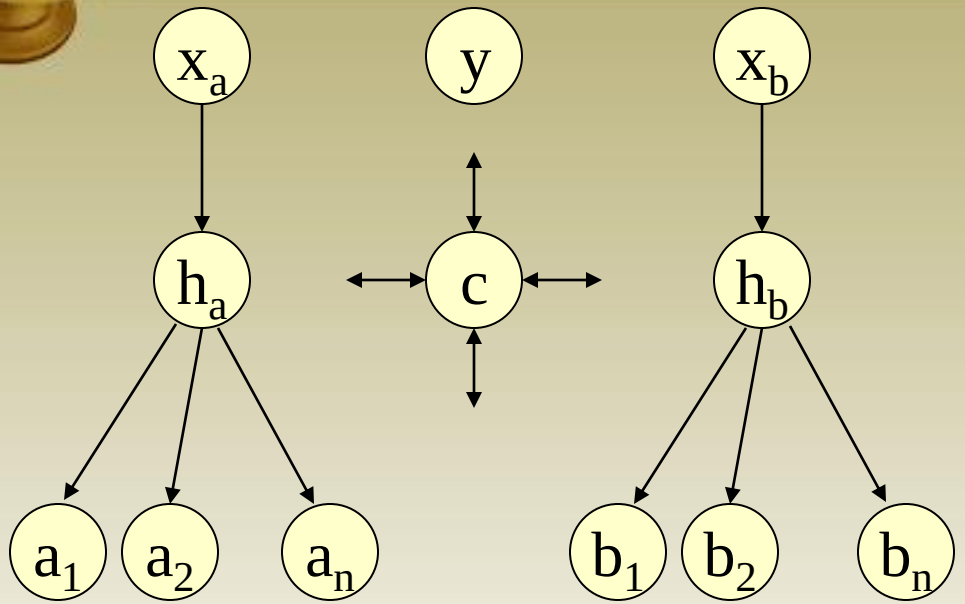
Σταθερότητα ranking του PageRank (3/6)

- Έστω ότι G=(V,E) είναι ένα κατευθυνόμενο γράφημα (που αναπαριστά κάποιο υπογράφημα του Web)
- Δυο κόμβοι p,q $\in$ V θα λέμε ότι είναι co-cited, εάν υπάρχει κάποιος κόμβος r που έχει υπερσύνδεσμο και προς τον κόμβο p και προς τον q
- Θα λέμε ότι οι κόμβοι p και q συνδέονται με ένα co-citation path, εάν υπάρχουν κόμβοι $p=v_0,v_1,\ldots v_{k-1},v_k=q$, τέτοιοι ώστε τα ζεύγη (v_{i-1},v_i) να είναι co-cited για κάθε i=1,2,…,k
- Έστω ότι συμβολίζουμε με V_{in} όλους τους κόμβους του V με τουλάχιστον έναν εισερχόμενο υπερσύνδεσμο
- **ΟΡΙΣΜΟΣ**. Ένα κατευθυνόμενο γράφημα G=(V,E) θα αποκαλείται authority-connected, εάν για όλους τους p,q $\in$ V_{in}, υπάρχει ένα co-citation path που συνδέει τους p και q

Σταθερότητα ranking του PageRank (4/6)

- Θα εξετάσουμε την rank stability PageRank όταν εφαρμόζεται πάνω σε authority-connected γραφήματα
- Γιατί μόνο σε τέτοιου είδους γραφήματα; Γιατί όταν ζητούμε από έναν αλγόριθμο ranking να κατατάξει ιστοσελίδες γραφημάτων που δεν είναι authority-connected, είναι σαν να ζητάμε από τον αλγόριθμο να κατατάξει σελίδες που δεν αναφέρονται στο ίδιο θέμα, π.χ., γεωγραφίας και αθλητικών
- **ΘΕΩΡΗΜΑ**. Ο PageRank δεν είναι rank-stable στην κλάση των authority-connected γραφημάτων
- **ΑΠΟΔΕΙΞΗ**. Με αντιπαράδειγμα (δείτε επόμενες δυο διαφάνειες)

Σταθερότητα ranking του PageRank (5/6)



$$V = \{c, x_a, y, x_b, h_a, h_b, a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n\}$$

$$E = \{x_a \to h_a, x_b \to h_b\} \cup \{h_a \to a_i, h_b \to b_i, \mid i = 1, \ldots, n\} \cup \{c \to v, v \to c \mid v \in V \backslash \{c\}\}$$

$$|V| = 2n + 6 \qquad |E| = 6(n + 2)$$

- Ορίζουμε τα γραφήματα:

$$G_a \triangleq (V, E \cup \{y \to h_a\})$$

$$G_b \triangleq (V, E \cup \{y \to h_b\})$$

- Τα G_a και G_b είναι authority-connected, διαμέσου του κόμβου c
- Έστω ότι $PR_a(v)$, $PR_b(v)$ ($v \in V$) είναι η PageRank τιμή του κόμβου v στα γραφήματα G_a και G_b, αντίστοιχα

Σταθερότητα ranking του PageRank (6/6)

- Από τον ορισμό του PageRank, εύκολα διαπιστώνουμε ότι:

$$0 < PR_a(x_a) = PR_a(y) = PR_a(x_b)$$

- και συνεπώς: $PR_a(h_a) > PR_a(h_b)$
- Επομένως, $PR_a(a_{ai}) > PR_a(a_{bi})$, για κάθε $1 \le i \le n$
- Όμοια, $PR_b(a_{ai}) < PR_b(a_{bi})$, για κάθε $1 \le i \le n$
- Επομένως:

$$d_r(PageRank(G_a), PageRank(G_b)) = \frac{n^2}{(2n+6)^2}, \quad d_e(G_a, G_b) = 2$$

που, για $N \to \infty$ τείνει στο ¼ και όχι στο 0 **(τέλος απόδειξης)**

- Παρατηρήστε ότι $\forall p \in \{h_a, h_b, a_1, \ldots, a_n, b_1, \ldots, b_n\}$, $PR(y) < PR(p)$, σε όποιο από τα δυο γραφήματα
- Επομένως, συντελέστηκε δραματική αλλαγή στο ranking με την αλλαγή ενός μόνο εξερχόμενου υπερσυνδέσμου του κόμβου y, ο οποίος τυγχάνει να έχει πολύ χαμηλό ranking!!!

Νόρμα Google G υπό αλλαγές (Απόδειξη)

ΑΠΟΔΕΙΞΗ. Έστω ότι ο **F** είναι ο πίνακας που αναπαριστά τη διαταραχή (perturbation) μεταξύ δυο στοχαστικών πινάκων **S** και $\hat{\mathbf{S}}$. Έτσι $\mathbf{F}= \mathbf{S} - \hat{\mathbf{S}}$. Τότε:

$$\begin{aligned}
\pi^T - \hat{\pi}^T &= \alpha\hat{\pi}^T\mathbf{S} - \alpha\pi^T\hat{\mathbf{S}} \\
&= \alpha\pi^T\mathbf{S} - \alpha(\hat{\pi}^T - \pi^T + \pi^T)\hat{\mathbf{S}} \\
&= \alpha\pi^T\mathbf{S} - \alpha\pi^T\hat{\mathbf{S}} + \alpha(\pi^T - \hat{\pi}^T)\hat{\mathbf{S}} \\
&= \alpha\pi^T\mathbf{F} + \alpha(\pi^T - \hat{\pi}^T)\hat{\mathbf{S}}
\end{aligned}$$

Επιλύοντας ως προς $\boldsymbol{\pi}^T$ - $\tilde{\mathbf{n}}^T$ έχουμε:

$$\pi^T - \hat{\pi}^T = \alpha\pi^T\mathbf{F}(\mathbf{I} - \alpha\hat{\mathbf{S}})^{-1}$$

Νόρμα Google G υπό αλλαγές (Απόδειξη)

Υπολογίζοντας νόρμες, έχουμε:

$$||\pi^T - \hat{\pi}^T||_1 \leq \alpha||\pi^T\mathbf{F}||_1||(\mathbf{I} - \alpha\hat{\mathbf{S}})^{-1}||_\infty$$
$$= \frac{\alpha}{1-\alpha}||\pi^T\mathbf{F}||_1$$

Ισχύει ότι ο $\mathbf{I}$-α$\hat{\mathbf{S}}$ είναι μη-ιδιόμορφος (nonsingular) και έχει αθροίσματα γραμμών ίσα προς 1/(1-α). Τώρα, αναδιατάσσουμε τον $\mathbf{F}$ (και $\boldsymbol{\pi}^T$) έτσι ώστε οι γραμμές που αντιστοιχούν στις ανανεωμένες σελίδες (μη μηδενικές γραμμές) να έρθουν στην κορυφή του πίνακα. Τότε

$$\pi^T\mathbf{F} = \begin{pmatrix}\pi_1^T & \pi_2^T\end{pmatrix}\begin{pmatrix}\mathbf{F}_1 \\ \mathbf{0}\end{pmatrix} = \pi_1^T\mathbf{F}_1$$

Νόρμα Google G υπό αλλαγές (Απόδειξη)

Επομένως: $$||\pi^T \mathbf{F}||_1 = ||\pi_1^T \mathbf{F}_1||_1 \leq ||\pi_1^T||_1 ||\mathbf{F}_1||_\infty$$

και $$||\mathbf{F}_1||_\infty = ||\mathbf{S}_1 - \hat{\mathbf{S}}_1||_\infty \leq ||\mathbf{S}_1||_\infty + ||\hat{\mathbf{S}}_1||_\infty = 2$$

όπου $\mathbf{S}_1$ και $\hat{\mathbf{S}}_1$ επίσης αντιστοιχούν σε ενημερωμένες σελίδες

Επομένως: $$||\pi^T \mathbf{F}||_1 \leq 2 \sum_{i \in U} \pi_i$$

Τελικά: $$||\pi^T - \hat{\pi}^T||_1 \leq \frac{2\alpha}{1-\alpha} \sum_{i \in U} \pi_i$$