

# ΜΥΕ003: Ανάκτηση Πληροφορίας

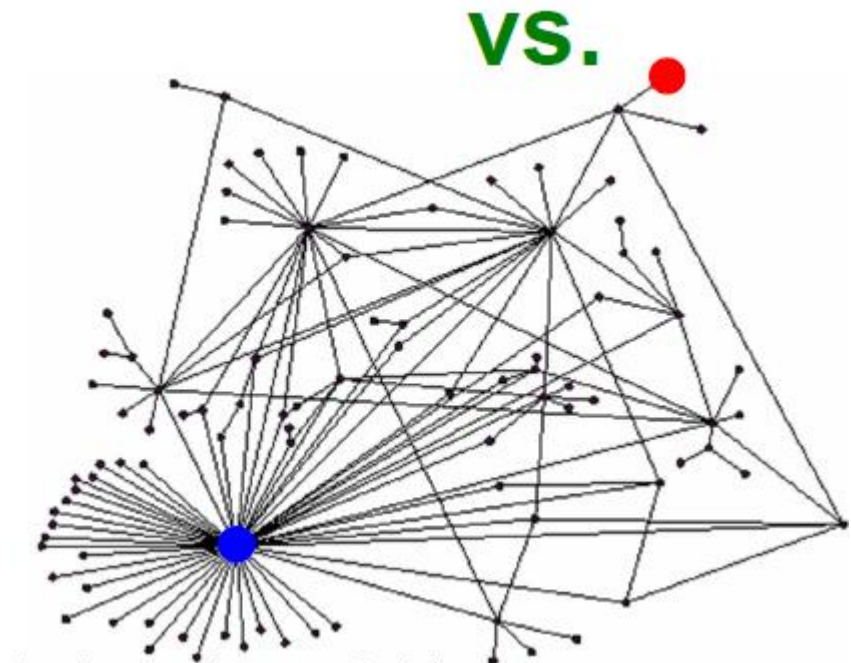
*Διδάσκουσα: Ευαγγελία Πιτουρά*

Κεφάλαιο 21: Ανάλυση Συνδέσμων.

# Τι θα δούμε σήμερα

*Πως μπορούμε να χρησιμοποιήσουμε το δίκτυο στη διάταξη των αποτελεσμάτων*

Δεν είναι όλες οι σελίδες (κόμβοι) ίσες.  
Ποιες σελίδες είναι «σημαντικές»?



# Τι θα δούμε σήμερα

## Ανάλυση συνδέσμων (link analysis)

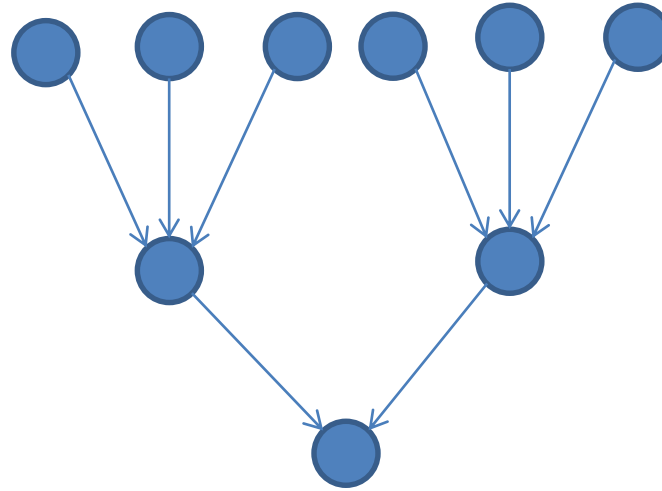
- PageRank
- HITS (Κομβικές σελίδες και σελίδες κύρους)
- SALSA

Υπολογισμός μιας τιμής ανά σελίδα (κόμβο) που εκφράζει το πόσο σημαντική είναι στο δίκτυο

# Διάταξη με βάση τη δημοτικότητα

Διάταξη των σελίδων με βάση τον αριθμό των εισερχόμενων ακμών (**in-degree**, **degree centrality**)

# Αρκεί η δημοτικότητα;



- Δεν είναι σημαντικό *πόσοι κόμβοι* δείχνουν σε μια σελίδα αλλά το *πόσο σημαντικοί* είναι αυτοί οι κόμβοι

# PageRank

# PageRank

- Βασική ιδέα: Μια σελίδα είναι σημαντική αν δείχνουν σε αυτήν σημαντικές σελίδες
- Η αξία (PageRank) ενός κόμβου είναι το **άθροισμα** της αξίας των φίλων του

# PageRank: Βασική ιδέα

Έχουμε μια «μονάδα κύρους» που τη λέμε PageRank ( $r$ ) και την μοιράζουμε στις σελίδες.

Κάθε σελίδα έχει ένα PageRank

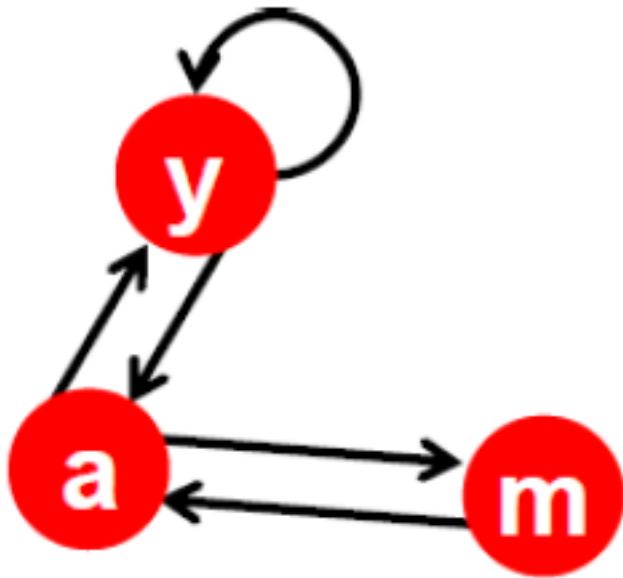
- Κάθε σελίδα *μοιράζει το PageRank στις σελίδες που δείχνει*
- Το PageRank μιας σελίδας είναι το *άθροισμα των PageRank των σελίδων που δείχνουν σε αυτήν*



# PageRank: Ορισμός

$$\text{PageRank}(v) = \sum_{u \in \text{inneighbors}(v)} \frac{\text{PageRank}(u)}{\text{outdegree}(u)}$$

# PageRank: παράδειγμα



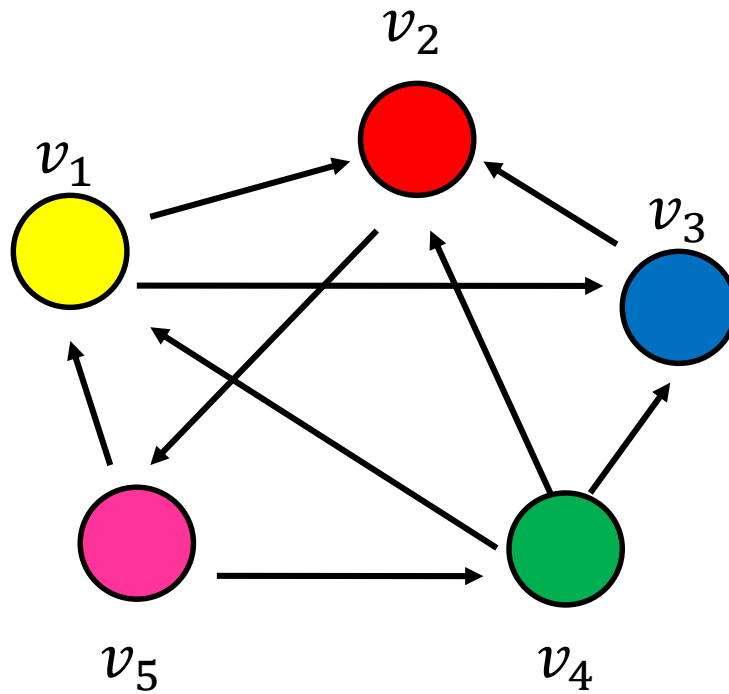
$$r_j = \sum_{i \rightarrow j} \frac{r_i}{d_i}$$

$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2 + r_m$$

$$r_m = r_a/2$$

# PageRank: παράδειγμα



# PageRank: παράδειγμα

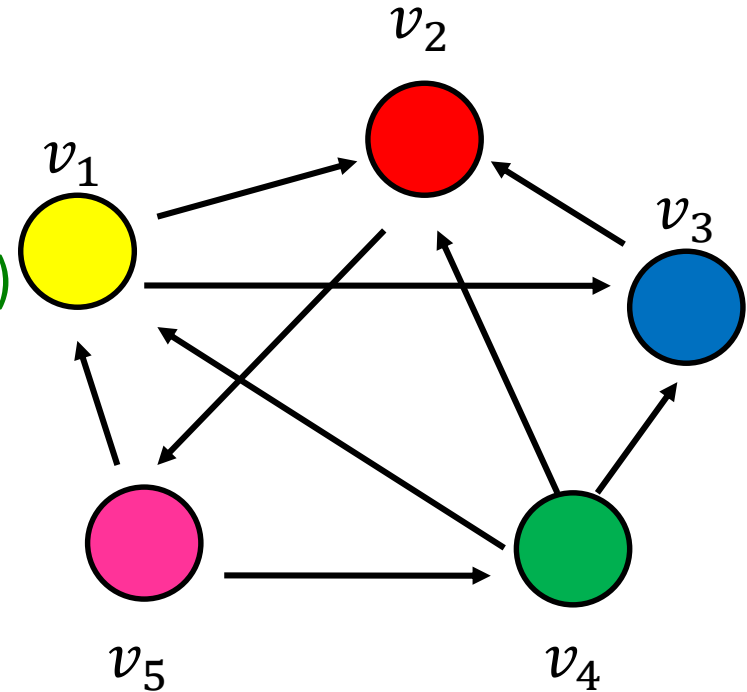
$$r(u_1) = 1/3 r(u_4) + 1/2 r(u_5)$$

$$r(u_2) = 1/2 r(u_1) + r(u_3) + 1/3 r(u_4)$$

$$r(u_3) = 1/2 r(u_1) + 1/3 r(u_4)$$

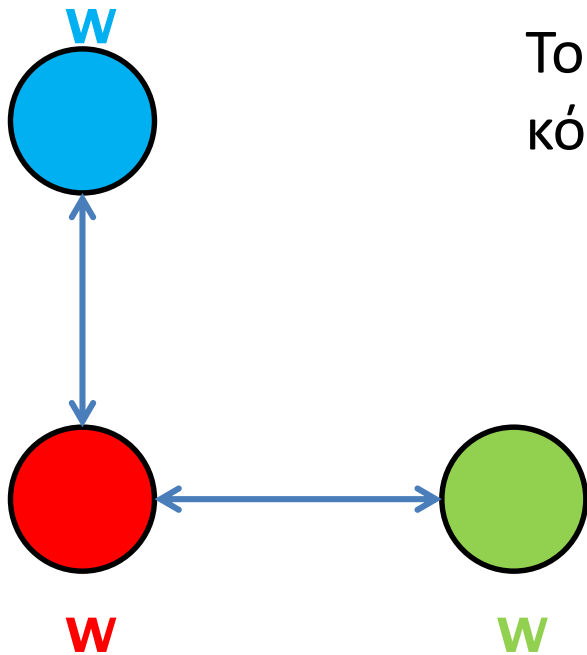
$$r(u_4) = 1/2 r(u_5)$$

$$r(u_5) = r(u_2)$$



$$r(u_1) + r(u_2) + r(u_3) + r(u_4) + r(u_5) = 1 \quad r(u): \text{PageRank}(u)$$

# Ένα απλό παράδειγμα: υπολογισμός



Το συνολικό PageRank μοιράζεται στους 3 κόμβους

$$w + w + w = 1$$

$$w = w + w$$

$$w = \frac{1}{2} w$$

$$w = \frac{1}{2} w$$

- Λύνοντας το σύστημα εξισώσεων παίρνουμε το PageRank των κόμβων

$$w = \frac{1}{2} \quad w = \frac{1}{4} \quad w = \frac{1}{4}$$

# PageRank: Επαναληπτικός Αλγόριθμος

Σε ένα γράφο με  $n$  nodes, αναθέτουμε σε όλους το ίδιο αρχικό PageRank =  $1/n$ .

- Εκτελούμε μια ακολουθία από  $k$  ενημερώσεις των PageRank τιμών με βάση των παρακάτω κανόνα:
  1. Κάθε σελίδα **μοιράζει** την τρέχουσα PageRank τιμή της ισόποσα στις *εξερχόμενες ακμές και τις περνά στους αντίστοιχους κόμβους*
  2. Κάθε σελίδα **ανανεώνει** την PageRank τιμή της ώστε να είναι ίση με το άθροισμα των ποσών που δέχεται μέσω των εισερχόμενων ακμών της.

# PageRank: Αλγόριθμος

Επαναληπτικός υπολογισμός

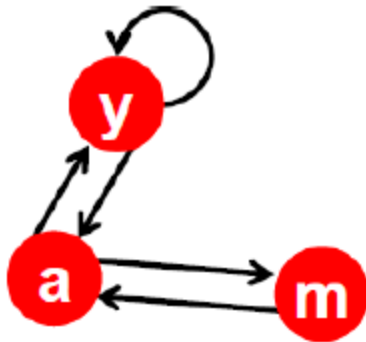
Initialize all PageRank weights to  $\frac{1}{n}$

Repeat:

$$w_v = \sum_{u \rightarrow v} \frac{1}{d_{out}(u)} w_u$$

Until the weights do not change

# Υπολογισμός: Παράδειγμα



$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2 + r_m$$

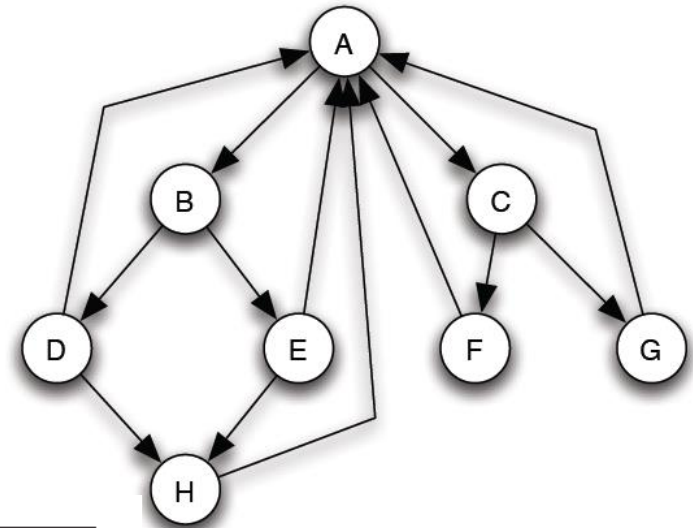
$$r_m = r_a/2$$

$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{matrix} 1/3 & 1/3 & 5/12 & 9/24 & & 6/15 \\ 1/3 & 3/6 & 1/3 & 11/24 & \dots & 6/15 \\ 1/3 & 1/6 & 3/12 & 1/6 & & 3/15 \end{matrix}$$



# Ένα μεγαλύτερο παράδειγμα

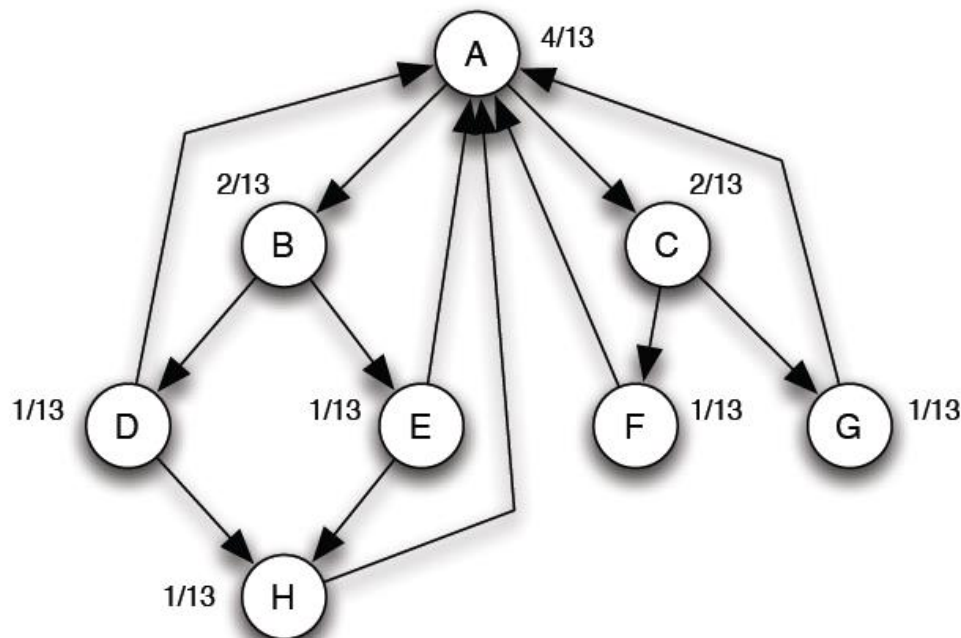
Αρχικά όλοι οι κόμβοι  
PageRank  $1/8$



Step	A	B	C	D	E	F	G	H
1	$1/2$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/16$	$1/8$
2	$3/16$	$1/4$	$1/4$	$1/32$	$1/32$	$1/32$	$1/32$	$1/16$

- Ένα είδος ροής (“fluid”) που κινείται στο δίκτυο
- Το συνολικό PageRank στο δίκτυο παραμένει σταθερό (δε χρειάζεται κανονικοποίηση)

# Ισορροπία



- Ένας απλός τρόπος να ελέγξουμε αν το σύνολο PageRank τιμών αντιστοιχεί σε **ισορροπία**: οι τιμές αθροίζουν σε 1 και **δεν αλλάζουν** αν εφαρμόσουμε τον κανόνα ενημέρωσης

# PageRank: Διανυσματική αναπαράσταση

Stochastic Adjacency Matrix – (Στοχαστικός) Πίνακας  
Γειτνίασης  $M$

Πίνακας  $M$  – πίνακας γειτνίασης του web

Αν  $j \rightarrow i$ , τότε  $M_{ij} = 1/\text{outdegree}(j)$

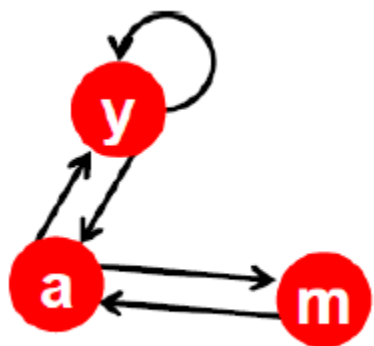
Αλλιώς,  $M_{ij} = 0$

Page Rank Vector  $r$

Ένα διάνυσμα με μία τιμή για κάθε σελίδα (το PageRank της σελίδας)

$$\sum r_i = 1$$

# PageRank: Διανυσματική αναπαράσταση



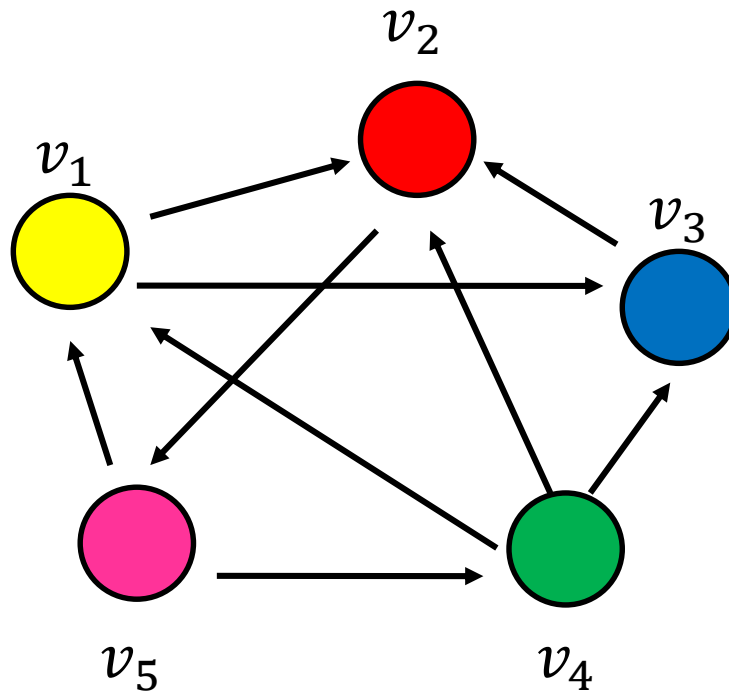
	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	1
m	0	$\frac{1}{2}$	0

Column stochastic: οι τιμές στις στήλες αθροίζουν στο 1

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix}$$

$$\mathbf{r} = \mathbf{M} \cdot \mathbf{r}$$

# PageRank: παράδειγμα



# PageRank: Διανυσματική αναπαράσταση

$$r_j^{(t+1)} = \sum_{i \rightarrow j} \frac{r_i^{(t)}}{d_i} \quad \text{or equivalently} \quad r = Mr$$

- Συγκλίνει;
- Συγκλίνει σε αυτό που θέλουμε;
- Ποια είναι η φυσική σημασία;

# Τυχαίος Περίπατος (Random Walks)

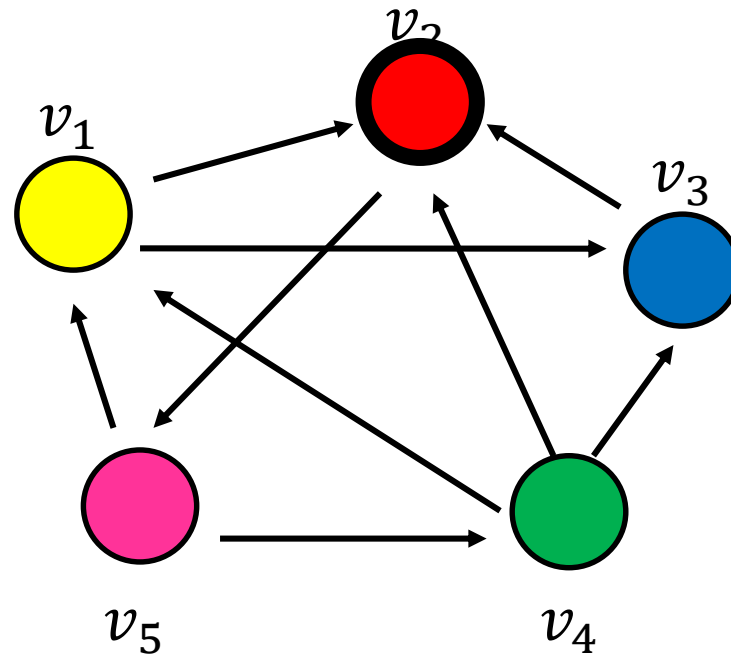
Ο αλγόριθμος προσομοιώνει ένα τυχαίο περίπατο (random walk) στο γράφο

Τυχαίος περίπατος:

- Ξεκίνα από κάποιον τυχαίο κόμβο (επιλεγμένο uniformly at random) με πιθανότητα  $1/n$
- Επέλεξε τυχαία (uniformly at random) μια από τις εξερχόμενες ακμές του κόμβου
- Ακολούθησε την ακμή
- Επανάλαβε

# Παράδειγμα

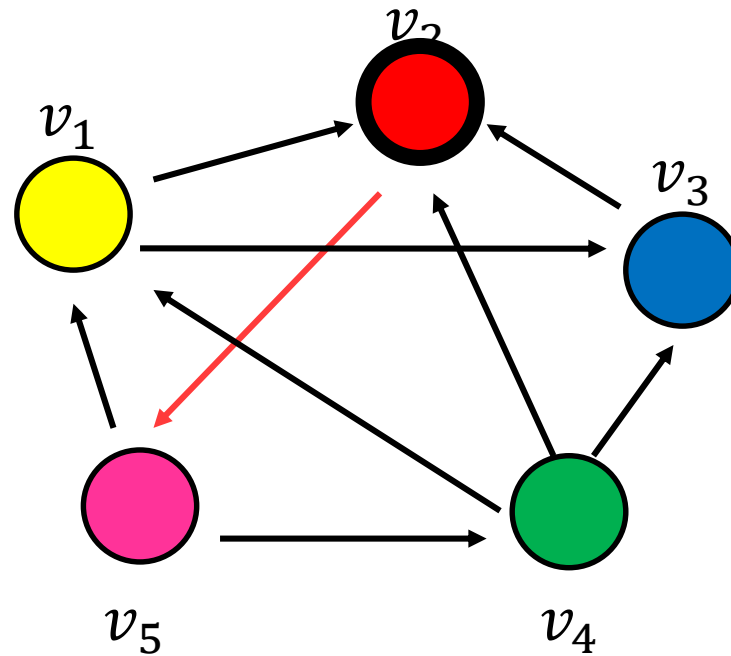
- Step 0





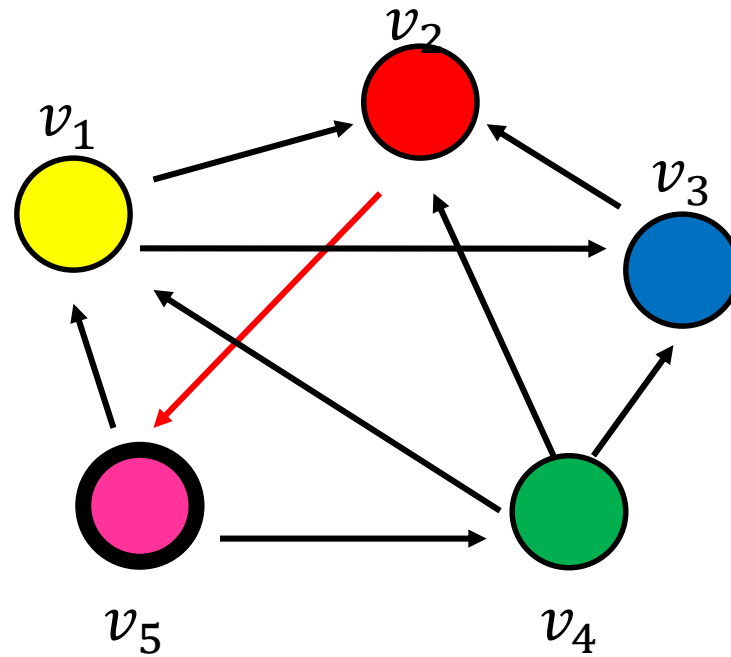
# Παράδειγμα

- Step 0



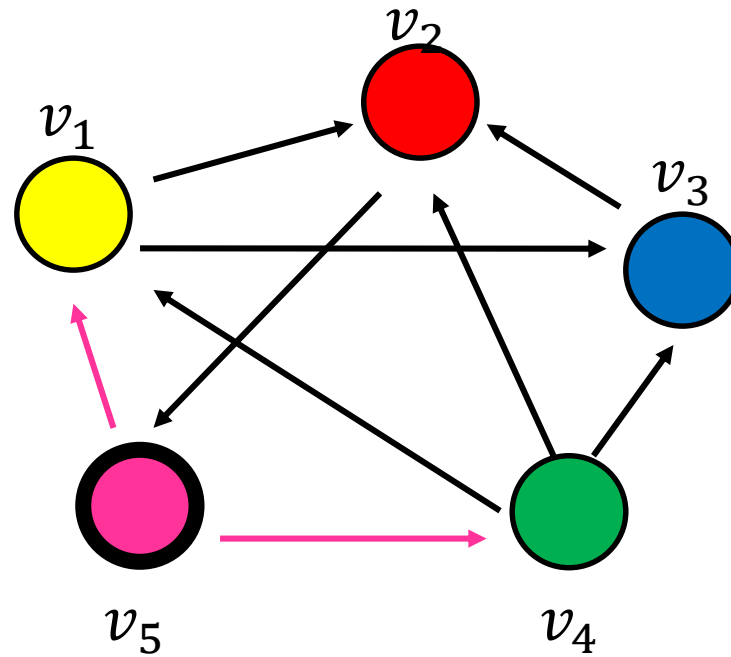
# Παράδειγμα

- Step 1



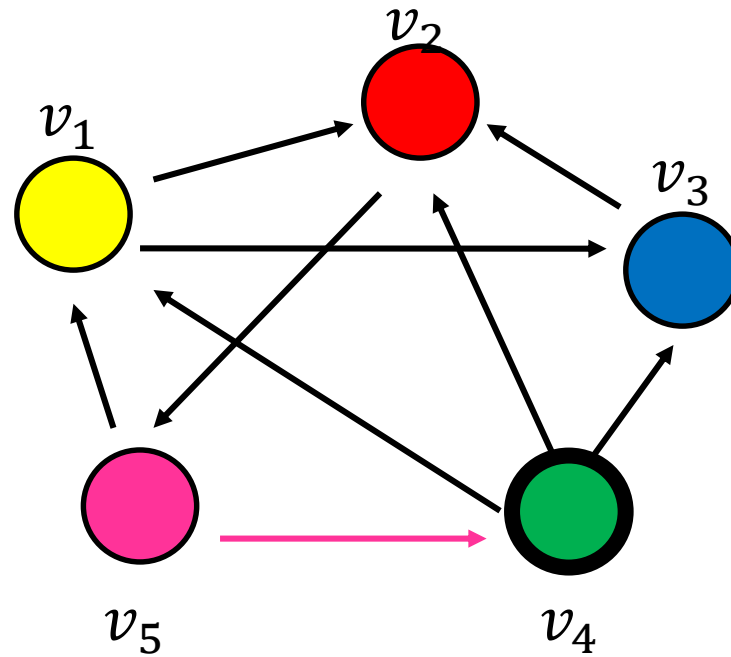
# Παράδειγμα

- Step 1



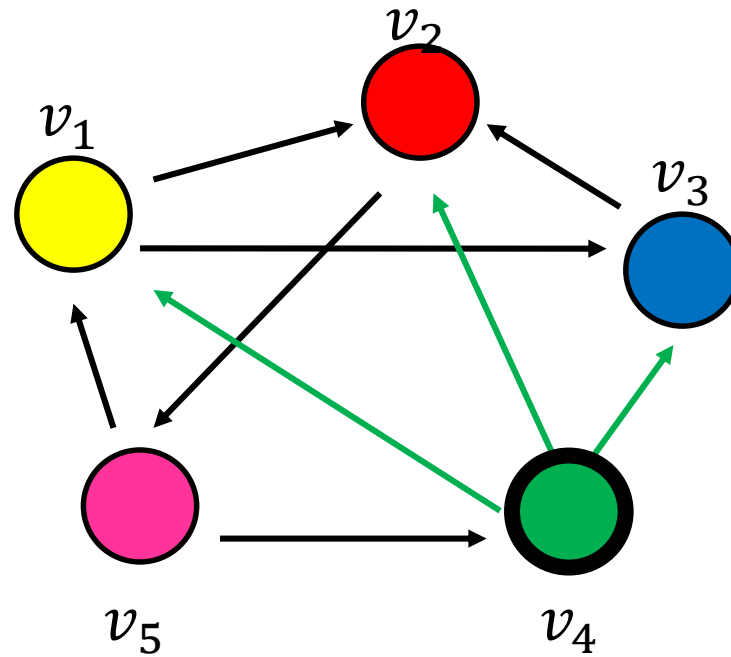
# Παράδειγμα

- Step 2



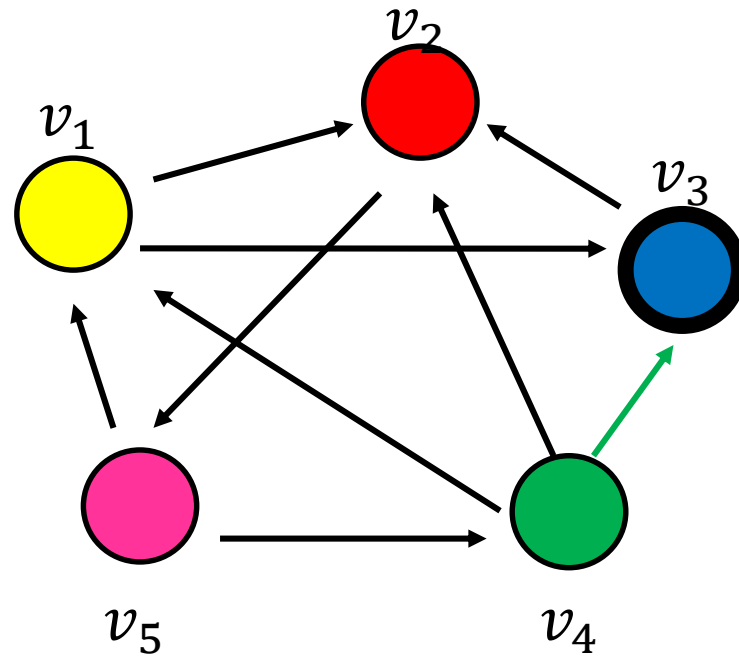
# Παράδειγμα

- Step 2



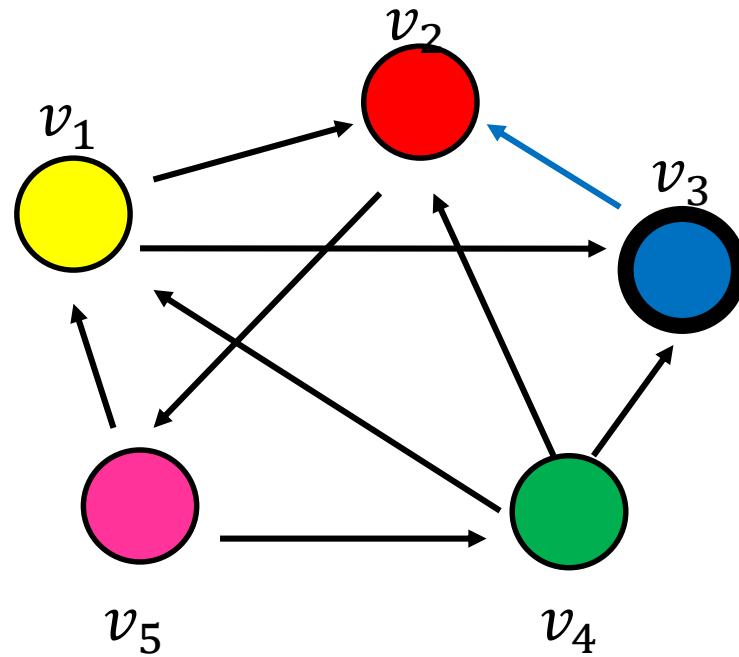
# Παράδειγμα

- Step 3



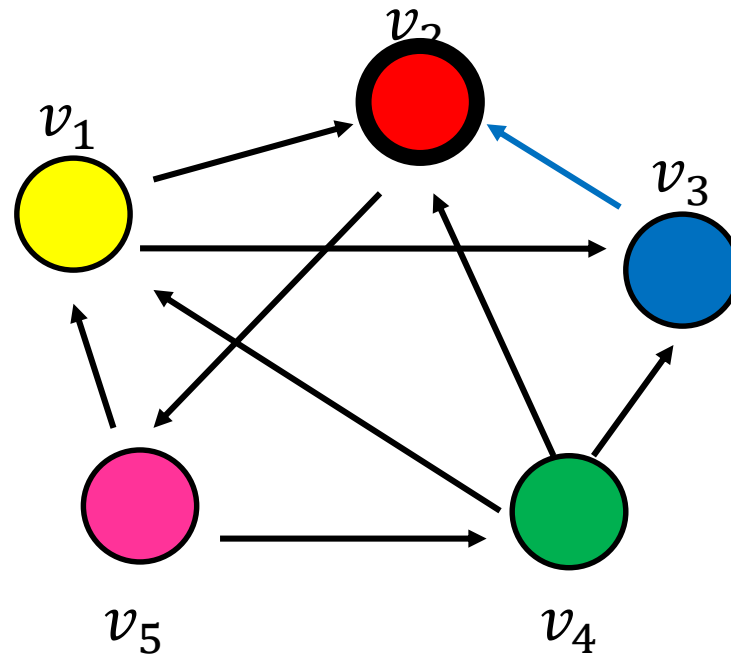
# Παράδειγμα

- Step 3



# Παράδειγμα

- Step 4...





# Τυχαίος Περίπατος

*Η πιθανότητα να είσαι στη σελίδα  $X$  μετά από  $k$  βήματα του τυχαίου περιπάτου είναι το  $PageRank$  της σελίδας  $X$  μετά από  $k$  επαναλήψεις του υπολογισμού του  $PageRank$*

Το μοντέλο του **Random Surfer**

Του χρήστη που τριγυρνά στο web, ξεκινώντας από μια τυχαία σελίδα και συνεχίζει ακολουθώντας τυχαία κάποια από τις συνδέσεις της σελίδας

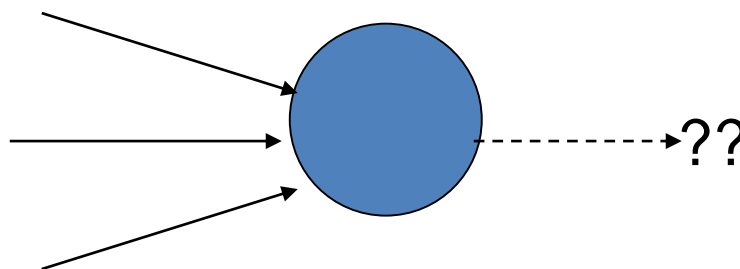
# PageRank: Επεκτάσεις

Δύο προβλήματα

1. **Dead ends**: σελίδες χωρίς εξερχόμενες ακμές  
Έχουν ως αποτέλεσμα να ξεφεύγει (leak out) to PageRank
2. **Spider traps**: Ομάδα σελίδων που όλες οι εξερχόμενες ακμές είναι μεταξύ τους  
Τελικά απορροφούν όλο το PageRank

# PageRank: Αδιέξοδα

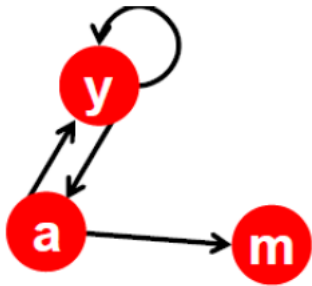
**Αδιέξοδα (dead ends):** σελίδες που δεν έχουν εξερχόμενες ακμές



Ο τυχαίος περίπατος μπορεί να κολλήσει σε ένα τέτοιο κόμβο

Λέγονται και **sink nodes**

# PageRank: Αδιέξοδα



	y	a	m
y	1/2	1/2	0
a	1/2	0	0
m	0	1/2	0

$$r_y = r_y/2 + r_a/2$$

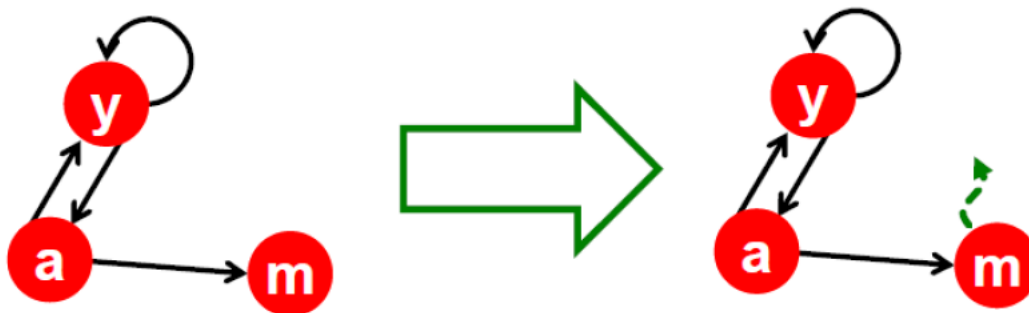
$$r_a = r_y/2$$

$$r_m = r_a/2$$

$$\begin{pmatrix} r_y \\ r_a \\ r_m \end{pmatrix} = \begin{pmatrix} 1/3 & 2/6 & 3/12 & 5/24 & \dots & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 1/6 & 1/12 & 2/24 & \dots & 0 \end{pmatrix}$$

# PageRank: Αδιέξοδα

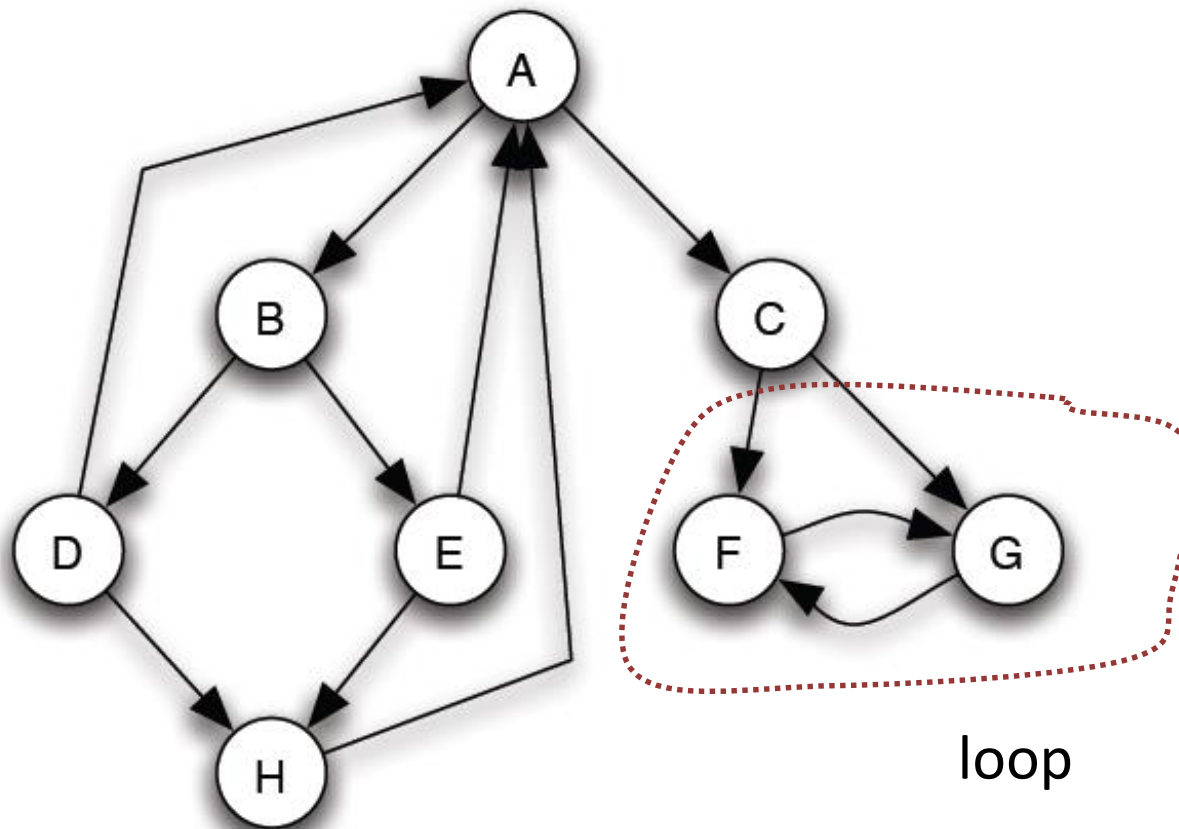
**Teleports:** ακολουθήσε με πιθανότητα 1 τυχαία links από τους αδιέξοδους κόμβους



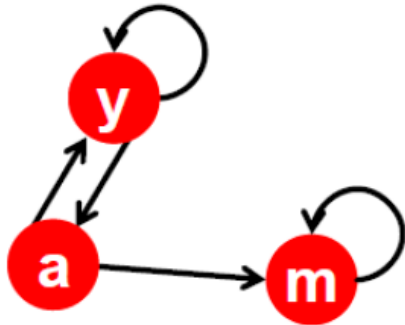
Αντίστοιχη τροποποίηση  
του πίνακα

	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$
a	$\frac{1}{2}$	0	$\frac{1}{3}$
m	0	$\frac{1}{2}$	$\frac{1}{3}$

# PageRank: Spider Traps



# PageRank: Spider Traps



	y	a	m
y	$\frac{1}{2}$	$\frac{1}{2}$	0
a	$\frac{1}{2}$	0	0
m	0	$\frac{1}{2}$	1

$$r_y = r_y/2 + r_a/2$$

$$r_a = r_y/2$$

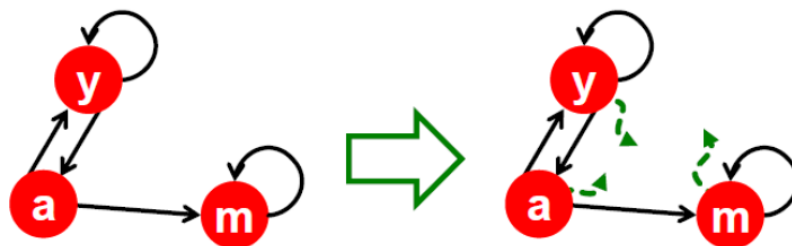
$$r_m = r_a/2 + r_m$$

$$\begin{bmatrix} r_y \\ r_a \\ r_m \end{bmatrix} = \begin{matrix} 1/3 & 2/6 & 3/12 & 5/24 & & 0 \\ 1/3 & 1/6 & 2/12 & 3/24 & \dots & 0 \\ 1/3 & 3/6 & 7/12 & 16/24 & & 1 \end{matrix}$$

# Random Walks with Jumps

## Τυχαία περίπατοι με «άλματα»

Με πιθανότητα  $\beta$ , ο περιπατητής ακολουθεί μια τυχαία εξερχόμενη ακμή όπως πριν και με πιθανότητα  $1-\beta$  επιλέγει (jumps) σε μια τυχαία σελίδα στο δίκτυο, επιλεγμένη με πιθανότητα  $(1/n)$





# Random Walks with Jumps

Brin-Page, 1998

$$r_j = \sum_{i \rightarrow j} \beta \frac{r_i}{d_i} + (1 - \beta) \frac{1}{n} \quad bM + (1-b) \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

Προτεινόμενη τιμή 0,8 – 0,9

## Μοντέλο του Random Surfer

Του χρήστη που περιπλανιέται στο web, ξεκινώντας από μια τυχαία σελίδα συνεχίζει ακολουθώντας τυχαία συνδέσεις ή με κάποια πιθανότητα βαριέται και πάει (jumps) σε μια άλλη τυχαία σελίδα

5 links και 1 jump

# PageRank, τυχαίοι περίπατοι και αλυσίδες Markov

# Τυχαίος περίπατος

Ποια είναι η πιθανότητα  $p_i^t$  να είσαι στον κόμβο  $i$  μετά από  $t$  βήματα;

$$p_1^0 = \frac{1}{5}$$

$$p_2^0 = \frac{1}{5}$$

$$p_3^0 = \frac{1}{5}$$

$$p_4^0 = \frac{1}{5}$$

$$p_5^0 = \frac{1}{5}$$

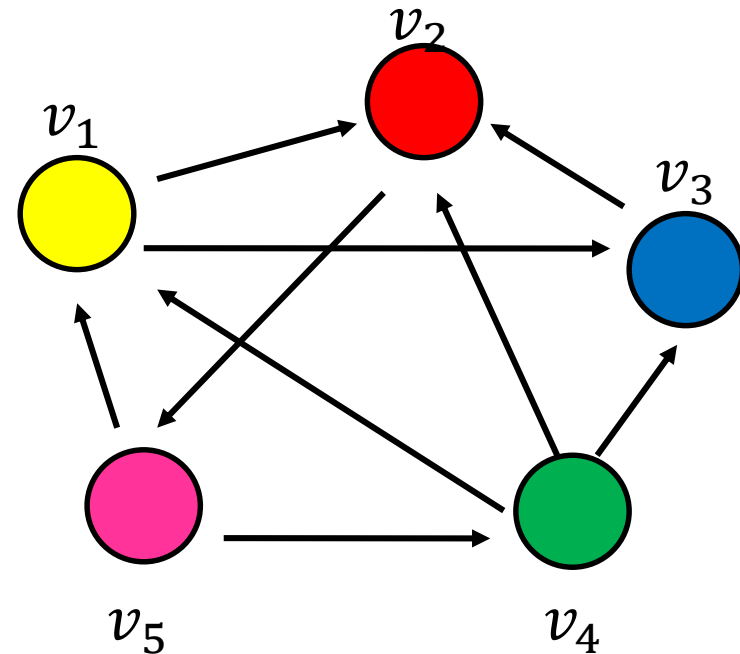
$$p_1^t = \frac{1}{3}p_4^{t-1} + \frac{1}{2}p_5^{t-1}$$

$$p_2^t = \frac{1}{2}p_1^{t-1} + p_3^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_3^t = \frac{1}{2}p_1^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_4^t = \frac{1}{2}p_5^{t-1}$$

$$p_5^t = p_2^{t-1}$$



# Αλυσίδες Markov

- Περιγράφουν μια **στοχαστική διαδικασία διακριτού χρόνου** σε ένα σύνολο από καταστάσεις  $S$   
$$S = \{s_1, s_2, \dots, s_n\}$$
με βάση έναν πίνακα πιθανότητας μεταβάσης (**transition probability matrix**)  $P$   
Όπου  $P[i, j]$  είναι η πιθανότητα  $s_i \rightarrow s_j$  να μεταβούμε στην κατάσταση  $s_j$  όταν είμαστε στην κατάσταση  $s_i$
- Οι γραμμές αθροίζουν σε 1 (row stochastic)
- **Memoryless**: η επόμενη κατάσταση εξαρτάται μόνο από την τωρινή κατάσταση και όχι από τον παρελθόν της διαδικασίας

# Αλυσίδες Markov

- State probability distribution vector

$$p^t = (p_1^t, p_2^t, \dots, p_n^t)$$

Διάνυσμα που αποθηκεύει την πιθανότητα να είμαστε στην κατάσταση  $p_i$  μετά από  $t$  βήματα

- Μπορούμε να το υπολογίσουμε ως:

$$p^t = P p^{t-1}$$

- Το state probability vector **συγκλίνει** σε μια μοναδική κατανομή αν η αλυσίδα είναι *μη περιοδική* (aperiodic) και *αμείωτη* (irreducible)

# Αλυσίδες Markov

*Irreducible*: υπάρχει πάντα μια ακολουθία μεταβάσεων με μη μηδενική πιθανότητα από μια οποιαδήποτε κατάσταση σε μία άλλη (connectivity)

*Aperiodicity*: οι καταστάσεις δε μπορούν να χωριστούν σε σύνολα τέτοια ώστε όλες οι μεταβάσεις να συμβαίνουν κυκλικά από το ένα σύνολο στο άλλο

# Τυχαίοι Περίπατοι

Οι τυχαίοι περίπατοι στους γράφους αντιστοιχούν σε Αλυσίδες Markov

- Το σύνολο των καταστάσεων  $S$  είναι οι κόμβοι του γράφου
- Ο πίνακας πιθανοτήτων μετάβασης είναι η πιθανότητα να ακολουθήσουμε μια ακμή από έναν κόμβο σε ένα άλλο

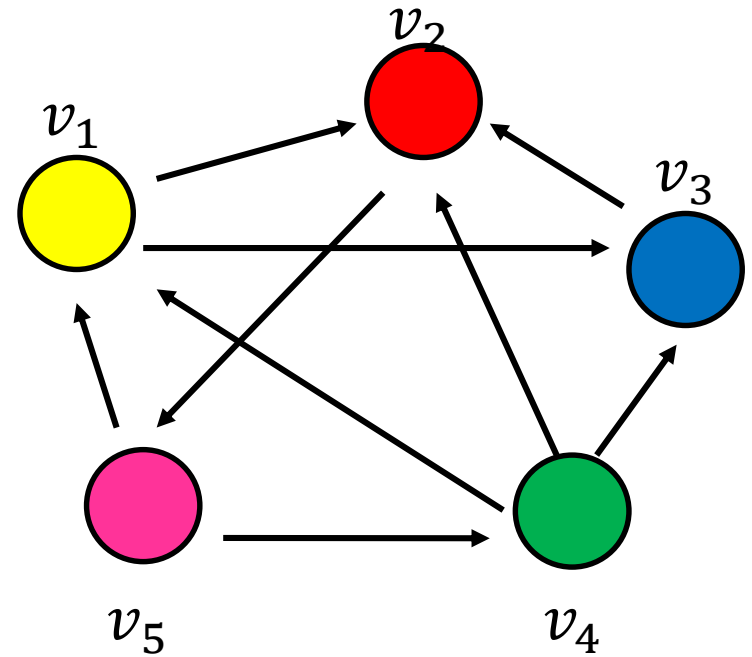
$$P[i, j] = 1/\text{outdegree}(i)$$

Στα επόμενα θα θεωρήσουμε τον ανάστροφο (transpose) του πίνακα  $M$

# Ένα παράδειγμα

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$





# Το διάνυσμα πιθανοτήτων

$$p^t = (p_1^t, p_2^t, \dots, p_n^t)$$

Διάνυσμα που αποθηκεύει την πιθανότητα να είμαστε στον κόμβο  $u_i$  μετά από  $t$  βήματα

- $p_i^0$  πιθανότητα να αρχίσουμε από τον κόμβο  $i$  (συνηθώς) ομοιόμορφη
- $p^t = P p^{t-1}$

# Ένα παράδειγμα

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

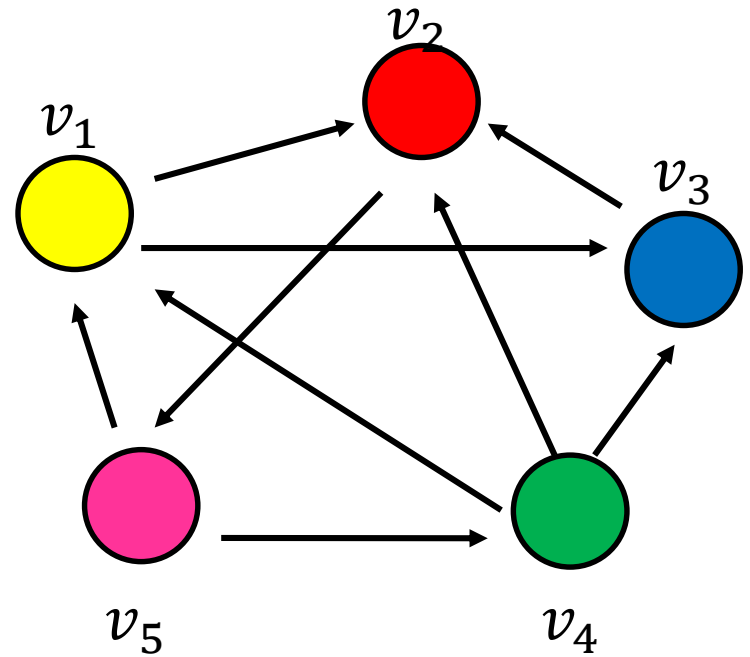
$$p_1^t = \frac{1}{3}p_4^{t-1} + \frac{1}{2}p_5^{t-1}$$

$$p_2^t = \frac{1}{2}p_1^{t-1} + p_3^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_3^t = \frac{1}{2}p_1^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_4^t = \frac{1}{2}p_5^{t-1}$$

$$p_5^t = p_2^{t-1}$$



# Stationary distribution

- Η stationary κατανομή ενός τυχαίου περίπατου με πίνακα μετάβασης  $P$  είναι η κατανομή πιθανοτήτων  $\pi$  τέτοια ώστε
  - $\pi = \pi P$
- Το **ιδιοδιάνυσμα** (principal left eigenvector) του πίνακα  $P$  (οι στοχαστικοί πίνακες έχουν μέγιστη ιδιοτιμή 1)
- Το ποσοστό των φορών που επισκεπτόμαστε την κατάσταση (κόμβο)  $i$  όταν  $t \rightarrow \infty$
- Θεωρία Αλυσίδων Markov: Ο τυχαίος περίπατος **συγκλίνει** σε μια μοναδική stationary distribution ανεξάρτητα από την αρχική κατάσταση αν ο γράφος είναι **ισχυρά συνεκτικός** και δεν είναι **διμερής**

# Υπολογισμός

- Power Method

Initialize  $q^0$  to some distribution  
Repeat  
 $q^t = q^{t-1}P$   
Until convergence

- Μετά από πολλές επαναλήψεις  $q^t \rightarrow \pi$  ανεξάρτητα από το αρχικό διάνυσμα  $q^0$
- Power method γιατί υπολογίζει το  $q^t = q^0 P^t$
- Ρυθμός σύγκλισης
  - Καθορίζεται από τη δεύτερη ιδιοτιμή  $\lambda_2^t$

# Stationary distribution

- Τι σημαίνει η stationary distribution  $\pi$  ενός τυχαίου περίπατου
- $\pi(i)$ : η πιθανότητα να είμαστε στον κόμβο  $i$  μετά από ένα πολύ μεγάλο (άπειρο) αριθμό από βήματα
- $\pi = p_0 P^\infty$ , όπου  $P$  ο πίνακας μετάβασης,  $p_0$  το αρχικό διάνυσμα
  - $P(i, j)$ : πιθανότητα μετάβασης από το  $i$  στο  $j$  σε ένα βήμα
  - $P^2(i, j)$ : πιθανότητα μετάβασης από το  $i$  στο  $j$  σε δύο βήματα (πιθανότητα όλων των μονοπατιών μήκους 2)
  - $P^\infty(i, j) = \pi(j)$ : πιθανότητα μετάβασης από το  $i$  στο  $j$  σε άπειρα βήματα – δεν έχει σημασία το αρχικό σημείο

# Περίληψη: ορισμός

Ένας κόμβος είναι σημαντικός αν δείχνουν σε αυτόν σημαντικοί κόμβοι

Κάθε κόμβος έχει ένα **PageRank**

- Κάθε κόμβος **μοιράζει** το PageRank του **στους κόμβους που δείχνει**
- Το PageRank ενός κόμβου είναι **το άθροισμα** των PageRank **των κόμβων που δείχνουν σε αυτόν**

$$\text{PageRank}(v) = \sum_{u \rightarrow v} \frac{\text{PageRank}(u)}{\text{outdegree}(u)}$$

# Περίληψη: παράδειγμα

$$r(u_1) = 1/3 r(u_4) + 1/2 r(u_5)$$

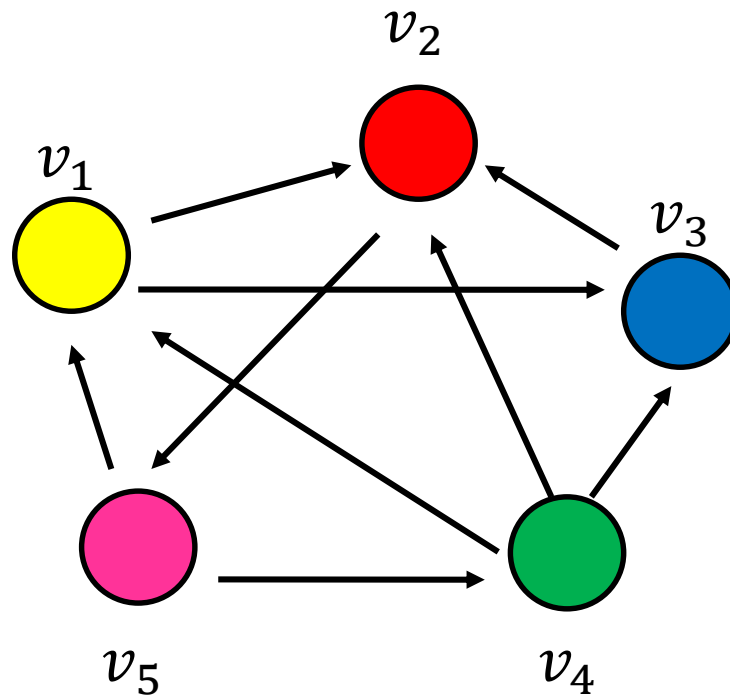
$$r(u_2) = 1/2 r(u_1) + r(u_3) + 1/3 r(u_4)$$

$$r(u_3) = 1/2 r(u_1) + 1/3 r(u_4)$$

$$r(u_4) = 1/2 r(u_5)$$

$$r(u_5) = r(u_2)$$

$r(u)$ : PageRank( $u$ )



# Περίληψη: αλγόριθμος

Επαναληπτικός υπολογισμός (power iteration method)

Initialize  $r^0(v) \leftarrow \frac{1}{n}$

$t = 1$

repeat

$$r^t(v) \leftarrow \sum_{u \rightarrow v} \frac{r^{t-1}(u)}{\text{outdegree}(u)}$$

$t = t + 1$

until convergence



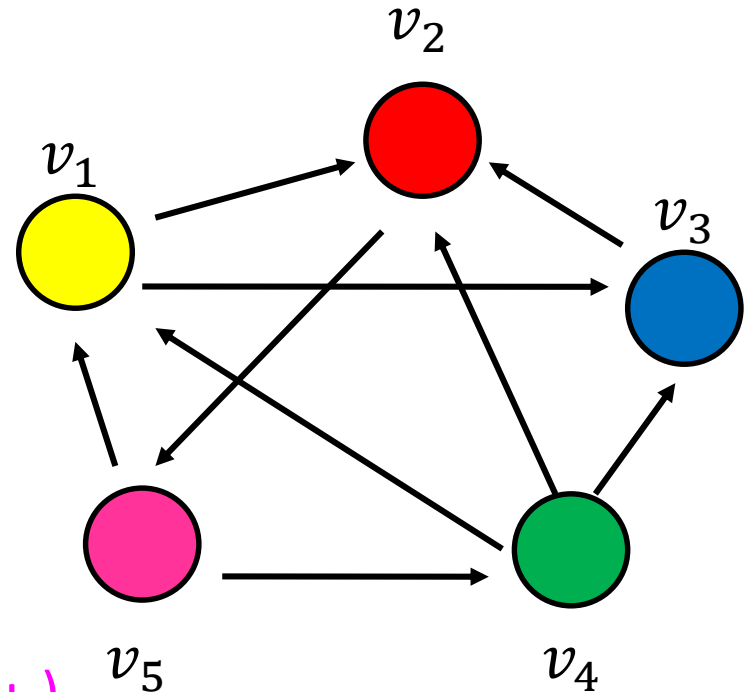
# Περίληψη: πίνακες

$M$ :  $M[i, j]$   $1/\text{outdegree}(j)$  αν  $j \rightarrow i$ , 0, αλλιώς

$r$  το διάνυσμα στήλη με τις PageRank τιμές

$$r = M r$$

$$M = \begin{bmatrix} 0 & 0 & 0 & 1/3 & 1/2 \\ 1/2 & 0 & 1 & 1/3 & 0 \\ 1/2 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$



$$r = \begin{bmatrix} r(u_1) \\ r(u_2) \\ r(u_3) \\ r(u_4) \\ r(u_5) \end{bmatrix}$$

$$r(u_1) = 1/3 r(u_4) + 1/2 r(u_5)$$

$$r(u_2) = 1/2 r(u_1) + r(u_3) + 1/3 r(u_4)$$

$$r(u_3) = 1/2 r(u_1) + 1/3 r(u_4)$$

$$r(u_4) = 1/2 r(u_5)$$

$$r(u_5) = r(u_2)$$

# Περίληψη: πίνακες

## Επεκτάσεις

Dead ends: η στήλη στο  $M$  που του αντιστοιχεί  $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1/n \\ 1/n \\ 1/n \end{bmatrix}$

Spider traps (loops): ακμές με πιθανότητα  $1-\beta$

$$\beta M + (1-\beta) \begin{bmatrix} 1/n & & \\ 1/n & & \\ & \dots & \\ 1/n & & \end{bmatrix}$$

# Περίληψη: random walks

Ξεκίνα από κάποιον τυχαίο κόμβο

Επέλεξε τυχαία μια από τις εξερχόμενες ακμές του  
Ακολούθησε την ακμή  
Επανάλαβε

*(με άλματα)*

Με πιθανότητα  $\beta$

Επέλεξε τυχαία μια από τις εξερχόμενες ακμές του  
Ακολούθησε την ακμή

Με πιθανότητα  $1-\beta$

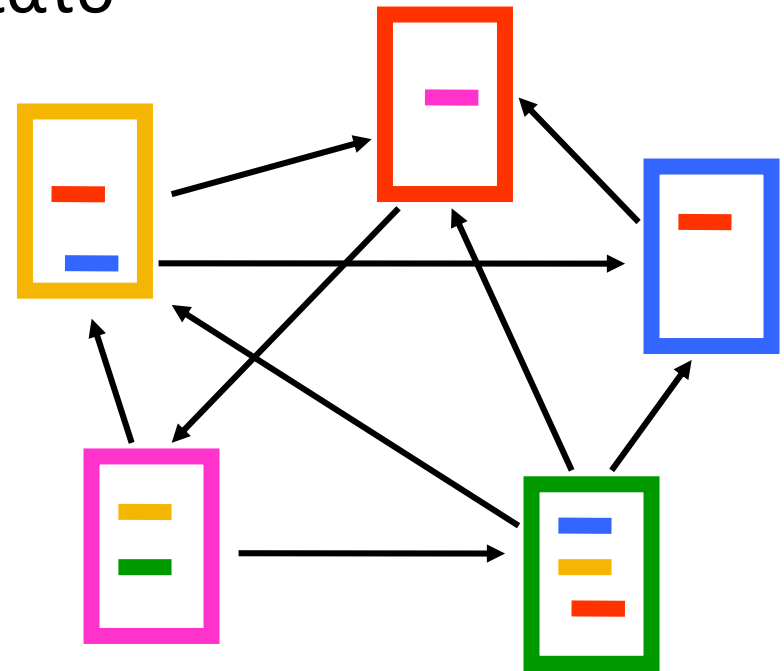
Επέλεξε ένα τυχαίο κόμβο

- Random surfer
- PageRank του κόμβου  $u$ : το ποσοστό των φορών (πιθανότητα) που επισκεπτόμαστε τον κόμβο όταν  $t \rightarrow \infty$

# PageRank: τυχαίος περίπατος (επανάληψη)

- Vanilla random walk
  - Κάνουμε το πίνακα γειτνίασης στοχαστικό και τρέχουμε ένα τυχαίο περίπατο

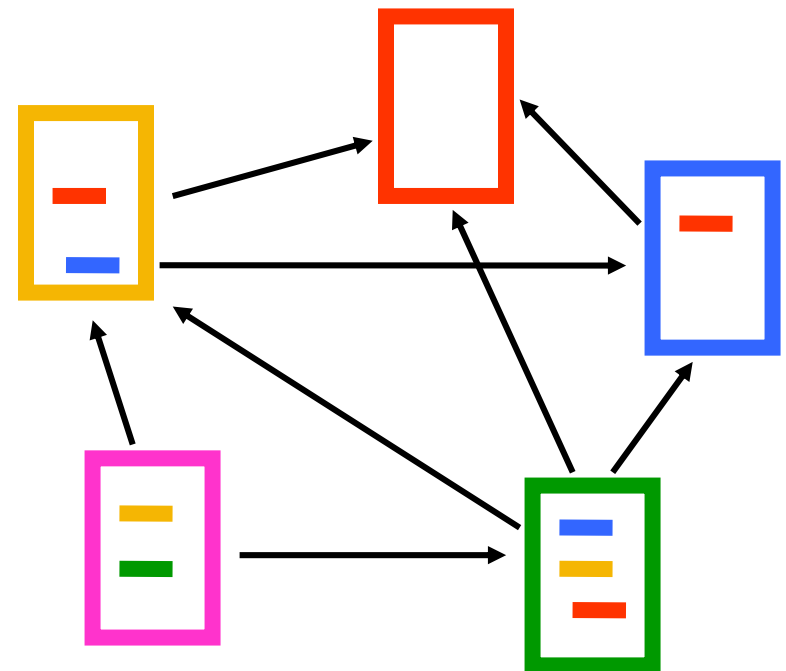
$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$



# PageRank: τυχαίος περίπατος

- Τι συμβαίνει με τα αδιέξοδα (**sink nodes**);
  - Όταν ο τυχαίος περίπατος φτάσει σε έναν κόμβο χωρίς εξερχόμενες ακμές;

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$



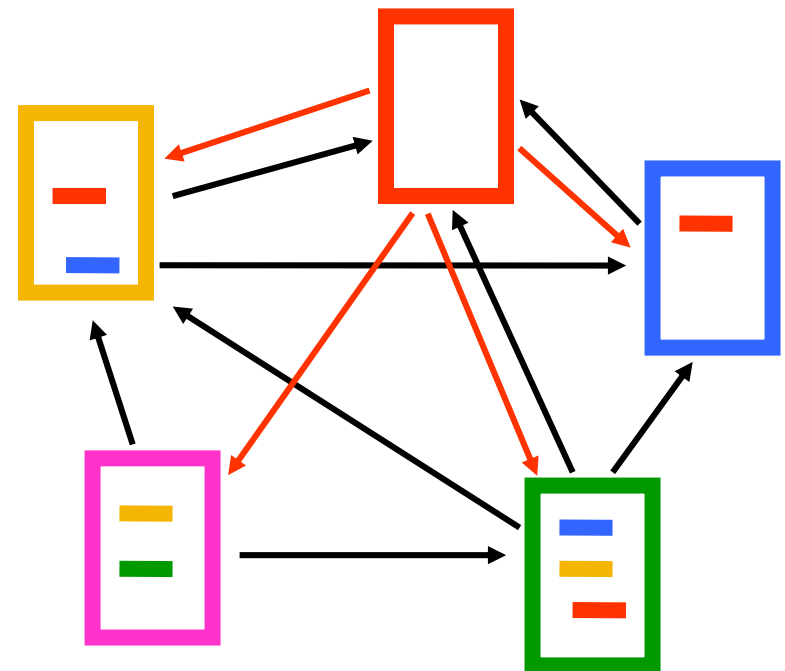
Στις διαφάνειες έχουμε τον ανάστροφο του  $M$

# PageRank: τυχαίος περίπατος

- Αντικαθιστούμε αυτά τα διανύσματα με ένα διάνυσμα  $v$ 
  - συνήθως, το ομοιόμορφο διάνυσμα

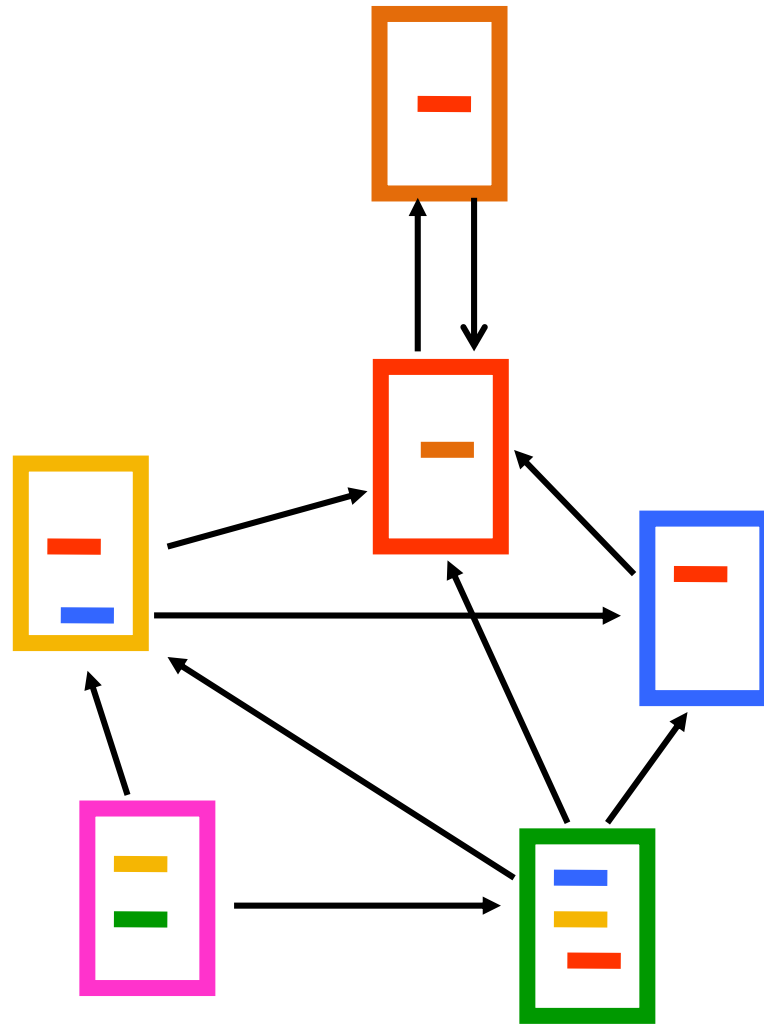
$$P' = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

$$P' = P + dv^T \quad d = \begin{cases} 1 & \text{if } i \text{ is sink} \\ 0 & \text{otherwise} \end{cases}$$



# PageRank: τυχαίος περίπατος

- Για τα loops (spider traps);



# PageRank: τυχαίος περίπατος

- Προσθέτουμε ένα τυχαίο άλμα σε ένα διάνυσμα  $\mathbf{v}$  με πιθανότητα  $1-\beta$ 
  - συνήθως, το ομοιόμορφο διάνυσμα
  - $\beta$  dumping factor
- Ο τυχαίος περίπατος ξαναρχίζει μετά από  $1/(1-\beta)$  βήματα in expectation
  - Εξασφαλίζει irreducibility

$$P'' = \beta \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 1/2 \end{bmatrix} + (1-\beta) \begin{bmatrix} 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

$$P'' = \beta P' + (1-\beta) \mathbf{u} \mathbf{v}^T, \text{ όπου } \mathbf{u} \text{ διάνυσμα με όλα } 1$$



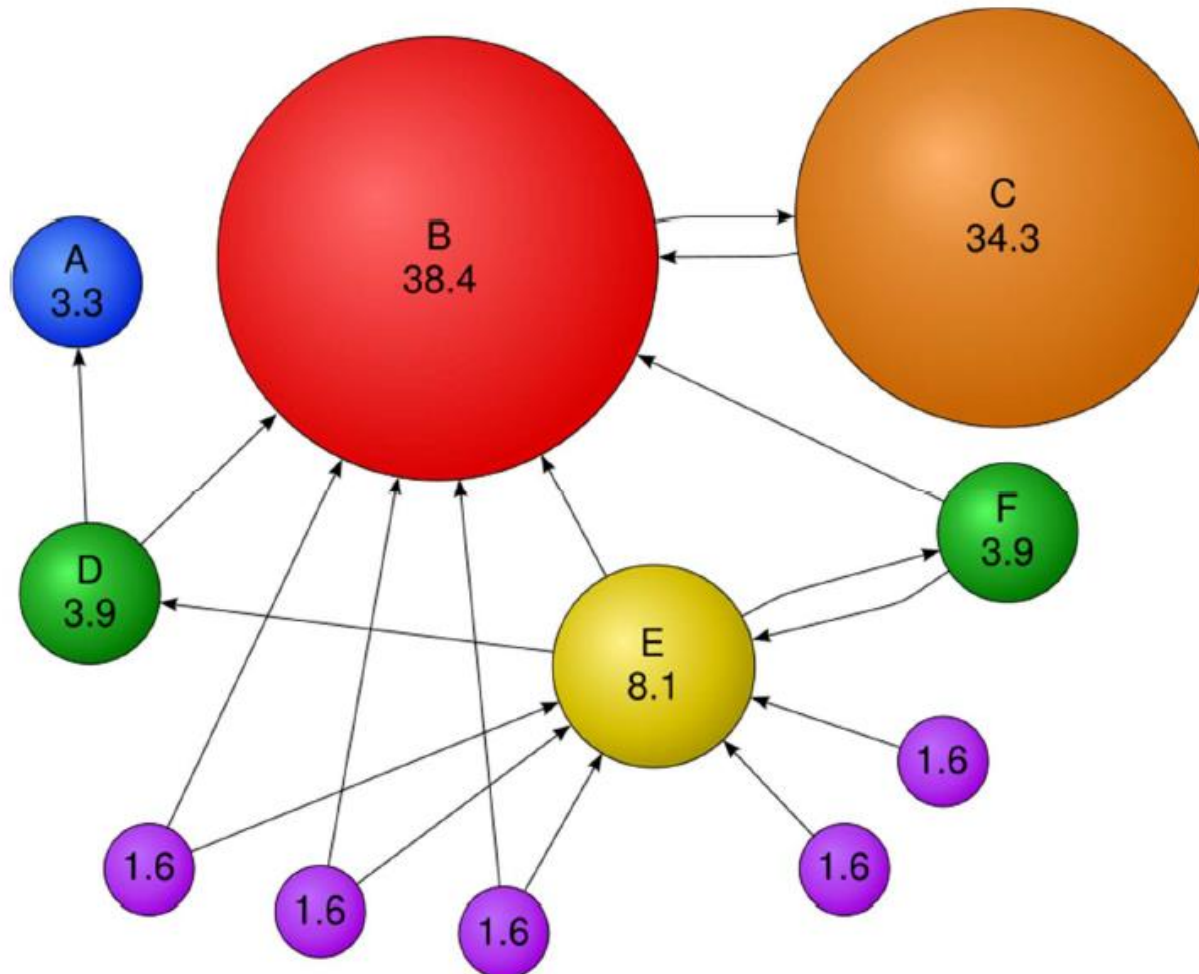
# PageRank: τυχαίος περίπατος

Ονομάζεται και τυχαίος περίπατος με επανεκκίνηση (random walk with restart)

Με τα random jumps, τα μικρότερα μονοπάτια, είναι πιο σημαντικά

Αν το διάνυσμα  $v$  που γίνεται το jump δεν είναι uniform, τότε μια προτίμηση σε συγκεκριμένους κόμβους (κόμβοι σε «μικρή» απόσταση από τους «restart» κόμβους)

# PageRank: Παράδειγμα



# Θεματικό PageRank

Η σημασία μιας σελίδας μόνο με βάση το δίκτυο – ανεξάρτητη από την ερώτηση (ή, το θέμα)

Πως μπορούμε να υπολογίσουμε «θεματικό» ή personalized PageRank;

# Θεματικό PageRank

Έστω  $S$  ένα σύνολο από σελίδες «συναφείς» με το θέμα

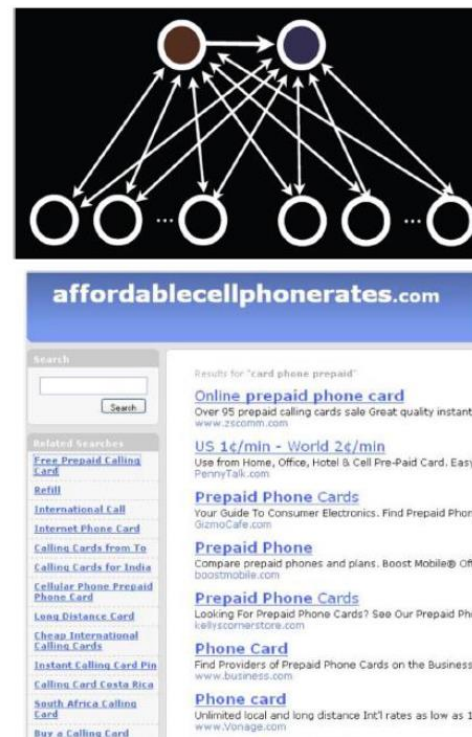
Προσθέτουμε συνδέσεις teleports σε αυτές αντί σε τυχαίους κόμβους

$$\begin{aligned}M'_{ij} &= (1 - \beta) M_{ij} + \beta / |S| && \text{if } i \in S \\ &= (1 - \beta) M_{ij} && \text{otherwise}\end{aligned}$$

Γειτνίαση με τις σελίδες στο σύνολο  $S$

# PageRank

**Link Farms:** δίκτυα από εκατομμύρια σελίδες που δείχνουν η μία στην άλλη με στόχο την αύξηση του PageRank κάποιων σελίδων



# PageRank: χρήση στην ανάκτηση

- Σελίδες με μεγάλο PageRank υψηλότερα στη διάταξη
- Τελικός βαθμός συνδυασμός πολλών χαρακτηριστικών (features)

# HITS

# Hyperlink-Induced Topic Search (HITS)

Την ίδια εποχή με το PageRank

Δύο βασικές διαφορές

1. Κάθε σελίδα έχει δύο βαθμούς:
  - ένα **βαθμό κύρους (authority rank)** και
  - ένα **κομβικό βαθμό (hub rank)**
2. Οι βαθμοί είναι **θεματικοί**



# HITS

**Authorities (σελίδες κύρους):** σελίδες που περιέχουν χρήσιμη πληροφορία (οι εξέχουσες, έγκριτες απαντήσεις στις ερωτήσεις)

Σελίδες εφημερίδων

Σελίδες μαθημάτων

Σελίδες κατασκευαστών αυτοκινήτων

**Hubs (κομβικές σελίδες):** σελίδες που δείχνουν σε αυθεντίες (λίστες μεγάλης αξίας)

Λίστες από εφημερίδες

Πρόγραμμα μαθημάτων

Λίστες από κατασκευαστές

Άλλη εφαρμογή: βιβλιογραφικές αναφορές

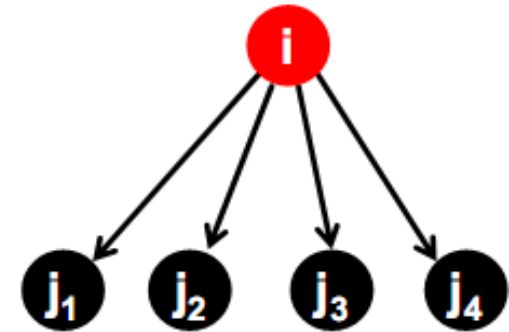
# HITS

- Ένα **hub** είναι τόσο καλό όσο καλά είναι τα authorities στα οποία δείχνει (εξερχόμενες ακμές σε πολλά καλά authorities)
- Ένα **authority** είναι τόσο καλό όσο τα hubs που δείχνουν σε αυτό (εισερχόμενες ακμές από πολλά καλά hubs)

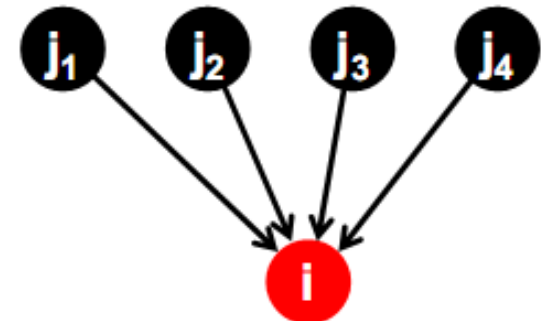
# HITS: Ορισμοί

Κάθε σελίδα  $p$ , έχει δύο σκορ

- **hub score ( $h$ )** ως ειδικός Άθροισμα των authority σκορ των σελίδων στις οποίες δείχνει
- **authority score ( $a$ )** ποιότητα περιεχομένου  
Άθροισμα των hub σκορ των σελίδων που δείχνουν σε αυτήν



$$h_i = \sum_{i \rightarrow j} a_j$$



$$a_i = \sum_{j \rightarrow i} h_j$$

# Χρήση στη αναζήτηση

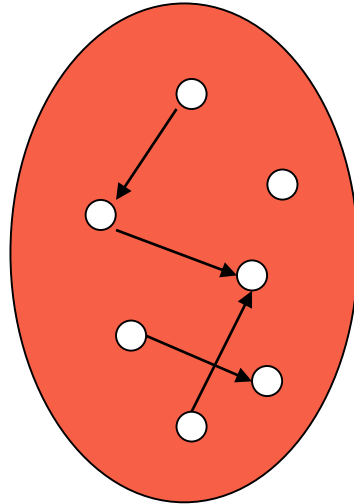
- Στην αρχική του μορφή, ο αλγόριθμος εφαρμόζεται σε *υποσύνολο του web (query dependent input)*
  1. Βρες από το web ένα **σύνολο βάσης (base set)** από σελίδες που θα μπορούσαν να είναι καλά hubs ή authorities.
  2. Χρησιμοποίησε αυτό το σύνολο για να υπολογίζεις τα scores και να βρεις ένα μικρό σύνολο από κορυφαίες hub και authority σελίδες (επαναληπτικό αλγόριθμο)

# Σύνολο βάσης

- Δοθείσας μια ερώτησης (πχ **Macron**), χρησιμοποίησε ένα ευρετήριο κειμένου και ανέκτησε όλες τις σελίδες που περιέχουν τον όρο **Macron**.
  - Ας ονομάσουμε αυτό το σύνολο, **σύνολο ρίζα (root set)**
- Πρόσθεσε οποιαδήποτε σελίδα:
  - είτε δείχνει σε μια σελίδα στο σύνολο ρίζα,
  - είτε μια σελίδα στο σύνολο ρίζα δείχνει σε αυτήνΚαι τις μεταξύ τους συνδέσεις
- Ονομάζουμε το σύνολο που προκύπτει **σύνολο βάσης**

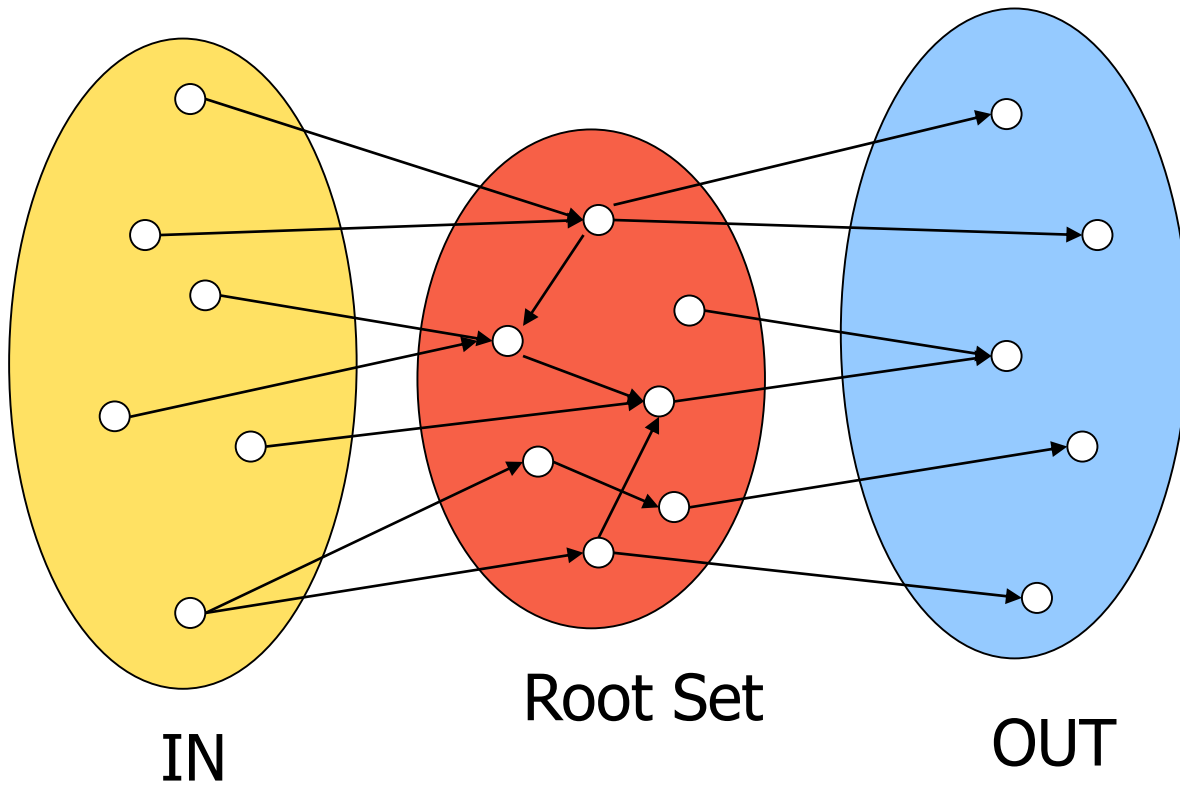
# Σύνολο βάσης

Σύνολο ρίζα που προκύπτει από μια μηχανή αναζήτησης που χρησιμοποιεί μόνο το κείμενο

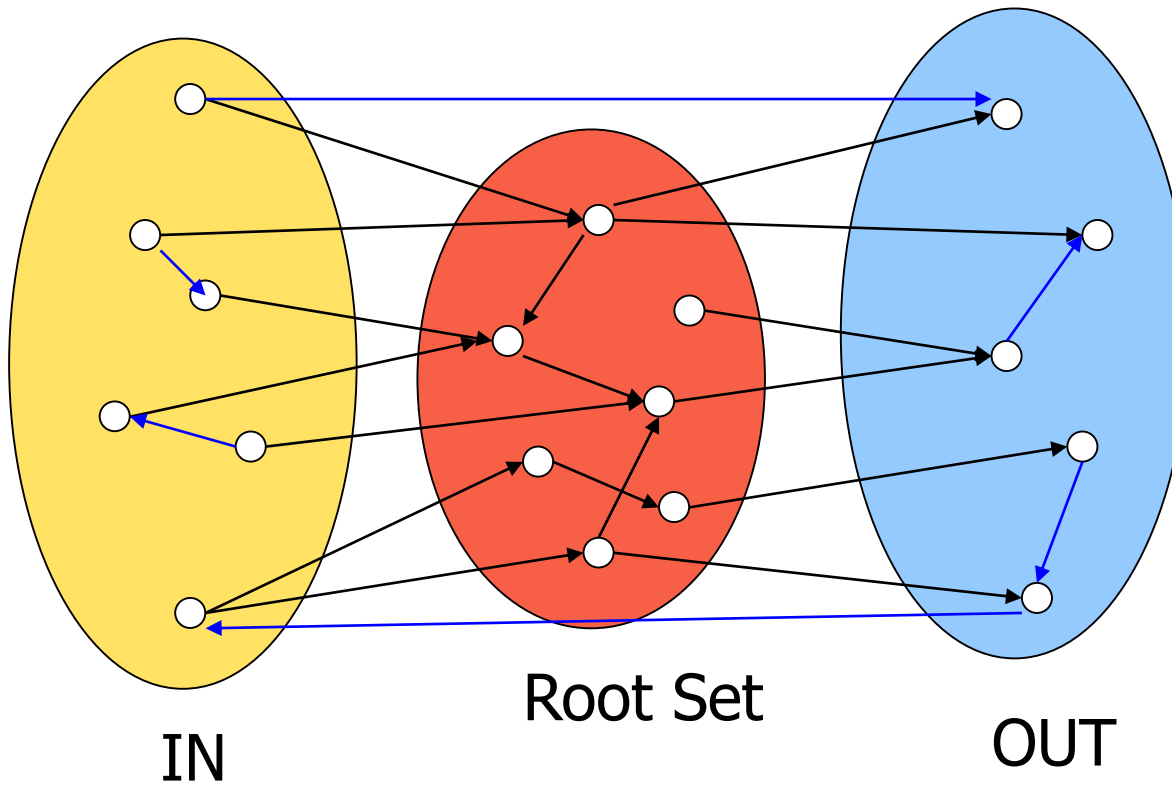


Root Set

# Σύνολο βάσης

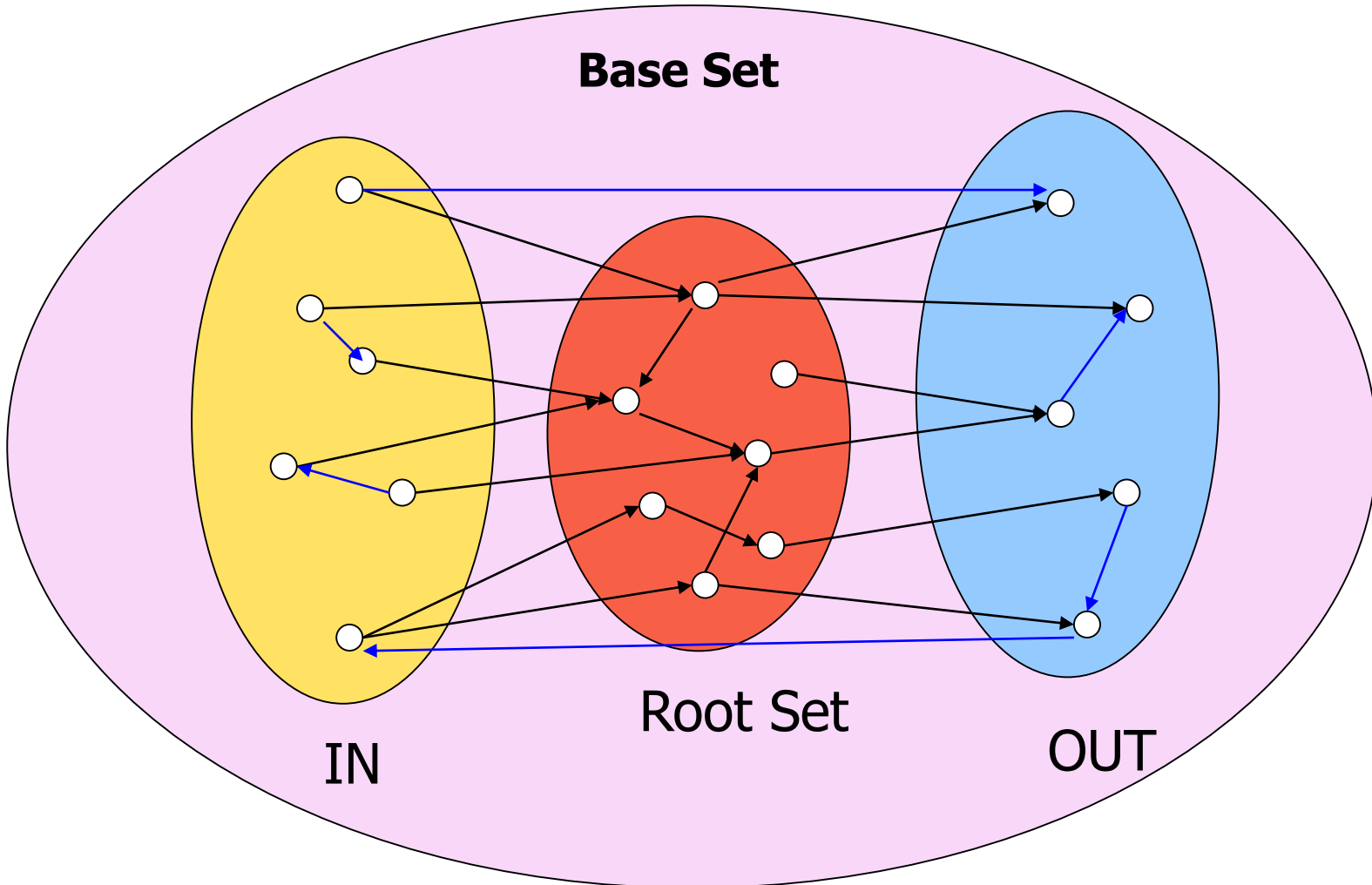


# Σύνολο βάσης





# Σύνολο βάσης



# Υπολογισμός

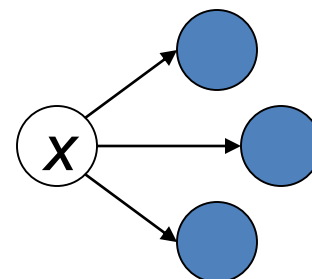
- Υπολόγισε για κάθε σελίδα  $x$  στο σύνολο βάσης ένα hub score  $h(x)$  και ένα authority score  $a(x)$ .
  - Initialize: for all  $x$ ,  $h(x) \leftarrow 1$ ;  $a(x) \leftarrow 1$ ;
  - Iteratively update all  $h(x)$ ,  $a(x)$ ;

# Επαναληπτικός υπολογισμός

- Επανάλαβε τις παρακάτω ενημερώσεις για κάθε κόμβο  $x$ :

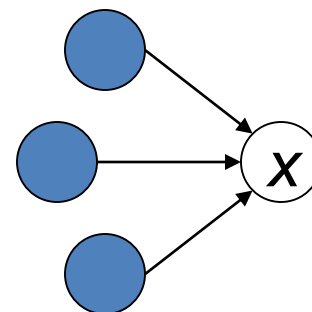
I operation

$$h(x) \leftarrow \sum_{x \mapsto y} a(y)$$



O operation

$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$

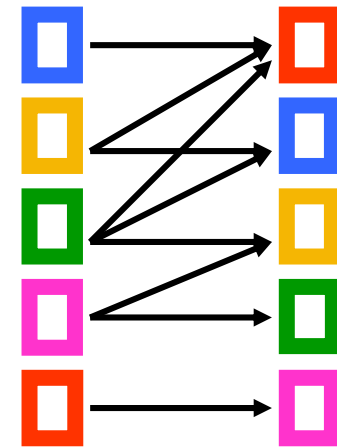
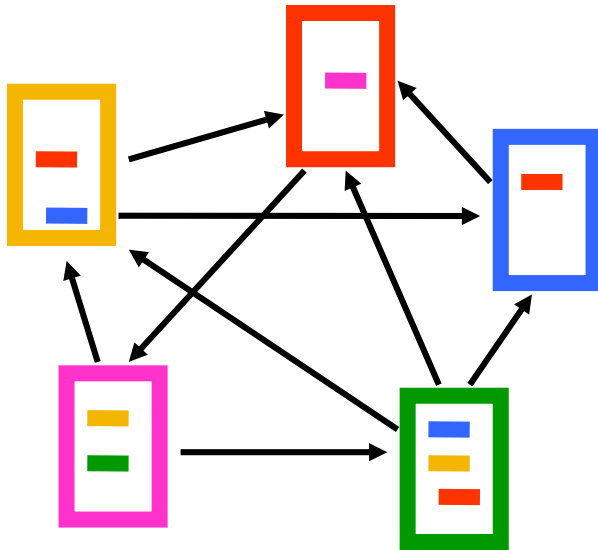


Normalize

# Κανονικοποίηση

- Για να αποφύγουμε οι τιμές  $h()$  and  $a()$  να γίνουν πολύ μεγάλες τις κλιμακώνουμε (scale down) μετά από κάθε επανάληψη
- Πως;
  - Δεν έχει σημασία γιατί αυτό που πραγματικά μας ενδιαφέρει είναι οι σχετικές τιμές τους
  - Διαίρεσε όλα τα hub scores με το άθροισμα των hub scores και όλα τα authority scores με το άθροισμα των authority scores

# Παράδειγμα

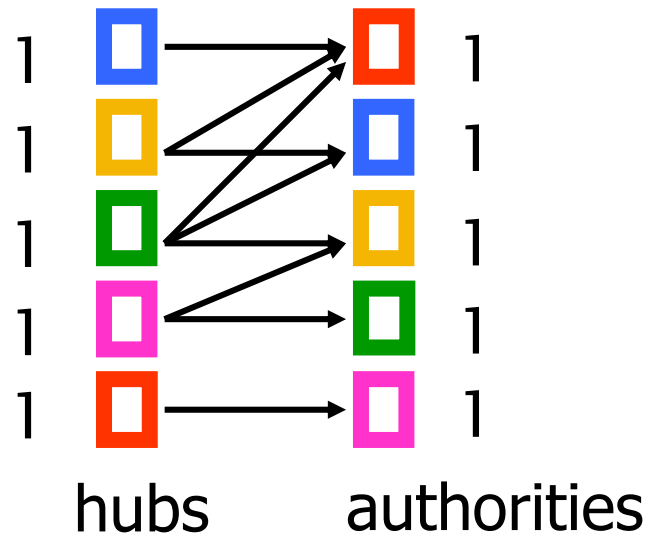


hubs

authorities

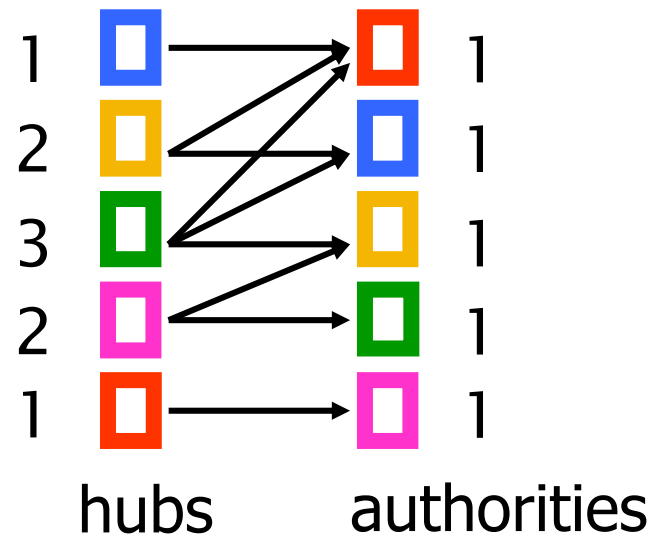
# Παράδειγμα

Initialize



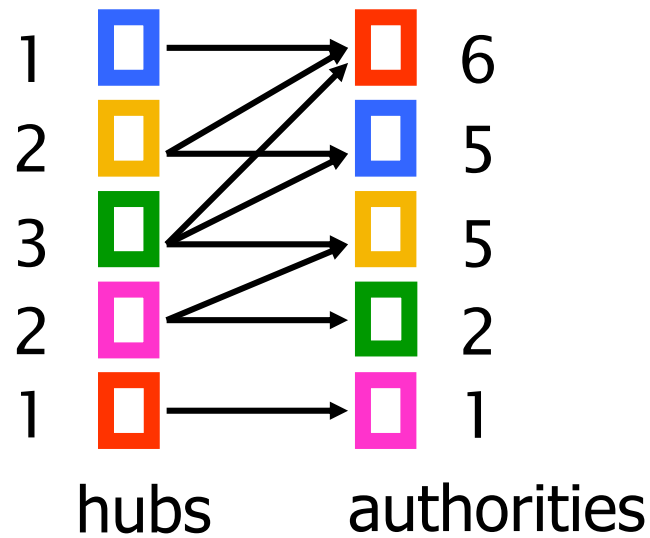
# Παράδειγμα

Step 1: O operation



# Παράδειγμα

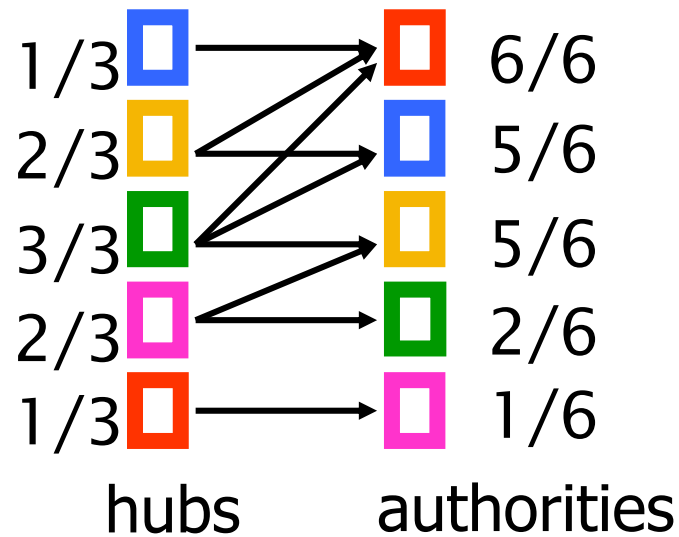
Step 1: I operation





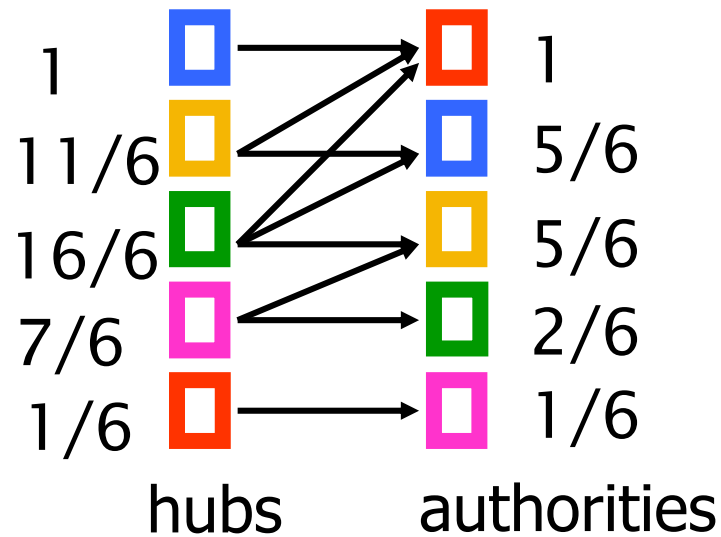
# Παράδειγμα

Step 1: Normalization (Max norm)



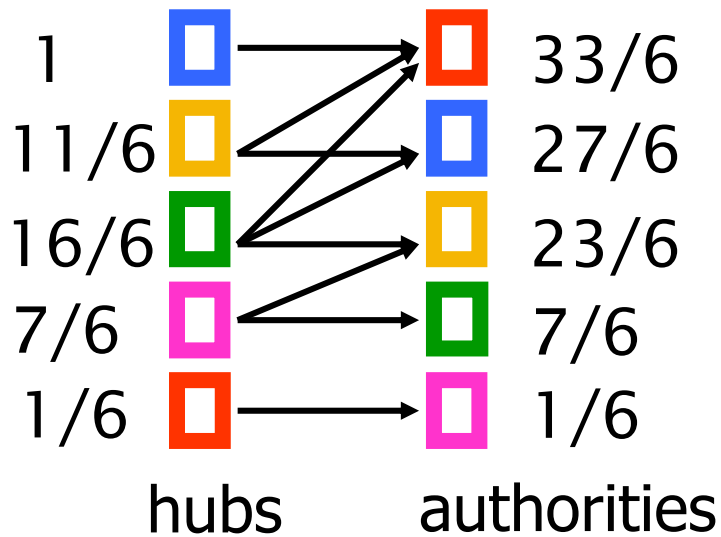
# Παράδειγμα

Step 2: 0 step



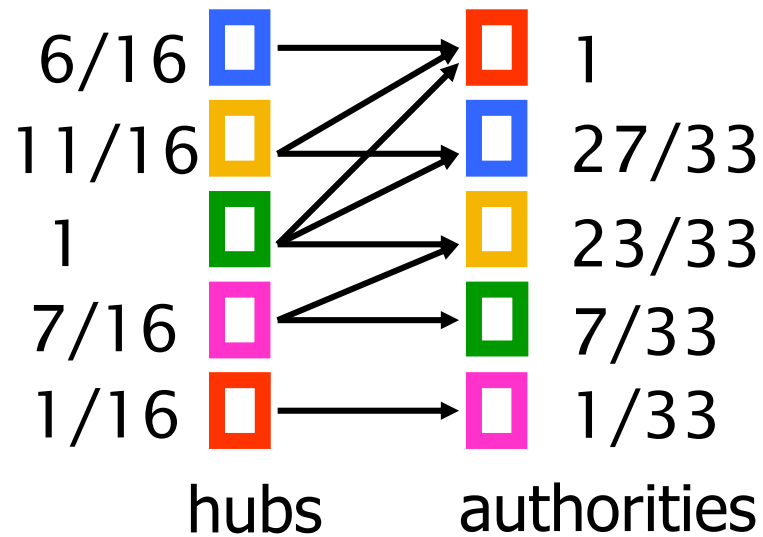
# Παράδειγμα

Step 2: 1 step



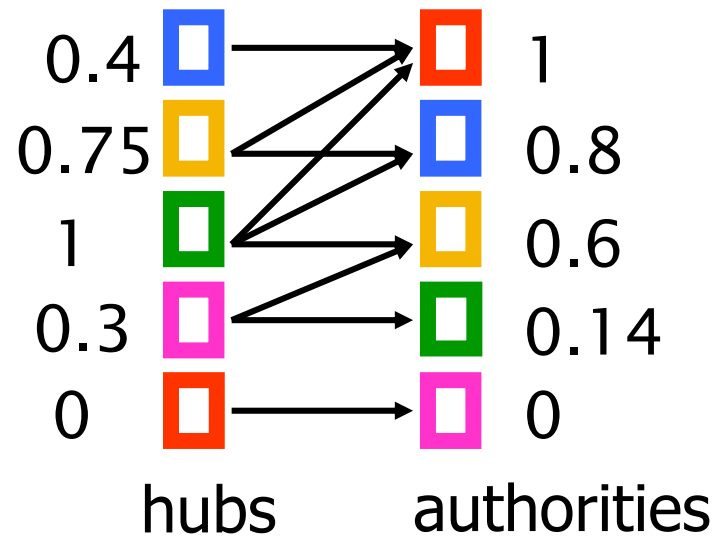
# Παράδειγμα

## Step 2: Normalization



# Παράδειγμα

Convergence



# Σύγκλιση

- Οι σχετικές τιμές συγκλίνουν μετά από λίγες επαναλήψεις
- Στην πράξη,  $\sim 5$  επαναλήψεις οι τιμές σχεδόν σταθεροποιούνται

# Japan Elementary Schools

## Hubs

- schools
- LINK Page-13
- “ú—{,ìŠwZ
- a%o,,ñŠwZfz[f fy[fW
- 100 Schools Home Pages (English)
- K-12 from Japan 10/...rnet and Education )
- http://www...iglobe.ne.jp/~IKESAN
- ,l,f,jñŠwZ,U”N,P’g•”œê
- ÒŠ—’—şÒŠ—“œñŠwZ
- Koulutus ja oppilaitokset
- TOYODA HOMEPAGE
- Education
- Cay's Homepage(Japanese)
- —y“iñŠwZ,ìfz[f fy[fW
- UNIVERSITY
- %oJ—³ñŠwZ DRAGON97-TOP
- Â%o^añŠwZ,T”N,P’g fz[f fy[fW
- ¶µ°é¼ÁÁ© ¥á¥È¥â¼¼ ¥á¥È¥â¼¼

## Authorities

- The American School in Japan
- The Link Page
- %o^es—ş^ä“cñŠwZfz[f fy[fW
- Kids' Space
- ^Àes—ş^Àé¼¼•”ñŠwZ
- <{éx³ç’âŠw•®ñŠwZ
- KEIMEI GAKUEN Home Page ( Japanese )
- Shiranuma Home Page
- fuzoku-es.fukui-u.ac.jp
- welcome to Miasa E&J school
- ¶“bìœşE%o;•ls—ş†ì¼¼ñŠwZ,ìfy
- http://www...p/~m\_maru/index.html
- fukui haruyama-es HomePage
- Torisu primary school
- goo
- Yakumo Elementary,Hokkaido,Japan
- FUZOKU Home Page
- Kamishibun Elementary School...

# Παρατηρήσεις

- Συγκεντρώθηκαν καλές σελίδες ανεξάρτητα από τη γλώσσα του περιεχομένου της σελίδας
- Η χρήση της ανάλυσης συνδέσμων γίνεται μετά τη δημιουργία του συνόλου βάσης
  - ο υπολογισμός των score εξαρτάται από το ερώτημα και γίνεται μετά την ανάκτηση με βάση το περιεχόμενο (σημαντική χρονική επιβάρυνση)

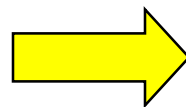
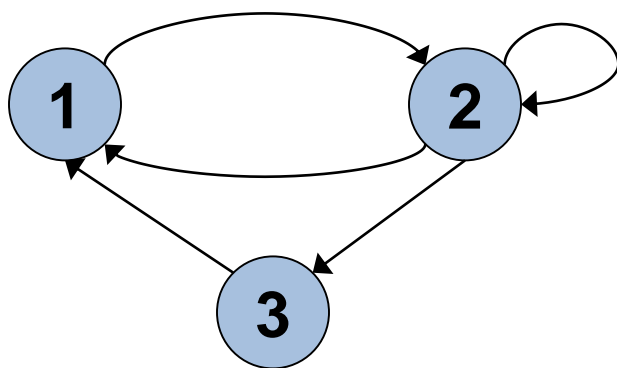


# Θέματα

- Topic Drift
  - Σελίδες εκτός θέματος (off-topic) μπορεί να οδηγήσουν στο να επιστραφούν εκτός θέματος authorities
    - Π.χ., οι γειτονικοί κόμβοι μπορεί να αναφέρονται σε κάποιο “super topic”
- Mutually Reinforcing Affiliates
  - Συνεργαζόμενες σελίδες μπορεί να αυξήσουν τα σκορ τους

# Διανυσματική αναπαράσταση

- $n \times n$  πίνακας γειτνίασης **A**:
  - Για το σύνολο βάση
  - $A_{ij} = 1$  αν η σελίδα page  $i$  δείχνει στο  $j$ , αλλιώς  $= 0$ .



	1	2	3
1	0	1	0
2	1	1	1
3	1	0	0

# Διανύσματα Hub/Authority

- Τα hub σκορ  $h()$  και authority σκορ  $a()$  ως  $n$ -διάστατα διανύσματα
- Οι επαναληπτικοί υπολογισμοί:

$$h(x) \leftarrow \sum_{x \mapsto y} a(y) \quad h_i = \sum_j A_{ij} \cdot a_j$$

$$a(x) \leftarrow \sum_{y \mapsto x} h(y)$$

$$h = A a$$

$$a = A^T h, \quad A^T \text{ transpose (ανάστροφος)}$$

# Διανυσματική αναπαράσταση

$$h = A a.$$

$$a = A^T h.$$

Αντικατάσταση:

$$h = A A^T h$$

$$a = A^T A a.$$

- $h$  ιδιοδιάνυσμα του  $AA^T$
- $a$  ιδιοδιάνυσμα του  $A^T A$
- Ο αλγόριθμος ανήκει στις *power iteration* μεθόδους υπολογισμού ιδιοδιανυσμάτων
- Συγκλίνει

# PageRank vs HITS

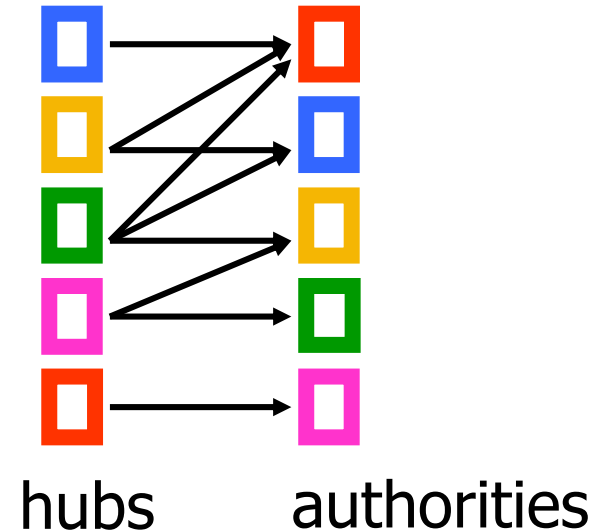
- Θα μπορούσαμε να εφαρμόσουμε το HITS σε όλο το web και το PageRank σε θεματικό υποσύνολο
- Στο web,
  - Ένα καλό hub είναι συνήθως και ένα καλό authority
  - Οι διαφορές στο rank με PageRank και HITS μικρές

# Τι άλλο θα δούμε

- Ανάλυση Συνδέσμων  
SALSA (συνδυασμός Random Walk και HITS)  
χρήση για συστάσεις φίλων στο Twitter

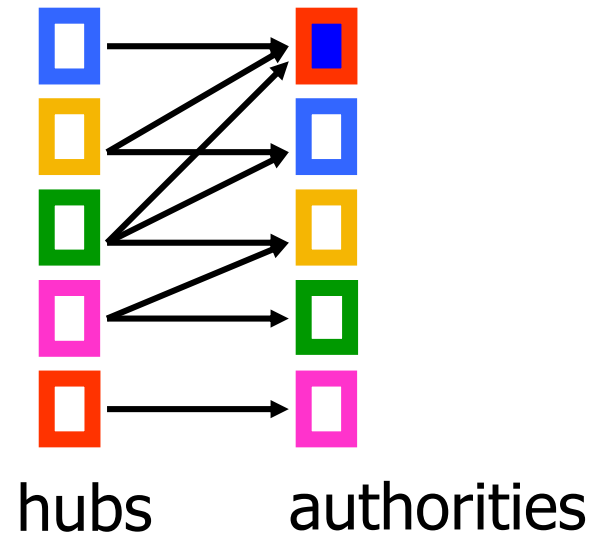
# Stochastic Approach for Link-Structure Analysis (SALSA)

- Τυχαίους περίπατους στο διμερή γράφο των hubs και των authorities εναλλάσσοντας ανάμεσα τους



# Ο αλγόριθμος SALSA

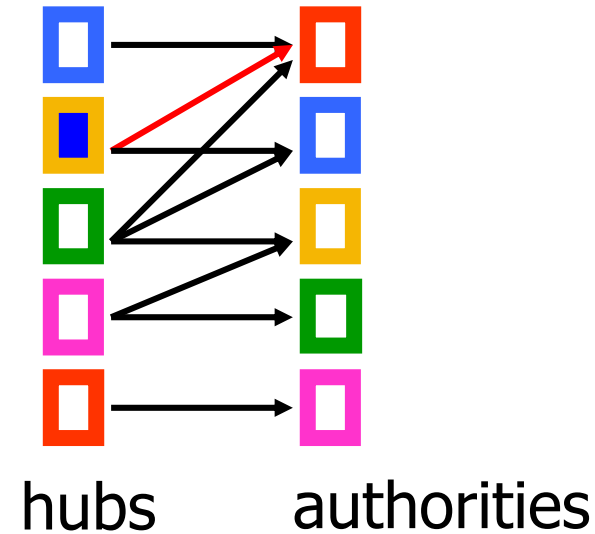
- Ξεκινάμε από ένα authority που επιλέγουμε τυχαία (uniformly at random)
  - Π.χ. το κόκκινο authority





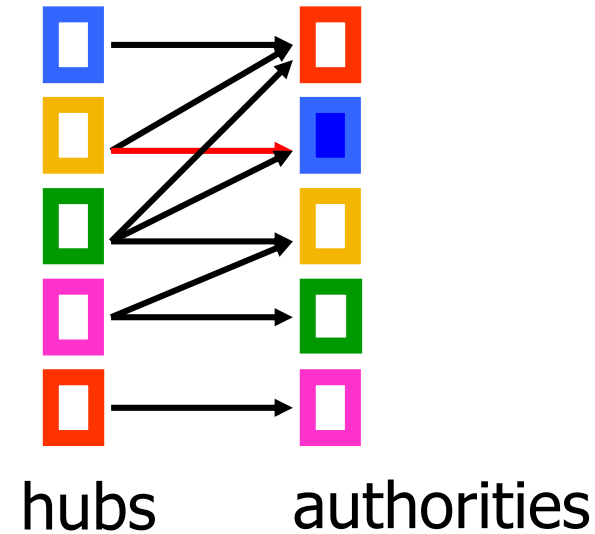
# Ο αλγόριθμος SALSA

- Επιλέγουμε τυχαία (uniformly at random) μια από τις εισερχόμενες ακμές του authority κόμβου και μετακινούμαστε σε ένα hub
  - π.χ. μετακινούμαστε στο κίτρινο hub με πιθανότητα  $1/3$



# Ο αλγόριθμος SALSA

- Επιλέγουμε τυχαία (uniformly at random) μια από τις εξερχόμενες ακμές του hub κόμβου και μετακινούμαστε σε ένα authority
  - π.χ. μετακινούμαστε στο μπλε authority με πιθανότητα  $1/2$



# Ο αλγόριθμος SALSA

- Τυπικά, έχουμε τις πιθανότητες:
  - $a_i$ : Πιθανότητα να βρίσκεται στο authority  $i$
  - $h_j$ : Πιθανότητα να βρίσκεται στο hub  $j$

# Ο αλγόριθμος SALSA

- Η πιθανότητα να βρίσκεται στο authority  $i$  υπολογίζεται ως:

$$a_i = \sum_{j \in N_{in}(i)} \frac{1}{d_{out}(j)} h_j$$

- Η πιθανότητα να βρίσκεται στο hub  $j$  υπολογίζεται ως:

$$h_j = \sum_{i \in N_{out}(j)} \frac{1}{d_{in}(i)} a_i$$

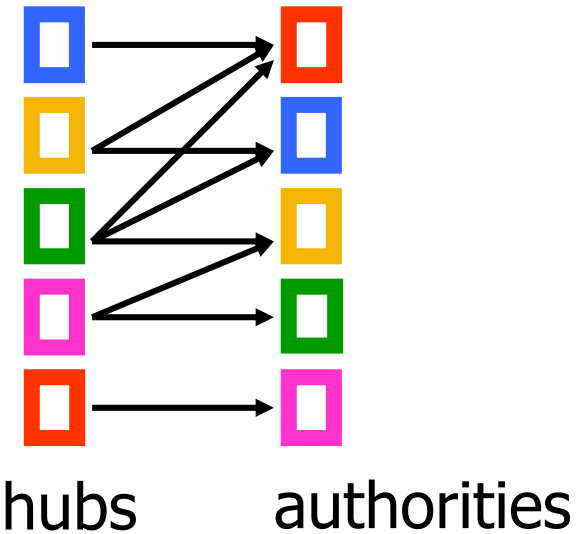
- Επαναλαμβανόμενοι υπολογισμοί συγκλίνουν

# Σε μορφή πινάκων

$A_c$  = πίνακας  $A$  όπου οι **στήλες** αθροίζουν σε 1

$A_r$  = πίνακας  $A$  όπου οι **γραμμές** αθροίζουν σε 1

- Ο υπολογισμός για τα hubs
  - $h = A_c a$
- Ο υπολογισμός για τα authorities
  - $a = A_r^T h = A_r^T A_c a$
- Ο πίνακας μετάβασης
  - $P = A_r A_c^T$



$$h_2 = 1/3 a_1 + 1/2 a_2$$

$$a_1 = h_1 + 1/2 h_2 + 1/3 h_3$$

Διαφορά από HITS θεωρούμε τα out-degrees και in-degrees

# SALSA (recommendations)

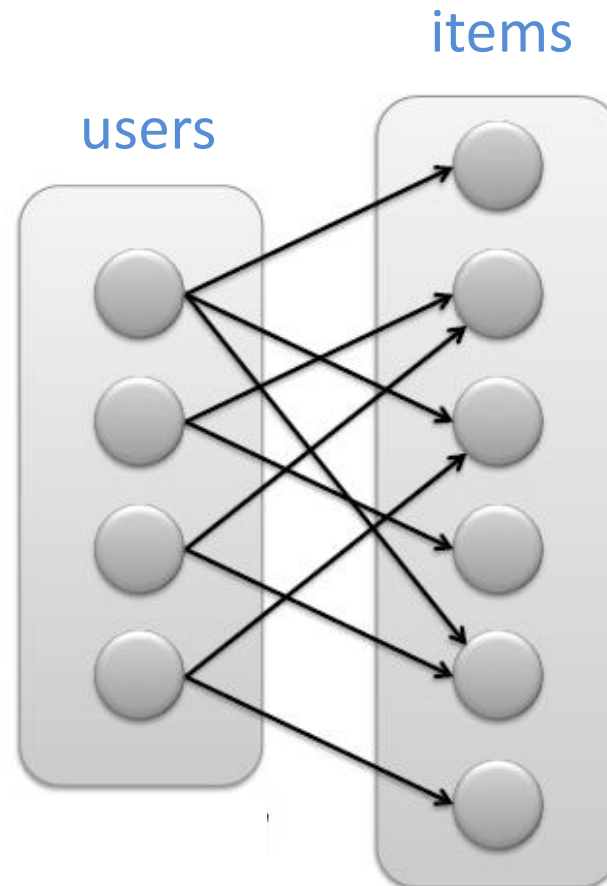
**Πρόβλημα:** θέλουμε να προτείνουμε αντικείμενα σε χρήστες

**Βασική ιδέα:** προτείνουμε αντικείμενα που «αρέσουν» στους φίλους των χρηστών

**Διμερή γράφο**

Users: φίλοι του χρήστη

Items: αντικείμενα που τους αρέσουν



# SALSA στο Twitter (as of 2013)

“circle of trust”: για κάθε χρήστη χρησιμοποιώντας μια μορφή personalized random walk

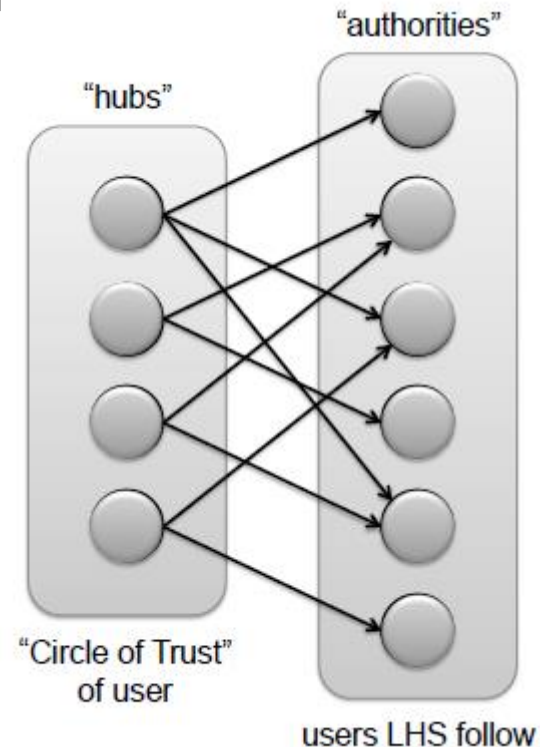
Παραλλαγή: κάθε βήμα περιλαμβάνει διάσχιση 2 ακμών (αρχή και τέλος στην ίδια πλευρά)

Authority score: συστάσεις (ποιους χρήστες να ακολουθήσει)

Hub score: ομοιότητα με το χρήστη (homophily)

**Hubs**: 500 κορυφαίοι χρήστες στο κύκλο εμπιστοσύνης του χρήστη

**Authorities**: χρήστες που ακολουθούν τα hubs



# SALSA στο Twitter (as of 2013)

“circle of trust”: κύκλο εμπιστοσύνης για κάθε χρήστη χρησιμοποιώντας μια μορφή personalized random walk

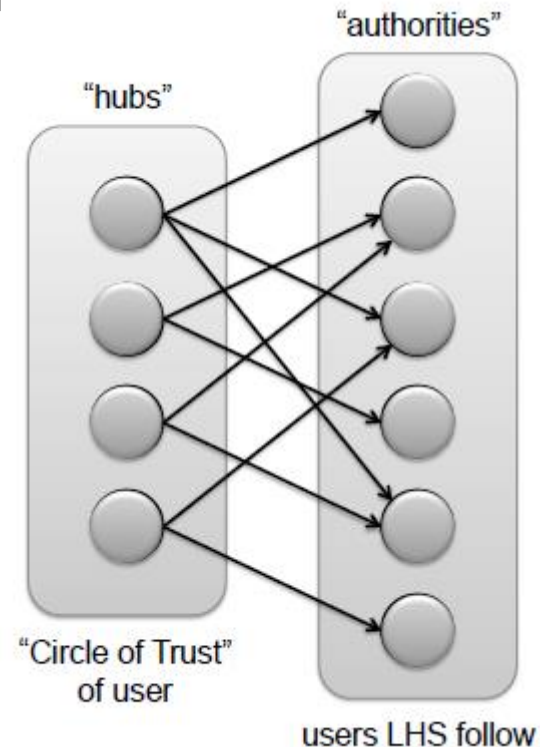
Παραλλαγή: κάθε βήμα περιλαμβάνει διάσχιση 2 ακμών (αρχή και τέλος στην ίδια πλευρά)

Authority score: συστάσεις (ποιους χρήστες να ακολουθήσει)

Hub score: ομοιότητα με το χρήστη (homophily)

**Hubs:** 500 κορυφαίοι χρήστες στο κύκλο εμπιστοσύνης του χρήστη

**Authorities:** χρήστες που ακολουθούν τα hubs





# SALSA στο Twitter (as of 2013)

- Ένας χρήστης  $u$  είναι πιθανόν να ακολουθήσει χρήστες που ακολουθούν χρήστες που είναι όμοιοι με τον  $u$ .
- Οι χρήστες είναι όμοιοι με τον  $u$  αν ακολουθούν τους ίδιους ή παρόμοιους χρήστες με αυτούς που ακολουθεί ο  $u$ .

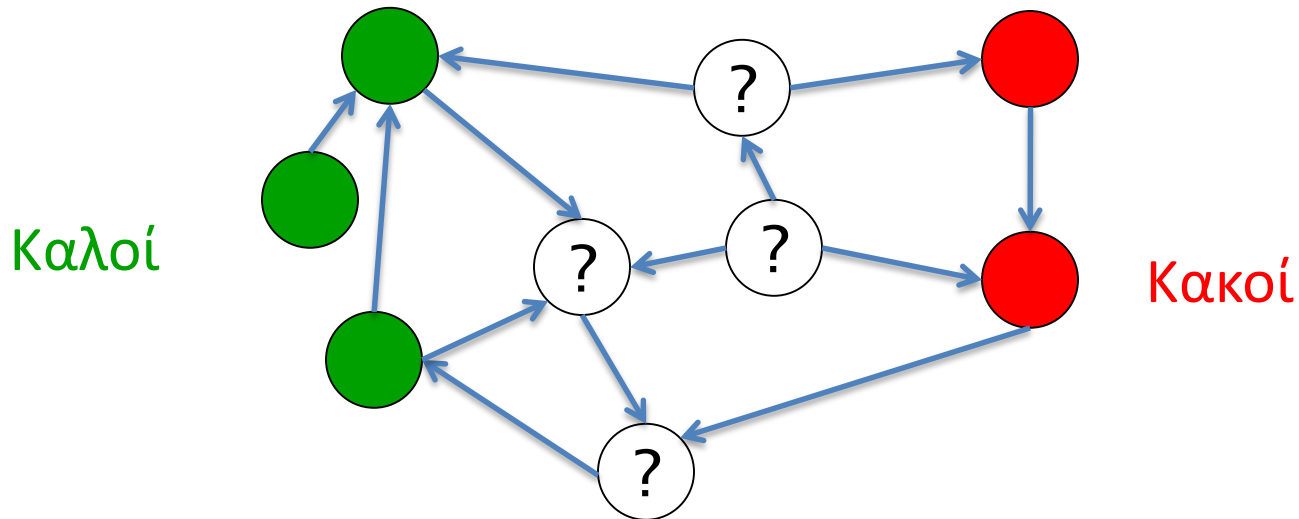
SALSA

Αριστερά: χρήστες όμοιους με τον  $u$

Δεξιά: όμοιους ακόλουθούς τους

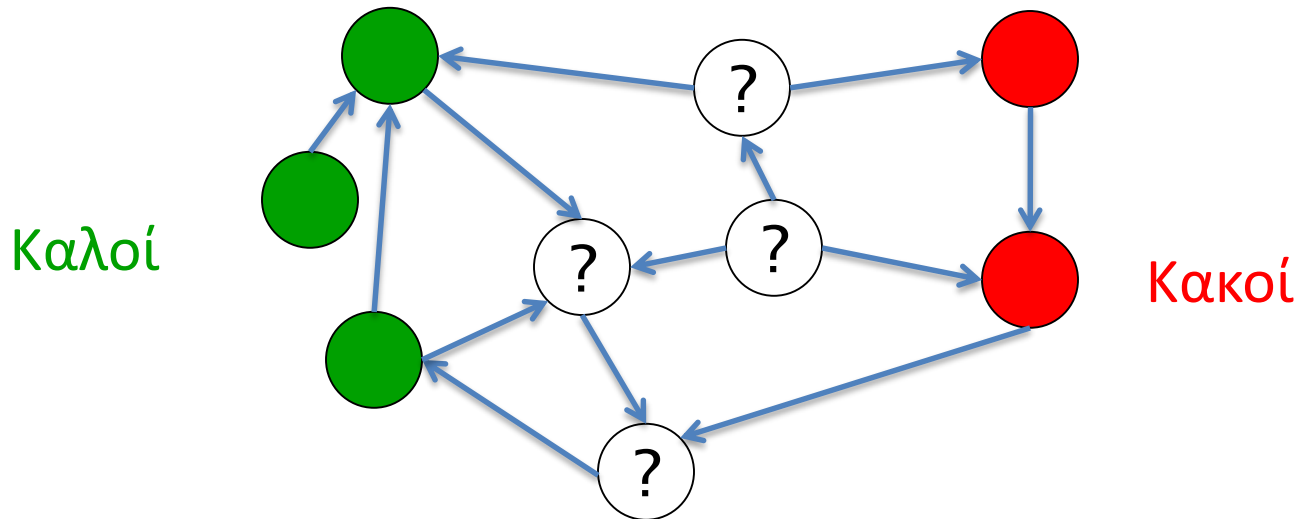
# Οι σύνδεσμοι είναι παντού!

- Πηγή πληροφορίας για το κύρος και γνησιότητα
  - Παράδειγμα - Mail spam: ποιοι email λογαριασμοί είναι spammer
- Οι Καλοί, Οι Κακοί and Οι Άγνωστοι



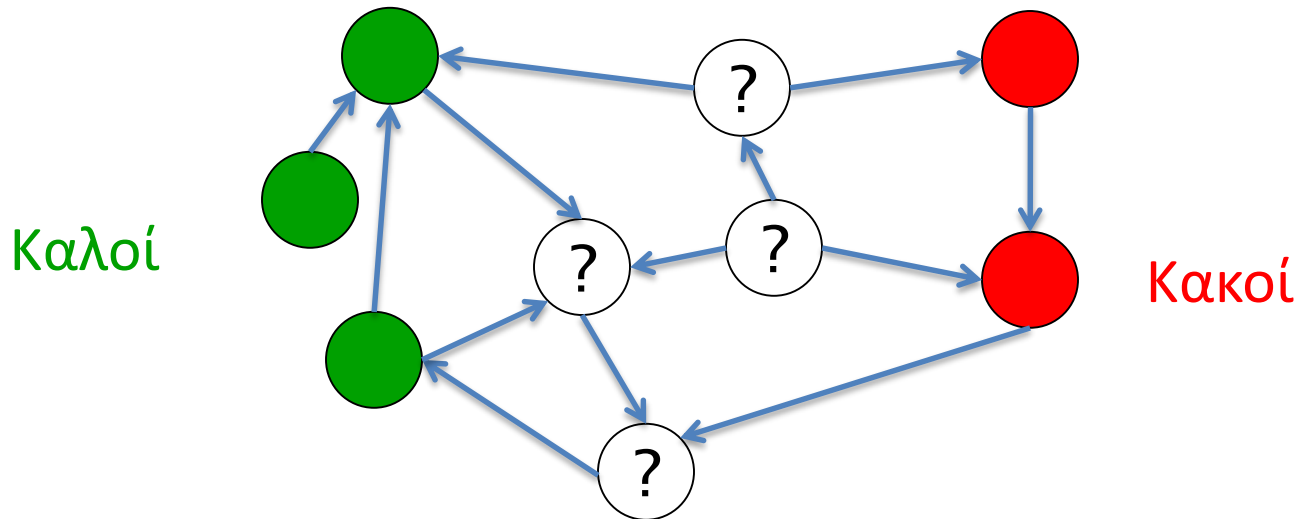
# Απλή επαναληπτική λογική

- Οι **Καλοί**, Οι **Κακοί** and Οι Άγνωστοι
  - Οι **Καλοί** κόμβοι δεν δείχνουν σε **Κακούς** κόμβους
  - Όλοι οι άλλοι συνδυασμοί επιτρέπονται



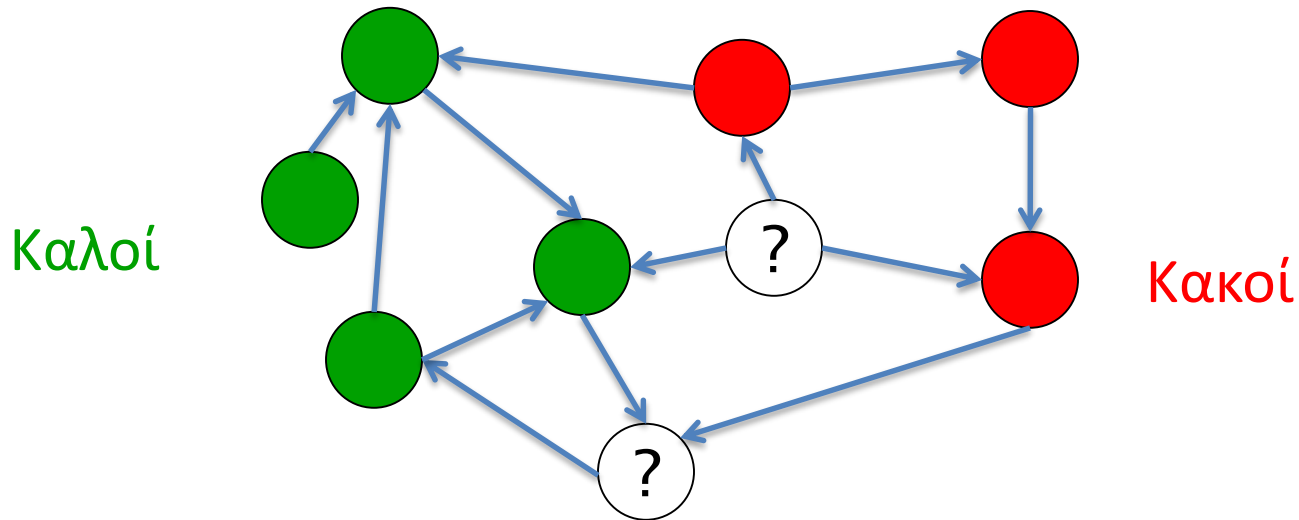
# Απλή επαναληπτική λογική

- Οι **Καλοί** κόμβοι δεν δείχνουν σε **Κακούς** κόμβους
  - Αν δείχνεις σε **Κακό** κόμβο, είσαι **Κακός**
  - Αν ένας **Καλός** κόμβος δείχνει σε σένα, είσαι **Καλός**



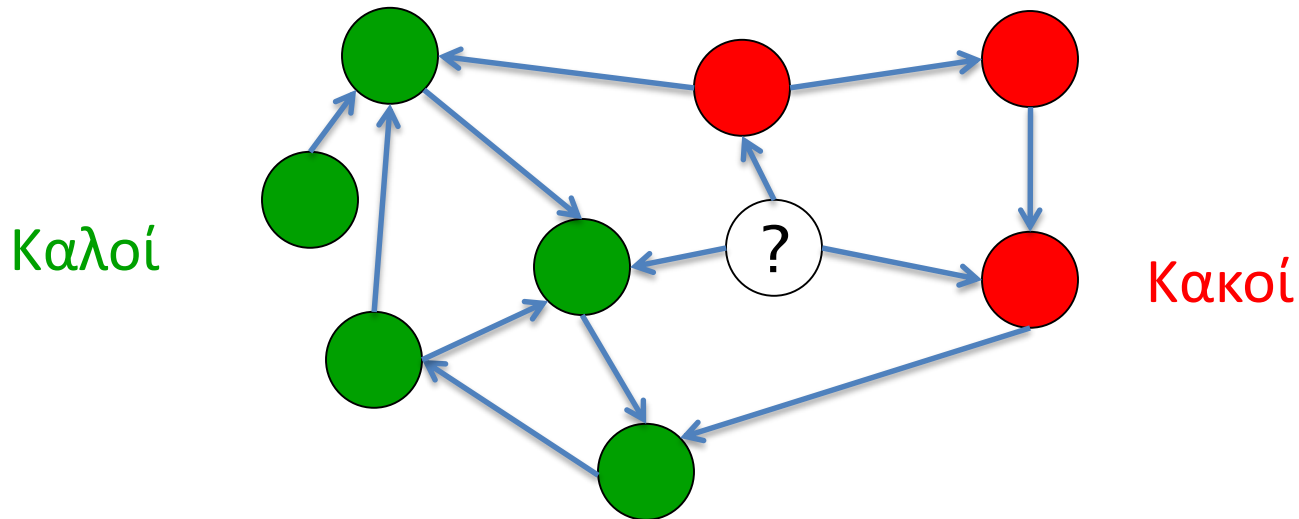
# Απλή επαναληπτική λογική

- Οι **Καλοί** κόμβοι δεν δείχνουν σε **Κακούς** κόμβους
  - Αν δείχνεις σε **Κακό** κόμβο, είσαι **Κακός**
  - Αν ένας **Καλός** κόμβος δείχνει σε σένα, είσαι **Καλός**



# Απλή επαναληπτική λογική

- Οι **Καλοί** κόμβοι δεν δείχνουν σε **Κακούς** κόμβους
  - Αν δείχνεις σε **Κακό** κόμβο, είσαι **Κακός**
  - Αν ένας **Καλός** κόμβος δείχνει σε σένα, είσαι **Καλός**



# Πολλές εφαρμογές της ανάλυσης συνδέσμων

- Homophily

Bibliometrics

e.g., citation analysis

# ΤΕΛΟΣ 21<sup>ου</sup> Κεφαλαίου

## Ερωτήσεις?

*Χρησιμοποιήθηκε κάποιο υλικό από:*

- ✓ *Pandu Nayak and Prabhakar Raghavan, CS276:Information Retrieval and Web Search (Stanford)*
- ✓ *Hinrich Schütze and Christina Lioma, Stuttgart IIR class*
- ✓ *Τις αντίστοιχες διαλέξεις του μεταπτυχιακού μαθήματος «Κοινωνικά Δίκτυα και Μέσα»*