Ανάκληση Πληροφορίας

Διδάσκων –
Δημήτριος Κατσαρός

Διάλεξη 17η

Η μέθοδος HITS

Hypertext Induced Topic Search

Hypertext Induced Topic Search (HITS)

- Επινοήθηκε από τον Jon Kleinberg το 1998
- Διαφορές των δυο κύριων αλγορίθμων ranking:
  - Ο PageRank είναι query-independent
  - Ο HITS είναι query-dependent
  - Ο PageRank παράγει ένα μέτρο "σημαντικότητας" που χαρακτηρίζει κάθε ιστοσελίδα
  - Ο HITS παράγει δυο τέτοιους αριθμούς
    - Το authority score & το hub score
- Ο HITS αναλύει τις σελίδες ως authorities και hubs
  - Μια authority είναι μια σελίδα με πολλούς εισερχόμενους υπερσυνδέσμους
  - Ένα hub είναι μια σελίδα με πολλούς εξερχόμενους υπερσυνδέσμους
- *Οι καλές authorities δείχνονται από καλά hubs, και τα καλά hubs δείχνουν σε καλές authorities*

Hypertext Induced Topic Search (HITS)

- Ο HITS ενσωματώθηκε στο έργο της IBM "CLEVER"
- Αποτέλεσε τη βάση για τη μηχανή αναζήτησης Teoma (αγοράστηκε από την Ask Jeeves, τώρα Ask.com
- Συνεπώς, κάθε σελίδα i έχει ένα **authority score $\mathbf{a_i}$** και ένα **hub score $\mathbf{h_i}$**
- Με e_{ij} συμβολίζουμε την ύπαρξη υπερσυνδέσμου από την ιστοσελίδα i στην j
- Υποθέτουμε ότι αρχικά έχουμε αναθέσει σε κάθε ιστοσελίδα ένα authority score $\mathbf{x_i}$ και ένα hub score $\mathbf{y_i}$
- Ο HITS υπολογίζει επαναληπτικά τις ποσότητες:

$$x_i^k = \sum_{j:e_{ji}\in E} y_j^{(k-1)} \qquad y_i^k = \sum_{j:e_{ij}\in E} x_j^{(k)} \qquad k = 0, 1, 2, \ldots$$

Hypertext Induced Topic Search (HITS)

- Έστω ο πίνακας γειτνίασης $\mathbf{L}_{ij}$ με στοιχεία ίσα με 1, εάν υπάρχει υπερσύνδεσμος από την ιστοσελίδα i στην j, και ίσα με 0, στην άλλη περίπτωση
- Οι προηγούμενες επαναληπτικές εξισώσεις μπορούν να γραφούν με τη βοήθεια πινάκων ως εξής:

$$\mathbf{x}^{(k)}=\mathbf{L}^T\mathbf{y}^{(k-1)} \quad \text{και} \quad \mathbf{y}^{(k)}=\mathbf{L}\mathbf{x}^{(k)}$$

Ο αρχικός αλγόριθμος HITS

- Αρχικοποίηση του $\mathbf{y}^{(0)}=\mathbf{e}$ (και άλλες επιλογές αρχικοποίησης είναι πιθανές)
- Μέχρι να επέλθει σύγκλιση:
  - $\mathbf{x}^{(k)}=\mathbf{L}^T\mathbf{y}^{(k-1)}$
  - $\mathbf{y}^{(k)}=\mathbf{L}\mathbf{x}^{(k)}$
  - ++k;
  - Κανονικοποίηση των $\mathbf{x}^{(k)}$ και $\mathbf{y}^{(k)}$

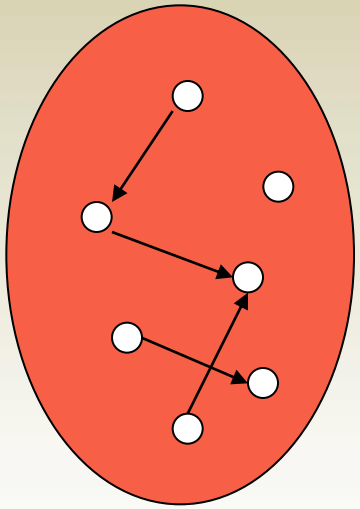
Hypertext Induced Topic Search (HITS)

- Οι προηγούμενες εξισώσεις μπορούν να απλοποιηθούν στις επόμενες:

$$\mathbf{x}^{(k)}=\mathbf{L}^T\mathbf{L}\mathbf{x}^{(k-1)} \quad \text{και} \quad \mathbf{y}^{(k)}=\mathbf{L}\mathbf{L}^T\mathbf{y}^{(k)}$$

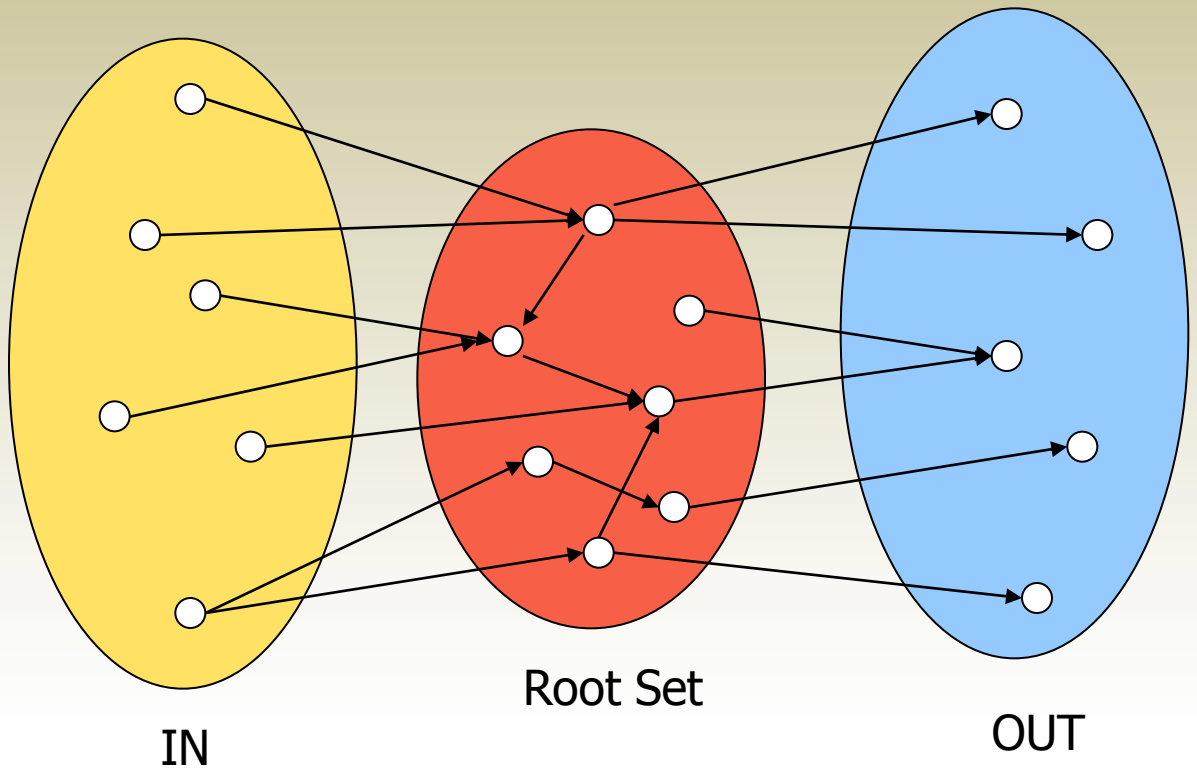
- Άρα, ορίζουν την επαναληπτική power μέθοδο για τον υπολογισμό των κυρίαρχων ιδιοδιανυσμάτων των πινάκων $\mathbf{L}^T\mathbf{L}$ και $\mathbf{L}\mathbf{L}^T$
- Είναι παρόμοια περίπτωση με τον υπολογισμό του PageRank, αλλά με διαφορετικό πίνακα συντελεστών
- Ο πίνακας $\mathbf{L}^T\mathbf{L}$ λέγεται **πίνακας authority**, αφού καθορίζει τα authority scores
- Ο πίνακας $\mathbf{L}\mathbf{L}^T$ λέγεται **πίνακας hub**, αφού καθορίζει τα hub scores
- Και οι δυο πίνακας είναι συμμετρικοί, θετικοί και semidefinite

Υλοποίηση του HITS (1/5)

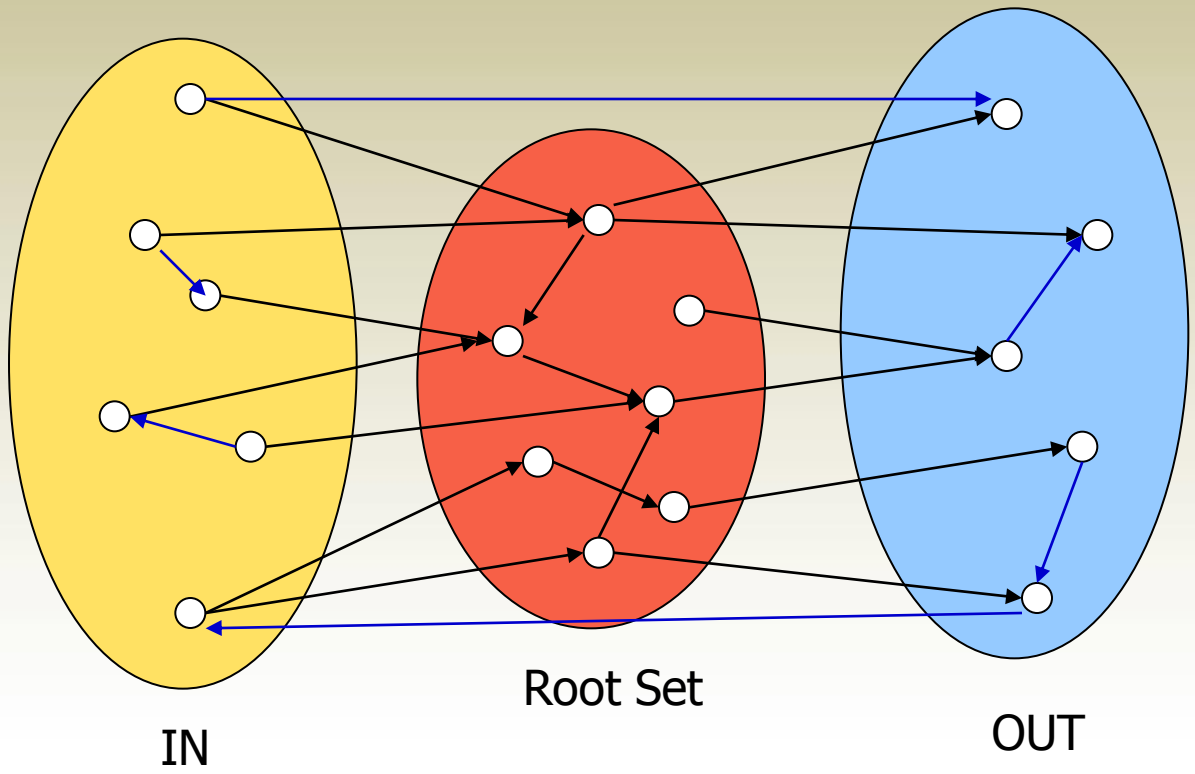


Root Set

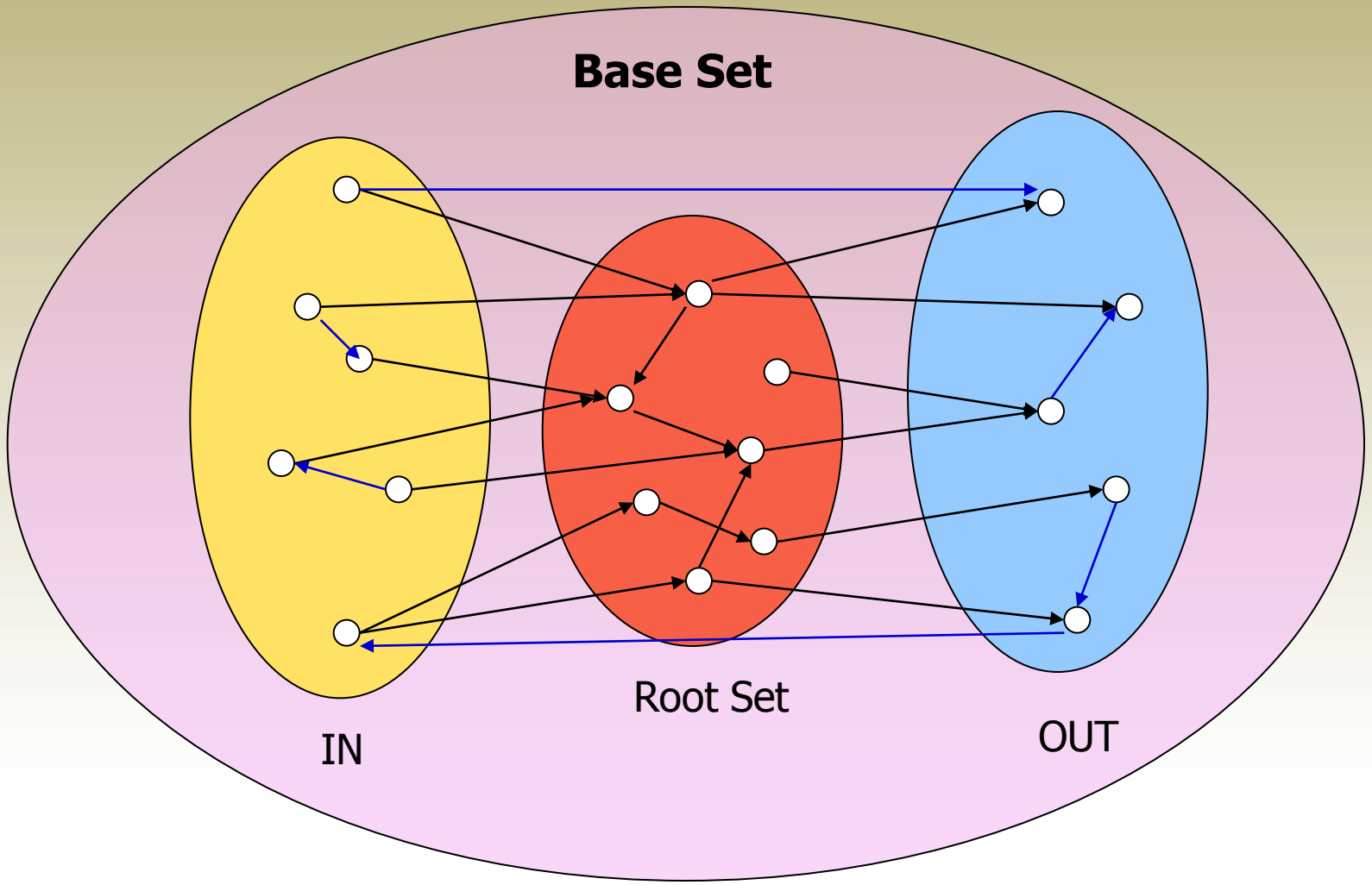
Υλοποίηση του HITS (2/5)



Υλοποίηση του HITS (3/5)



Υλοποίηση του HITS (4/5)



Υλοποίηση του HITS (5/5)

- Το Root Set βρίσκεται με τη βοήθεια μιας μηχανής αναζήτησης με keyword-search
- Το Base Set είναι ο Neighborhood Graph
- Ενσωματώνονται και τεχνικές σημασιολογικής συγγένειας στην επιλογή των σελίδων που θ' απαρτίζουν το Base Set
- Για να μην μεγαλώσει σε τεράστιο μέγεθος το Base Set, ορίζουμε έναν μέγιστο αριθμό κόμβων (με εισερχόμενους/εξερχόμενους) υπερσυνδέσμους για τον κάθε κόμβο του Root Set τους οποίους ενσωματώνουμε στο Base Set
- Δεν χρειάζεται να υπολογίσουμε το κυρίαρχο ιδιοδιάνυσμα και για τον πίνακα $\mathbf{L}^T\mathbf{L}$ αλλά και για τον $\mathbf{L}\mathbf{L}^T$, αλλά μόνο για τον ένα πίνακα, αφού ισχύει: $\mathbf{y}=\mathbf{L}\mathbf{x}$

Σύγκλιση του HITS (1/6)

- Ο επαναληπτικός αλγόριθμος για τον υπολογισμό του HITS είναι συνήθως η power μέθοδος πάνω στους πίνακες $\mathbf{L}^T\mathbf{L}$ και $\mathbf{L}\mathbf{L}^T$
- Για έναν διαγωνιοποιήσιμο πίνακα $\mathbf{B}$ με διακριτές ιδιοτιμές $\{\lambda_1,\lambda_2,\ldots,\lambda_k\}$ όπου $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_k|$ η power μέθοδος υπολογίζει επαναληπτικά το εξής:

$$\mathbf{x}^{(k)} = \mathbf{B}\mathbf{x}^{(k-1)} \qquad \mathbf{x}^{(k)} \Longleftarrow \frac{\mathbf{x}^{(k)}}{m(\mathbf{x}^{(k)})}$$

  όπου $m(\mathbf{x}^{(k)})$ είναι "σταθερά" κανονικοποίησης παραγόμενη από το $\mathbf{x}^{(k)}$
- Συνήθως, είναι η (προσημασμένη) συνιστώσα με το μέγιστο μέγεθος. Στην περίπτωση αυτή, το $m(\mathbf{x}^{(k)})$ συγκλίνει στην κυρίαρχη ιδιοτιμή λ_1 και το $\mathbf{x}^{(k)}$ στο αντίστοιχο κανονικοποιημένο ιδιοδιάνυσμα

Σύγκλιση του HITS (2/6)

- Εάν απαιτείται μόνο το ιδιοδιάνυσμα και όχι και η αντίστοιχη ιδιοτιμή, τότε η "σταθερά" κανονικοποίησης μπορεί να είναι η $m(\mathbf{x}^{(k)}) = ||\mathbf{x}^{(k)}||$
- Εάν $\lambda_1 < 0$, τότε η $m(\mathbf{x}^{(k)}) = ||\mathbf{x}^{(k)}||$ δεν μπορεί να συγκλίνει στην λ_1, αλλά η $\mathbf{x}^{(k)}$ συγκλίνει στο ιδιοδιάνυσμα που αντιστοιχεί στην λ_1
- Επειδή οι πίνακες $\mathbf{L}^T\mathbf{L}$ και $\mathbf{L}\mathbf{L}^T$ είναι συμμετρικοί, θετικοί, semidefinite και μη-αρνητικοί, οι διακριτές τους ιδιοτιμές $\{\lambda_1,\lambda_2,\ldots,\lambda_k\}$ είναι πραγματικοί αριθμοί και μη-αρνητικοί, και ισχύει ότι με $\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_k \geq 0$
- Έτσι, ο HITS με κανονικοποίηση πάντα συγκλίνει
- Ο ρυθμός σύγκλισης δίνεται από τον ρυθμό με τον οποίο ισχύει ότι $[\lambda_2(\mathbf{L}^T\mathbf{L})/\lambda_1(\mathbf{L}^T\mathbf{L})]^k \rightarrow 0$

Σύγκλιση του HITS (3/6)

- Δυστυχώς, δεν μπορούμε να δώσουμε ικανοποιητική προσέγγιση για τον ασυμπτωτικό ρυθμό σύγκλισης του HITS
- Πολλά πειράματα δείχνουν ότι η διαφορά των πρώτων ιδιοτιμών (λ_1-λ_2) είναι αρκετά μεγάλη, και συνεπώς απαιτούνται μόνο μερικές επαναλήψεις (10-15) για να συγκλίνει
- Παρά την ταχεία σύγκλιση όμως, υπάρχει πρόβλημα με τη μοναδικότητα των διανυσμάτων authority και hub που προκύπτουν ως λύση με την power μέθοδο, π.χ., για τον $\mathbf{L}$

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \qquad \mathbf{L}^T\mathbf{L} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Σύγκλιση του HITS (4/6)

- Έχει δυο διακριτές ιδιοτιμές, τις 2 και 0, οι οποίες έχουν πολλαπλότητα 2
- Εάν ξεκινήσουμε με το $\mathbf{x}^{(0)} = ¼\ \mathbf{e}^T$ καταλήγουμε στο διάνυσμα authority $\mathbf{x}^{(\infty)} = (1/3 \quad 1/3 \quad 1/3 \quad 0)^T$
- Εάν ξεκινήσουμε με το $\mathbf{x}^{(0)} = (1/4 \quad 1/8 \quad 1/8 \quad 1/2$ καταλήγουμε στο διάνυσμα authority $\mathbf{x}^{(\infty)} = (1/2 \quad 1/4 \quad 1/4 \quad 0)^T$
- Αιτία του προβλήματος μοναδικότητας είναι η reducibility
- Θα λέμε ότι ένας πίνακας **B** είναι reducible εάν υπάρχει πίνακας μετάθεσης **Q** τέτοιος ώστε (οι **X** και **Z** είναι τετραγωνικοί):

$$\mathbf{Q}^T\mathbf{B}\mathbf{Q} = \begin{pmatrix} \mathbf{X} & \mathbf{Y} \\ \mathbf{0} & \mathbf{Z} \end{pmatrix}$$

Σύγκλιση του HITS (5/6)

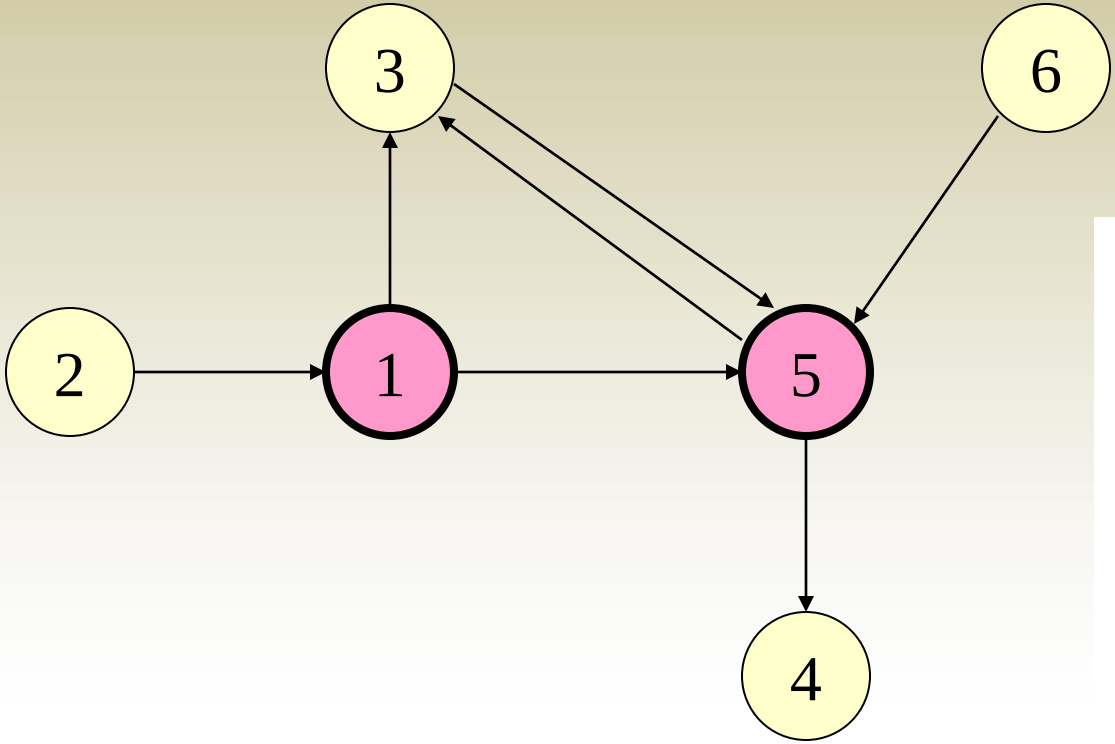
- Η reducibility ενός πίνακα σημαίνει ότι υπάρχουν καταστάσεις “καταβόθρες”, ενώ η irreducibility σημαίνει ότι από οποιαδήποτε κατάσταση μπορώ να μεταβώ σε οποιαδήποτε άλλη κατάσταση
- Το θεώρημα Perron-Frobenious εγγυάται ότι ένας irreducible, μη-αρνητικός πίνακας έχει ένα μοναδικό, κανονικοποιημένο θετικό κυρίαρχο ιδιοδιάνυσμα, το λεγόμενο *διάνυσμα Perron*
- Συνεπώς, η reducibility του $\mathbf{L}^{\mathrm{T}}\mathbf{L}$ είναι υπεύθυνη για η σύγκλιση σε περισσότερα του ενός διανύσματα-λύσεις
- Το ίδιο πρόβλημα αντιμετώπισε και ο πίνακας S στο PageRank, αλλά με μια μετατροπή του πίνακα σε irreducible επιλύθηκε
- Το ίδιο μπορεί να γίνει και με τον HITS

Σύγκλιση του HITS (6/6)

- Αντί για τον αρχικό πίνακα authority, χρησιμοποιούμε τον πίνακα $\xi\mathbf{L}^T\mathbf{L} + (1-\xi)/n\ \mathbf{e}\mathbf{e}^T$
- Όμοια για τον πίνακα hub
- Αυτός ο τροποποιημένος HITS, λέγεται Exponential HITS
- Τέλος, ανεξάρτητα από το εάν η κυρίαρχη ιδιοτιμή του πίνακα επανάληψης $\mathbf{B}$ είναι απλή ή πολλαπλή, η σύγκλιση σε ένα μη-μηδενικό διάνυσμα εξαρτάται από εάν το αρχικό διάνυσμα $\mathbf{x}^{(0)}$ δεν βρίσκεται στην εμβέλεια του $(\mathbf{B}-\lambda_1\mathbf{I})$
- Εάν το $\mathbf{x}^{(0)}$ παράγεται τυχαία, τότε με (σχεδόν) βεβαιότητα δεν θα υφίσταται το πρόβλημα

Παράδειγμα εφαρμογής του HITS (1/4)

- Έστω ότι, σε απάντηση ενός ερωτήματος σε μια παραδοσιακή keyword-based μηχανή αναζήτησης, επιστρέφονται οι ιστοσελίδες που αντιστοιχούν στους κόμβους 1 και 5



$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Παράδειγμα εφαρμογής του HITS (2/4)

$$\mathbf{L}^{\mathrm{T}}\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{L}\mathbf{L}^{\mathrm{T}} = \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

- Τα κανονικοποιημένα κύρια ιδιοδιανύσματα με τα authority και hub scores είναι:

$$\mathbf{x}^{\mathrm{T}} = (0 \quad 0 \quad 0.3660 \quad 0.1340 \quad 0.5 \quad 0)$$

$$\mathbf{y}^{\mathrm{T}} = (0.3660 \quad 0 \quad 0.2113 \quad 0 \quad 0.2113 \quad 0.2113)$$

Παράδειγμα εφαρμογής του HITS (3/4)

- Μπορεί να συμβούν δυο τύπων ισοπαλίες
  - Στις τιμές 0
    - Μπορεί να αποφευχθούν με την primitivity τροποποίηση
  - Στις θετικές τιμές
    - αυτές είναι σπάνιες σε μεγάλους θετικά-ορισμένους πίνακες
    - μπορούν να επιλυθούν με FCFS
- Authority ranking = (6 3 5 1 2 10)
- Hub ranking = (1 3 6 10 2 5)
- Για λόγους σύγκρισης, υπολογίζουμε ξανά τα διανύσματα authority και hub, αλλά χρησιμοποιώντας τον irreducible πίνακα $\xi \mathbf{L}^T\mathbf{L} + (1-\xi)/n\ \mathbf{e}\mathbf{e}^T$ ως πίνακα authority και τον irreducible πίνακα $\xi \mathbf{L}\mathbf{L}^T + (1-\xi)/n\ \mathbf{e}\mathbf{e}^T$ ως πίνακα hub

Παράδειγμα εφαρμογής του HITS (4/4)

- Για ξ=0.95, έχουμε

$\mathbf{x}^T$ = (0.0032 0.0023 0.3634 0.1351 0.4936 0.0023)
$\mathbf{y}^T$ = (0.3628 0.0032 0.2106 0.0023 0.2106 0.2106)

- Στο παράδειγμα αυτό, η μετατροπή τους σε irreducible πίνακες δεν άλλαξε το ranking, ούτε το authority ούτε το hub ranking
- Όμως, απέφυγε τις ισοπαλίες στο 0

Πλεονεκτήματα/Μειονεκτήματα του HITS

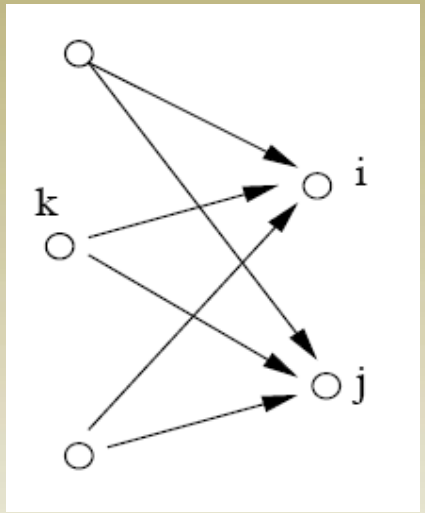
- Παρουσιάζει δυο λίστες στο χρήστη
  - Authoritative σελίδες, για εις βάθος αναζήτηση σε κάποιο αντικείμενο
  - Hub σελίδες, δηλ., portal σελίδες, για αναζήτηση σε εύρος
- Λύνει ένα μικρό πρόβλημα, σε μέγεθος πινάκων
- Είναι query-dependent
  - Σε run-time, δηλ., για κάθε ερώτημα του χρήστη, χτίσιμο του Base Set, και εύρεση ενός ιδιοδιανύσματος
- Μπορεί να γίνει query-independent, δηλ., να εκτελεστεί πάνω σε όλο το γράφημα του Web
- Είναι πολύ επιρρεπής σε spamming
  - Με προσθήκη συνδέσμων από/προς την ιστοσελίδα μας
  - Φυσικά, είναι ευκολότερο να αυξήσουμε το hub score της ιστοσελίδας μας, αλλά εξ' αιτίας της αλληλεξάρτησής τους μπορεί να αυξηθεί και το authority score ως αποτέλεσμα της αύξησης του hub score

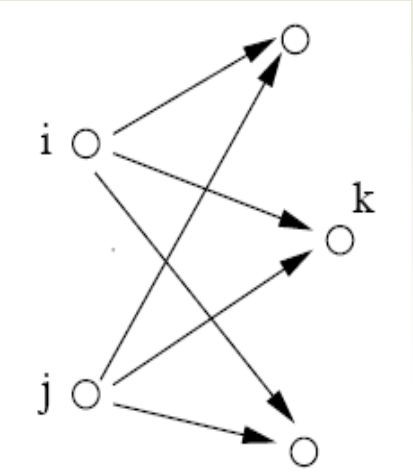
Πλεονεκτήματα/Μειονεκτήματα του HITS

- Παρουσιάζει το φαινόμενο **topic drift**
  - Κατά το χτίσιμο του Base Set, είναι πιθανό ότι μια πολύ authoritative σελίδα, αλλά εκτός αντικειμένου της αναζήτησης, να εμφανιστεί στο Base Set επειδή έχει σύνδεσμο από/προς κάποια σελίδα του Root Set
  - Αυτή η σελίδα μπορεί να έχει τόσο βάρος, ώστε αυτή και όλες οι "γειτονικές" της να εμφανίζονται στην κορυφή της λίστας των αποτελεσμάτων, με συνέπεια η λίστα των αποτελεσμάτων να κυριαρχηθεί από σελίδες εκτός του ζητούμενου αντικειμένου
  - Το πρόβλημα μπορεί να αντιμετωπιστεί εάν "ζυγίσουμε" τα hub και authority scores των σελίδων με βάση τη σχετικότητά της κάθε σελίδας ως προς το αντικείμενο της αναζήτησης, δηλ., αντί για τον δυαδικό πίνακα **L**, να έχουμε κάτι ανάλογο με τον πίνακα του intelligent surfer

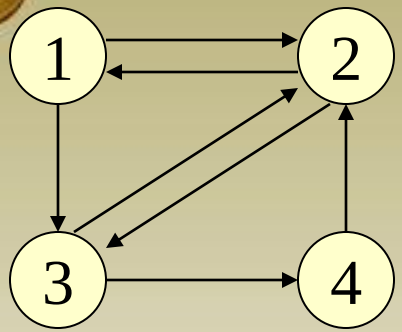
Σχέση του HITS με τη Βιβλιομετρία (1/2)

- **Co-citation**: Δυο σελίδες δέχονται υπερσύνδεσμο από την ίδια σελίδα
- Ο πίνακας authority $\mathbf{L^T L}$ σχετίζεται με την έννοια του co-citation
- Ranking με βάση το inlink παρέχει αξιοπρεπή προσέγγιση του HITS authority
- **Co-reference**: Δυο σελίδες έχουν υπερσύνδεσμο προς την ίδια σελίδα
- Ο πίνακας hub $\mathbf{L L^T}$ σχετίζεται με την έννοια του co-reference
- Ranking με βάση το outlink παρέχει αξιοπρεπή προσέγγιση του HITS hub





Σχέση του HITS με τη Βιβλιομετρία (2/2)



$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\mathbf{L}^{\mathrm{T}}\mathbf{L} = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 3 & 1 & 1 \\ 1 & 1 & 2 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} = \mathbf{D}_{in} + \mathbf{C}_{cit}$$

$$\mathbf{L}\mathbf{L}^{\mathrm{T}} = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 0 & 0 \\ 1 & 0 & 2 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} = \mathbf{D}_{out} + \mathbf{C}_{ref}$$

- $\mathbf{D_{in}}$: διαγώνιος πίνακας με το indegree των κόμβων και $\mathbf{C_{cit}}$ ο πίνακας co-citation
  - Το στοιχείο (3,3) δείχνει ότι ο κόμβος 3 έχει indegree 2
  - Το στοιχείο (4,3) δείχνει ότι οι κόμβοι 4 και 3 δεν έχουν κοινό inlink
- $\mathbf{D_{out}}$: διαγώνιος πίνακας με το outdegree των κόμβων και $\mathbf{C_{ref}}$ ο πίνακας co-reference
  - Το στοιχείο (1,2) δείχνει ότι οι κόμβοι 1 και 2 έχουν 1 κοινό σύνδεσμο (προς τον κόμβο 3)
  - Το στοιχείο (4,2) δείχνει ότι οι κόμβοι 4 και 2 δεν έχουν κοινό σύνδεσμο προς κάποιο κόμβο

Query-independent HITS (1/4)

Ο query-independent αλγόριθμος HITS

- Αρχικοποίηση του $\mathbf{x}^{(0)}=\mathbf{e}/n$ (και άλλες επιλογές αρχικοποίησης είναι πιθανές)
- Μέχρι να επέλθει σύγκλιση:
  - $\mathbf{x}^{(k)} = \xi \mathbf{L}^T \mathbf{L} \mathbf{x}^{(k-1)} + (1-\xi)/n\ \mathbf{e}$
  - $\mathbf{x}^{(k)} = \mathbf{x}^{(k)} / \|\mathbf{x}^{(k)}\|_1$
  - $\mathbf{y}^{(k)} = \xi \mathbf{L} \mathbf{L}^T \mathbf{y}^{(k-1)} + (1-\xi)/n\ \mathbf{e}$
  - $\mathbf{y}^{(k)} = \mathbf{y}^{(k)} / \|\mathbf{y}^{(k)}\|_1$
  - ++k;
  - Θέτουμε $\mathbf{x} = \mathbf{x}^{(k)}$ και $\mathbf{y} = \mathbf{y}^{(k)}$

- Συγκλίνει στα μοναδικά θετικά διανύσματα hub και authority, ανεξάρτητα από τη reducibility του πίνακα
- Ο **L** είναι ο πίνακας γειτνίασης όλου του Web γραφήματος

Query-independent HITS (2/4)

Μέθοδος	Πολλαπλασιασμοί	Προσθέσεις
HITS	0	$2nnz(\mathbf{L})$
Modified HITS	0	$4nnz(\mathbf{L}) + 2n$
Random surfer PageRank	n	$nnz(\mathbf{L}) + n$
Intelligent surfer PageRank	$nnz(\mathbf{H})$	$nnz(\mathbf{H}) + n$

- **ΘΕΩΡΗΜΑ.** Έστω ότι M = είναι ο τροποποιημένος πίνακας authority. Έστω ότι $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ είναι οι ιδιοτιμές του $\mathbf{L}^T\mathbf{L}$ και $\gamma_1 \geq \gamma_2 \geq \ldots. \geq \gamma_n$ είναι οι ιδιοτιμές του M. Τότε ισχύει η σχέση:

$$\gamma_1 \geq \alpha\lambda_1 \geq \gamma_2 \geq \alpha\lambda_2 \geq \ldots. \geq \geq \gamma_n \geq \alpha\lambda_n$$

Υπάρχουν scalars $\beta_i \geq 0$, $\sum\beta_i = 1$, ώστε $\gamma_i = \xi\lambda_i + (1-\xi)\beta_i$

Query-independent HITS (3/4)

- Τα όρια γ_2/γ_1 παράγονται εξετάζοντας ακραία συμπεριφορά
- Στην καλύτερη περίπτωση, η τροποποίηση του $\mathbf{L^TL}$ αυξάνει μόνο το λ_2 σε μέγιστη τιμή $\lambda_2+1-\xi$ (δηλ., $\beta_2=1$, $\beta_i=0$, $i\neq2$)
- Στη χειρότερη περίπτωση, το λ_1 αυξάνει σε μέγιστη τιμή $\lambda_1+1-\xi$ (δηλ., $\beta_1=1$, $\beta_i=0$, $i\neq1$)
- Στην πράξη, πολλά β_i αυξάνουν ταυτόχρονα, αλλά η σχέση $\sum\beta_i=1$ εγγυάται ότι το αποτέλεσμα δεν έχει δραματικές επιπτώσεις, όπως στις δυο ακραίες περιπτώσεις

Μέθοδος	**Σύγκλιση**
HITS	λ_2/λ_1
Modified HITS	$(\xi\lambda_2)/(\xi\lambda_1+1-\xi) \leq \gamma_2/\gamma_1 \leq \lambda_2/\lambda_1 \leq (1-\xi)/(\xi\lambda_1)$
Random surfer PageRank	α
Intelligent surfer PageRank	α

Query-independent HITS (4/4)

- Ανεξάρτητα από τις ακριβείς τιμές των β_i, το ξ επιλέγεται συνήθως κοντά στο 1, οπότε $\gamma_2/\gamma_1 \approx \lambda_2/\lambda_1$
- Έτσι ο ασυμπτωτικός ρυθμός σύγκλισης του HITS και του τροποποιημένου HITS είναι ο ίδιος
- Πειράματα έχουν δείξει ότι $\lambda_2/\lambda_1 < 0.5$, το οποίο είναι πολύ μικρότερο του $\alpha=0.85$ του PageRank
- Με περίπου διπλάσιο κόστος ανά επανάληψη, ο query-independent-HITS απαιτεί λιγότερο από ¼ των επαναλήψεων του PageRank και παράγει 2 διανύσματα στο χρήστη

Επιτάχυνση του HITS

- Ο Kleinberg χρησιμοποίησε την power μέθοδο για τον υπολογισμό των κυρίαρχων δεξιών ιδιοδιανυσμάτων των πινάκων $\mathbf{L^TL}$ και $\mathbf{LL^T}$
- Οι πίνακες του HITS είναι πολύ μικροί σε σχέση με τους αντίστοιχους του PageRank, και κατά πάσα πιθανότητα, χρησιμοποιούνται τεχνικές που είναι memory-intensive για τον υπολογισμό των δυο πινάκων του HITS, π.χ., Lanczos
- Για τον query-dependent HITS δεν υπάρχει έρευνα σχετική με μεθόδους επιτάχυνσης
- Για τον query-independent HITS μπορούν να χρησιμοποιηθούν οι ίδιες τεχνικές που έχουμε συζητήσει για τον PageRank

Ευαισθησία του HITS

- **ΘΕΩΡΗΜΑ**. Έστω **E** ο πίνακας διαταραχής, ώστε $\mathbf{\acute{L}^T\acute{L}} = \mathbf{L^TL + E}$. Όταν η λ_1 είναι απλή, τότε

$$\sin \angle(\mathbf{x},\mathbf{x'}) \leq ||\mathbf{E}||_2/(\lambda_1-\lambda_2)$$

- Συνεπώς, εάν το χάσμα ιδιοδιανυσμάτων είναι μεγάλο, τότε ο πίνακας authority δεν είναι ευαίσθητος σε μικρές αλλαγές στο Web γράφημα

Η μέθοδος SALSA

Stochastic Approach for Link Structure Analysis

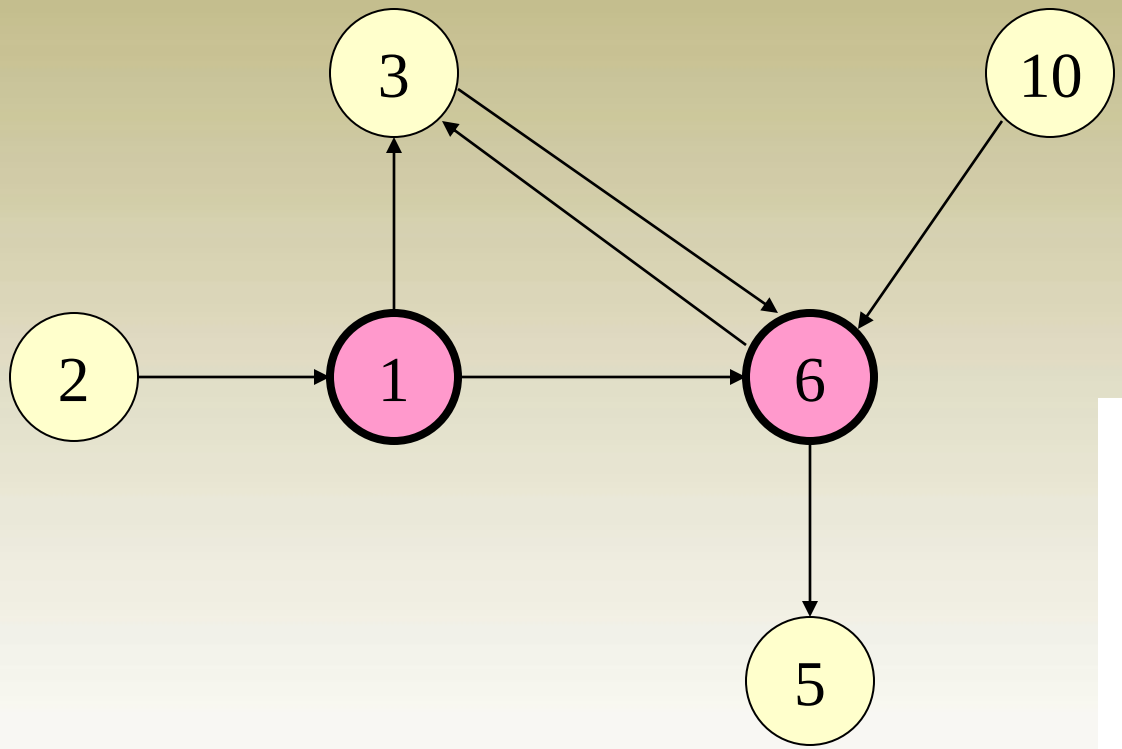
Ομοιότητες SALSA με HITS και PageRank

- Επινοήθηκε από τους Ronny Lempel και Shlomo Moran το 2000
- Συνδυασμός HITS και PageRank
- Ο SALSA χρησιμοποιεί authority και hub score
- Ο SALSA δημιουργεί ένα neighborhood graph χρησιμοποιώντας authority και hub ιστοσελίδες και υπερσυνδέσμους

Διαφορές SALSA με HITS και PageRank

- Η μέθοδος SALSA δημιουργεί ένα διμερές γράφημα (bipartite graph) των σελίδων authority και hub στο neighborhood γράφημα
- Το ένα σύνολο περιέχει τις hub σελίδες
- Το άλλο σύνολο περιέχει τις authority σελίδες
- Μια σελίδα μπορεί να περιέχεται και στα δυο σύνολα

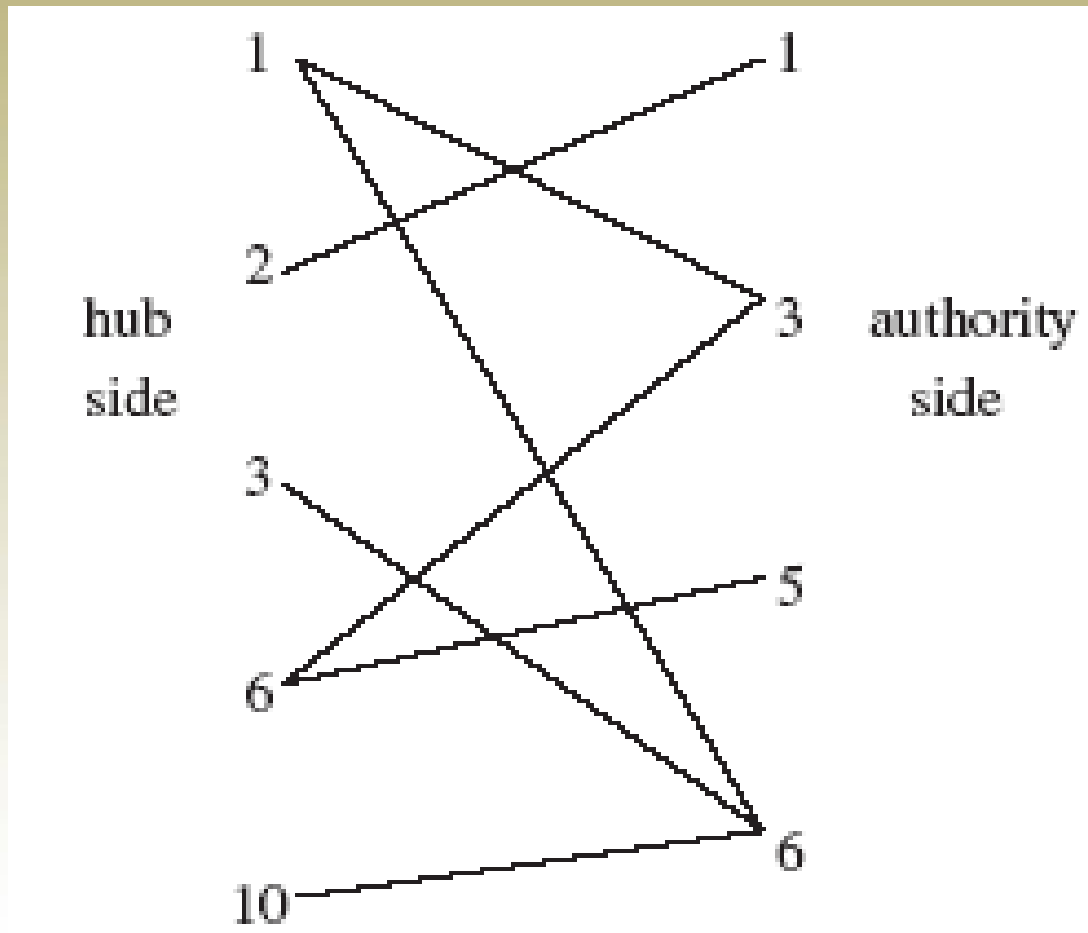
Neighborhood Graph N



$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Διμερές γράφημα G του Neighborhood Graph N

$$V_h = \{1, 2, 3, 6, 10\},$$
$$V_a = \{1, 3, 5, 6\}.$$



Markov αλυσίδες

- Από το διμερές γράφημα G σχηματίζονται δυο πίνακες
  - Μια hub Markov chain με πίνακα **H**
  - Μια authority Markov chain με πίνακα **A**
- Οι πίνακες **H** και **A** μπορεί να παραχθούν από τον πίνακα γειτνίασης (adjacency matrix) **L** που έχουμε δει στον υπολογισμό του HITS και του PageRank
- Ο HITS χρησιμοποιεί τον unweighted matrix **L**
- Ο PageRank χρησιμοποιεί τη row-weighted έκδοση του πίνακα **L**
- SALSA χρησιμοποιεί row και column weighting

Πώς υπολογίζονται οι **H** και **A**;

- Έστω ότι $\mathbf{L}_r$ είναι ο **L** με κάθε μη-μηδενική γραμμή του να διαιρείται με το άθροισμά της
- Έστω ότι $\mathbf{L}_c$ είναι ο **L** με κάθε μη-μηδενική στήλη του να διαιρείται με το άθροισμά της

Παράδειγμα των $\mathbf{L}_r$ και $\mathbf{L}_c$

$$\mathbf{L} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}, \qquad \mathbf{L}_r = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array},$$

and

$$\mathbf{L}_c = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{3} & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{3} & 0 \end{pmatrix} \end{array}.$$

Οι πίνακες **H** και **A**

- Ο πίνακας **H**, SALSA hub matrix, αποτελείται από τις μη-μηδενικές γραμμές και στήλες του $\mathbf{L}_r\mathbf{L}_c^T$

- Ο πίνακας **A**, SALSA authority matrix, αποτελείται από τις μη-μηδενικές γραμμές και στήλες του πίνακα $\mathbf{L}_c^T\mathbf{L}_r$

Οι πίνακες $\mathbf{L}_r\mathbf{L}_c{}^T$ και $\mathbf{L}_c{}^T\mathbf{L}_r$

$$\mathbf{L}_r\mathbf{L}_c^T = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} \frac{5}{12} & 0 & \frac{2}{12} & 0 & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{4} & 0 & 1 & 0 & \frac{3}{4} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix} \end{array},$$

$$\mathbf{L}_c^T\mathbf{L}_r = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 5 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 5 & 6 & 10 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & \frac{5}{6} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

Οι πίνακες **H** και **A**

$$\mathbf{H} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 6 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 6 & 10 \end{array} \\ \begin{pmatrix} \frac{5}{12} & 0 & \frac{2}{12} & \frac{3}{12} & \frac{2}{12} \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & \frac{3}{4} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix} \end{array}.$$

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ 3 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 3 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{6} & 0 & \frac{5}{6} \end{pmatrix} \end{array}.$$

Ιδιοδιανύσματα

- $Av = \lambda v$
- $v^T A = \lambda\, v^T$
- Αριθμητικός υπολογισμός: Power μέθοδος

Η power μέθοδος

- $X_{k+1} = AX_k$
- $X_{k+1}^T = X_k^T A$
- Συγκλίνει στο κυρίαρχο ιδιοδιάνυσμα (dominant eigenvector), δηλ., σ' αυτό που αντιστοιχεί στην κυρίαρχη ιδιοτιμή ($\lambda = 1$).

Η power μέθοδος

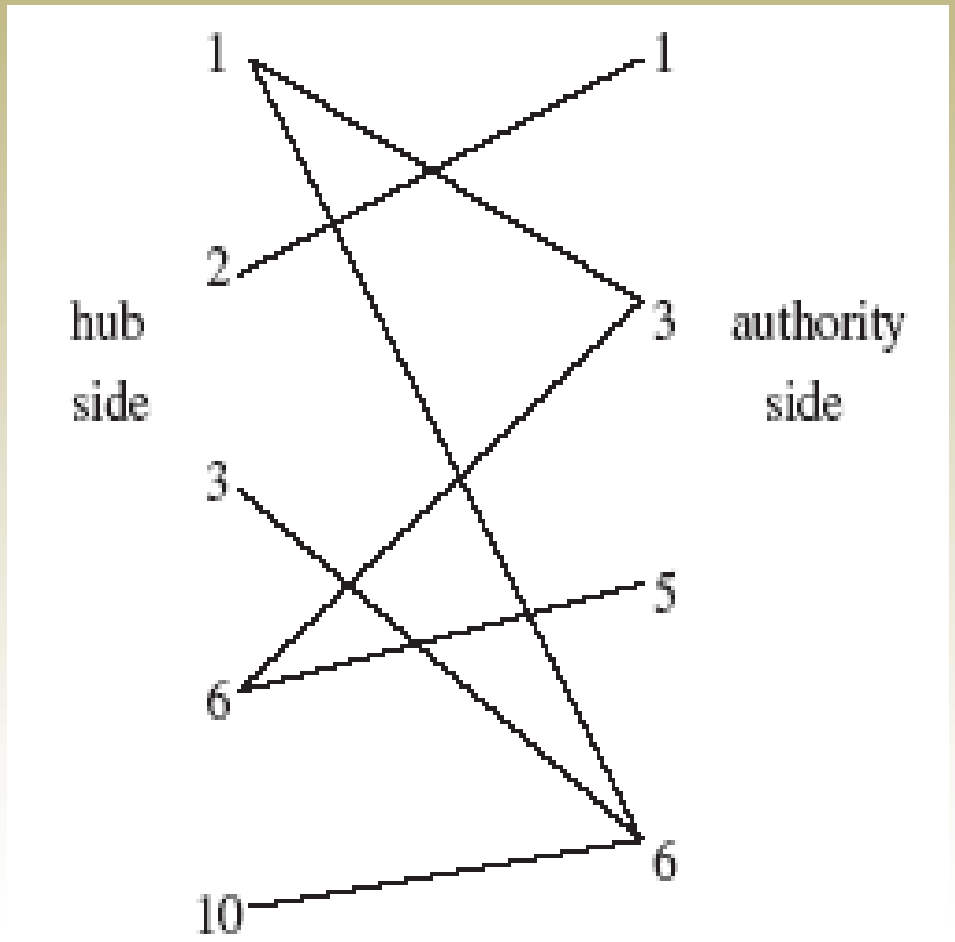
- Οι πίνακες **H** και **A** πρέπει να είναι irreducible, ώστε να συγκλίνει η power μέθοδος σε ένα unique eigenvector, ξεκινώντας από κάποια αρχική τιμή

- Εάν το neighborhood γράφημα **G** είναι συνδεδεμένο (connected), τότε ο **H** και ο **A** είναι irreducible

- Εάν το **G** δεν είναι συνδεδεμένο, τότε η εκτέλεση της power μεθόδου στον **H** και **A** δεν θα έχει ως αποτέλεσμα τη σύγκλιση σε ένα μοναδικό κυρίαρχο ιδιοδιάνυσμα

Στο παράδειγμα το G δεν είναι συνδεδεμένο

- Είναι προφανές ότι το γράφημα δεν είναι συνδεδεμένο, αφού η σελίδα 2 στο σύνολο hub συνδέεται μόνο με τη σελίδα 1 στο σύνολο authority, και αντίστροφα
- Οι **H** και **A** είναι reducible και επομένως περιέχουν multiple irreducible connected components



Connected Components

- Ο πίνακας **H** περιέχει δυο connected components, C = {2} και D = {1, 3, 6, 10}

- Ο πίνακας **A** περιέχει δυο connected components, E = {1} και F = {3, 5, 6}

Cutting και Pasting. Μέρος I

- Εκτελούμε την power method σε κάθε συνιστώσα των **H** και **A**

$$\pi_h^T(C) = \begin{matrix} 2 \\ (1), \end{matrix} \quad \pi_h^T(D) = \begin{matrix} 1 & 3 & 6 & 10 \\ (\frac{1}{3} & \frac{1}{6} & \frac{1}{3} & \frac{1}{6}), \end{matrix}$$

$$\pi_a^T(E) = \begin{matrix} 1 \\ (1), \end{matrix} \quad \pi_a^T(F) = \begin{matrix} 3 & 5 & 6 \\ (\frac{1}{3} & \frac{1}{6} & \frac{1}{2}). \end{matrix}$$

Cutting και Pasting. Μέρος II

- Ενώνουμε τις δυο συνιστώσες για κάθε πίνακα
- Πρέπει να πολλαπλασιάσουμε κάθε στοιχείο του διανύσματος με το κατάλληλο weight

H:

$$\pi_h^T = \begin{matrix} 1 & 2 & 3 & 6 & 10 \end{matrix}$$
$$\left(\frac{4}{5}\cdot\frac{1}{3} \quad \frac{1}{5}\cdot 1 \quad \frac{4}{5}\cdot\frac{1}{6} \quad \frac{4}{5}\cdot\frac{1}{3} \quad \frac{4}{5}\cdot\frac{1}{6}\right)$$
$$\begin{matrix} 1 & 2 & 3 & 6 & 10 \end{matrix}$$
$$= \left(.2667 \quad .2 \quad .1333 \quad .2667 \quad .1333\right).$$

A:

$$\pi_h^T = \begin{matrix} 1 & 3 & 5 & 6 \end{matrix}$$
$$\left(\frac{1}{4}\cdot 1 \quad \frac{3}{4}\cdot\frac{1}{3} \quad \frac{3}{4}\cdot\frac{1}{6} \quad \frac{3}{4}\cdot\frac{1}{2}\right)$$
$$\begin{matrix} 1 & 3 & 5 & 6 \end{matrix}$$
$$= \left(.25 \quad .25 \quad .125 \quad .375\right).$$

SALSA hub ranking: 1/6 2 3/10

HITS hub ranking: 1 3/6/10 2 5

SALSA authority ranking: 6 1/3 5

HITS hub ranking: 6 3 5 1 2/10

Πλεονεκτήματα/Μειονεκτήματα του

- Δεν εμφανίζει το φαινόμενο topic drift, όπως ο HITS
- Παρέχει authority και hub scores
- Χειρίζεται το spamming καλύτερα από ότι ο HITS, αλλά όχι τόσο καλά όσο ο PageRank
- Είναι query-dependent