Ανάλυση Κοινωνικών Δικτύων Και Εφαρμογές

Τυχαίοι Περίπατοι σε Γράφους

Συμεών Παπαβασιλείου (papavass@mail.ntua.gr)
Βασίλειος Καρυώτης (vassilis@netmode.ntua.gr)

09 Ιουνίου 2022

Περιεχόμενα

- Ορισμοί τυχαίων περιπάτων σε γράφους
- Γιατί τυχαίους περιπάτους σε γράφους?
- Ιδιότητες
- Εφαρμογές

Ορισμοί

- $n \times n$ Adjacency matrix A
  - $A(i,j)$ = βάρος ακμής από τον κόμβο i στον j
  - Αν ο γράφος μη-κατευθυνόμενος $A(i,j) = A(j,i)$, ο A συμμετρικός
- $n \times n$ Transition matrix P
  - P είναι row stochastic
  - $P(i,j)$ = πιθανότητα να πάμε στον j από τον i $= \frac{A(i,j)}{\sum_i A(i,j)}$
- $n \times n$ Laplacian Matrix L
  - $L(i,j) = \sum_i A(i,j) - A(i,j)$
  - Symmetric positive semi-definite για μη-κατευθυνόμενους
  - Singular

Παραδείγματα Ορισμών

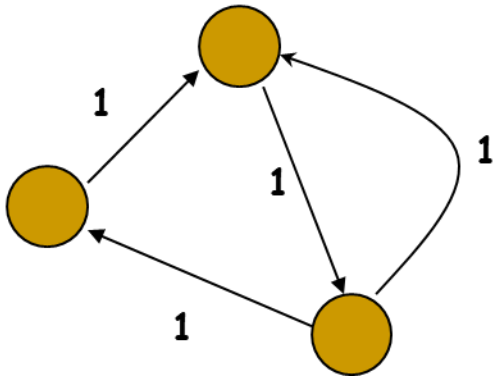
- Υπολογισμός πίνακα γειτονίας και πίνακα μεταβάσεων

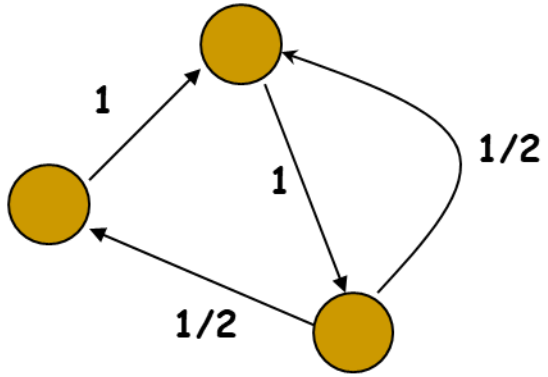
0	1	0
0	0	1
1	1	0

Adjacency matrix A

0	1	0
0	0	1
1/2	1/2	0

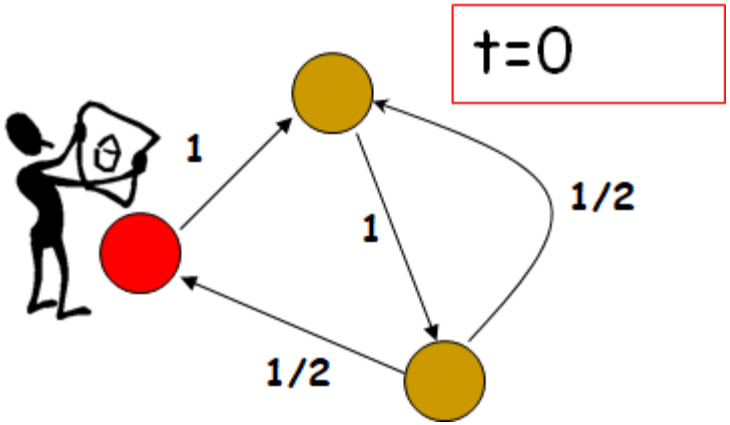
Transition matrix P

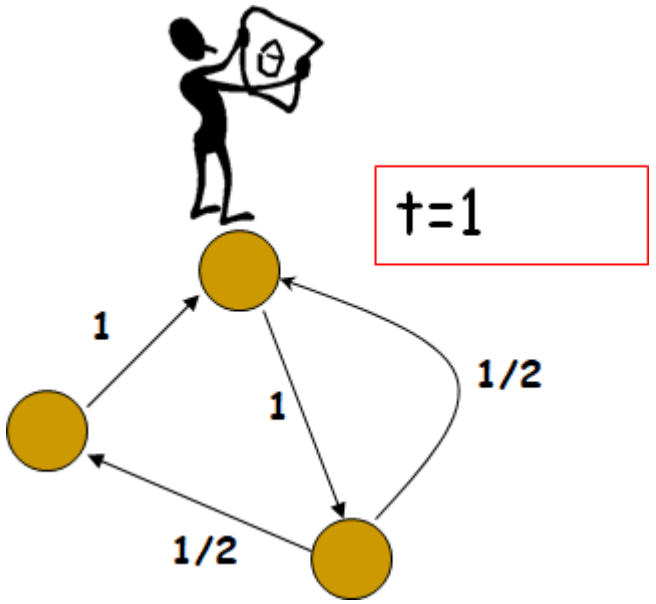


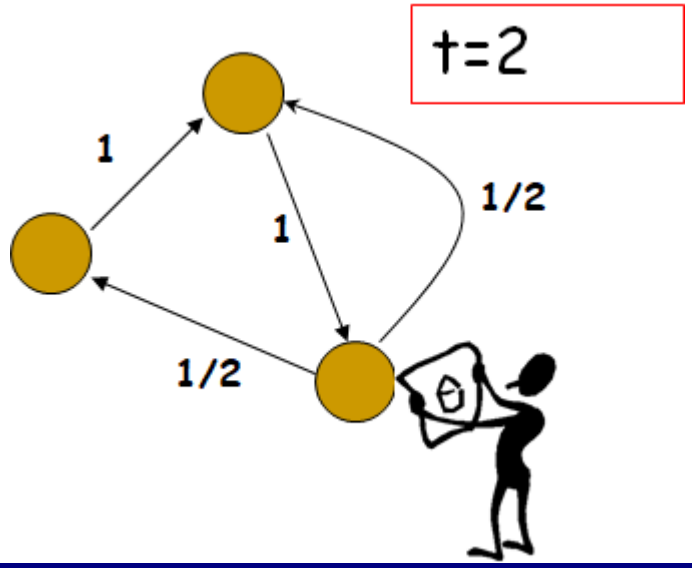


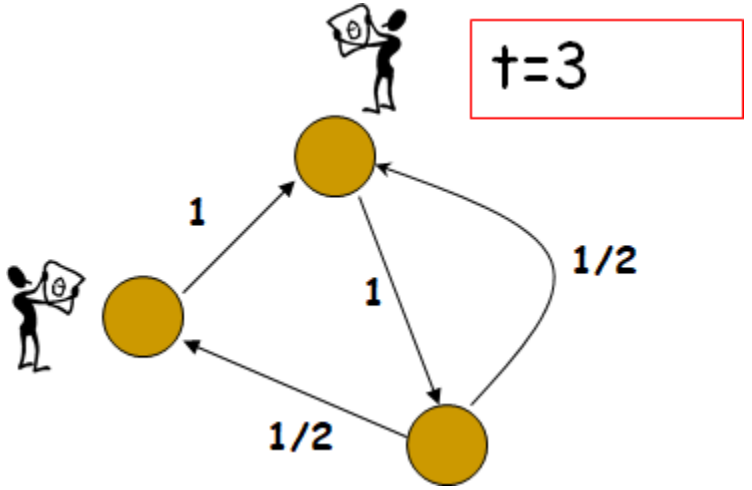
Τυχαίος Περίπατος – Random Walk

- Διαισθητικός ορισμός









Κατανομές Πιθανοτήτων

- $x_t(i)$ = πιθανότητα να είμαστε στον κόμβο i σε χρόνο t

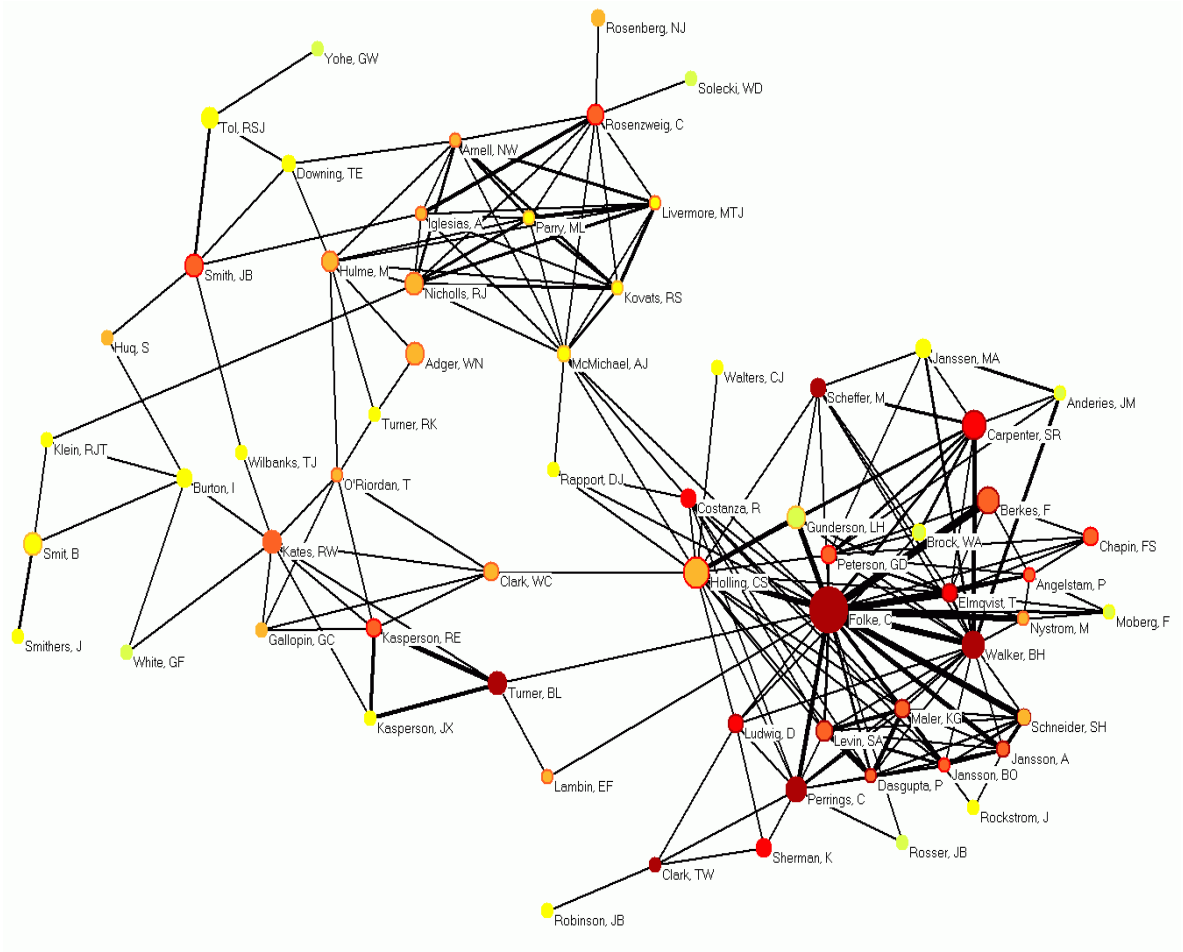
- $x_{t+1}(i) = \sum_i \text{Prob}(\text{Να είμαστε στον κόμβο j}) \times \text{Prob}(j \rightarrow$

Στάσιμη Κατανομή (Stationary Distribution)

- Όταν περιπλανιόμαστε για πολύ χρόνο
- Η κατανομή δεν αλλάζει πλέον
  - $x_{T+1} = x_T$
- Για καλά συμπεριφερόμενους γράφους, αυτό δεν εξαρτάται από την αρχική κατανομή
- **Διαισθητικά**: στάσιμη κατανομή κόμβου σχετίζεται με τον χρόνο που περνά ένα περιπλανώμενος επισκεπτόμενος τον κόμβο
  - $x_{t+1} = x_t P$
- **Μαθηματικά**: Για την στάσιμη κατανομή έχουμε $v_0 = v_0 P$
  - Αριστερό ιδιοδιάνυσμα του πίνακα μεταβάσεων!

Γιατί Τυχαίους Περιπάτους?

- Πρόβλεψη ακμών σε κοινωνικά δίκτυα



Γιατί Τυχαίους Περιπάτους?

- Βάση για συστάσεις

purnamrita's Amazon.com™ › Recommended for you
(If you're not purnamrita, click here.)

Recommendations Based on Activity

View & edit Your Browsing History

Recommendations by Category

Your Favorites Edit

Books

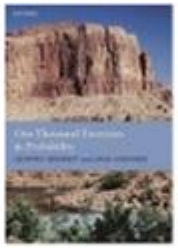
More Categories

Apparel & Accessories
Baby
Beauty
Camera & Photo
Computer & Video Games
Computers & PC Hardware
DVD
Electronics
Gourmet Food
Health & Personal Care
Industrial & Scientific

These recommendations are based on items you own and more.

view: All | New Releases | Coming Soon

More results

1. 

One Thousand Exercises in Probability
by Geoffrey R. Grimmett, David R. Stirzaker
Average Customer Review:
In Stock
Publication Date: August 2, 2001

Our Price: $53.95
Used & new from $42.74

Add to cart
Add to Wish List

I Own It Not interested x Rate it

Recommended because you purchased Probability and Random Processes (edit)

2. 

The Elements of Statistical Learning
by T. Hastie, et al.
Average Customer Review:
In Stock
Publication Date: July 30, 2003

Our Price: $64.76
Used & new from $55.00

Add to cart
Add to Wish List

I Own It Not interested x Rate it

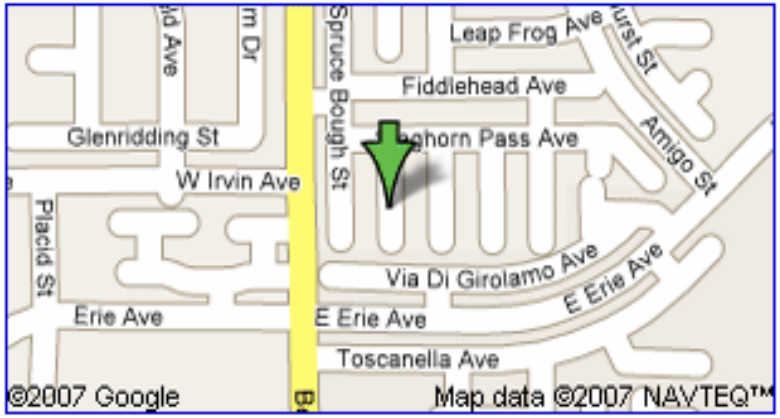
Γιατί Τυχαίους Περιπάτους?

- Εξατομικευμένη αναζήτηση

Web Images Video News Maps Gmail more ▼

Google my car keys Search Advanced Search Preferences

Web



In the front door, where you left them last night.

Where Are **My Car Keys**?

Τι Ψάχνουμε?

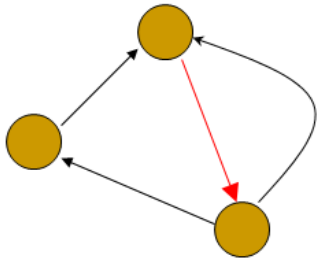
- Να κατατάξουμε κόμβους ενός γραφήματος που αναπαριστά το σύστημα της εκάστοτε εφαρμογής με βάση ένα συγκεκριμένο query
  - Top-k ταυτίσεις για “Random Walks” από Citeseer
  - Ποιοι είναι οι πιο πιθανοί συν-συγγραφείς του “Manuel Blum”
  - Top-k συστάσεις βιβλίων για κάποιον από το Amazon
  - Top-k websites που ταιριάζουν με το “Sound of Music”
  - Top-k συστάσεις φίλων για ένα πρόσωπο όταν δημιουργεί λογαριασμό στο “Facebook”
  - Και πολλά άλλα…..

Ερωτήματα Ενδιαφέροντος για Τ.Π.

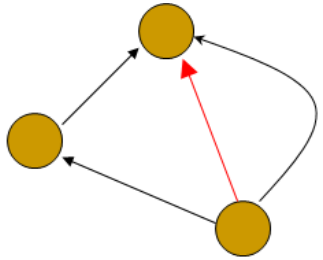
- Υπάρχει πάντα μια stationary distribution? Είναι μοναδική?
  - Ναι, αν ο γράφος συμπεριφέρεται καλά
- Τι θα πει ο γράφος συμπεριφέρεται καλά?
  - Θα το δούμε στη συνέχεια
- Πόσο γρήγορα θα φτάσει ο τυχαίος περίπατος τη στάσιμη κατανομή (αν φυσικά υπάρχει)?
  - Mixing Time!

Καλά Συμπεριφερόμενοι Γράφοι

- **Irreducible**: Υπάρχει μονοπάτι από κάθε ένα κόμβο προς κάθε άλλο

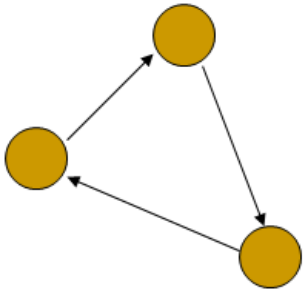


Irreducible

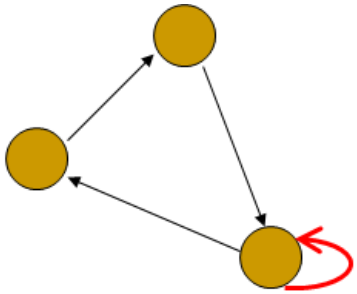


Not irreducible

- **Aperiodic**: Ο GCD (Μ.Κ.Δ.) κάθε μήκους κύκλου είναι 1
  - Το GCD ονομάζεται περίοδος



Periodicity is 3



Aperiodic

Θεώρημα Perron-Frobenius

- Αν μια αλυσίδα Markov είναι irreducible και aperiodic τότε η μεγαλύτερη ιδιοτιμή του πίνακα μεταβάσεων θα είναι ίση με 1 και όλες οι άλλες ιδιοτιμές αυστηρά μικρότερες του 1
  - Έστω οι ιδιοτιμές του P είναι $\{\sigma_i | i = 0 \dots n-1\}$ σε μη-αυξανόμενη κατάταξη των σ_i
  - $\sigma_0 = 1 > \sigma_1 > \sigma_2 >= \dots \dots >= \sigma_n$

- Τα παραπάνω υπονοούν ότι σε ένα καλά συμπεριφερόμενο γράφο υπάρχει μοναδική stationary distribution
- Αυτό σχετίζεται πολύ με το Pagerank (google search)

Χρήσιμες Ιδιότητες για Μη-κατευθυνόμενους

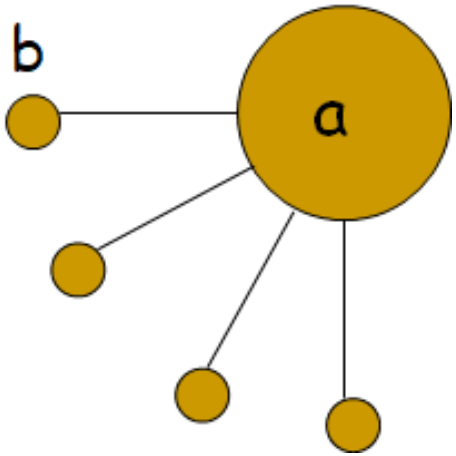
- Ένας συνδεδεμένος και μη-κατευθυνόμενος γράφος είναι και irreducible

- Ένας συνδεδεμένος, μη-διμερής (non-bipartite) και μη-κατευθυνόμενος γράφος έχει stationary distribution ανάλογη με την κατανομή βαθμού!

- Διαισθητικά είναι σωστό γιατί όσο μεγαλύτερος ο βαθμός ενός κόμβου, τόσο πιο πιθανό ο τυχαίος περίπατος να επιστρέψει σε αυτόν

Μετρικές Τυχαίων Περιπάτων

- Πόσος χρόνος χρειάζεται για να βρεθεί ο τυχαίος περίπατος στο b ξεκινώντας από το a?
  - Hitting time
- Πόσος χρόνος χρειάζεται για να βρεθεί ο τυχαίος περίπατος στο b και να επιστρέψει πίσω στο a?
  - Commute time

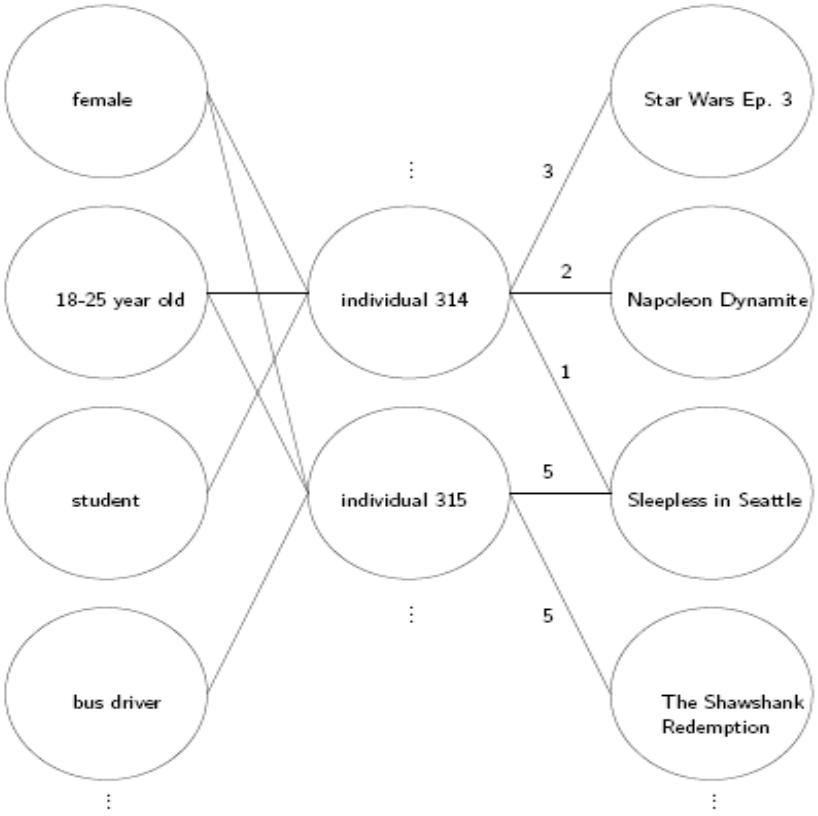


Hitting - Commute times

- Hitting time από κόμβο i στον κόμβο j
  - Αναμενόμενος αριθμός από hops για να βρεθεί ο τυχαίος περίπατος στον κόμβο j ξεκινώντας στον κόμβο i
  - **Δεν είναι συμμετρικός**: $h(a,b) > h(a,b)$
  - $h(i,j) = 1 + \sum_{k \in Neigh(i)} p(i,k)h(k,j)$
- Commute time between node i and j
  - Αναμενόμενος χρόνος να βρεθεί ο τυχαίος περίπατος στον κόμβο j και να επιστρέψει στον κόμβο i
  - $c(i,j) = h(i,j) + h(j,i)$
  - **Είναι συμμετρικός**: $c(a,b) = c(b,a)$

Εφαρμογές: Συστήματα Συστάσεων (1)

An example association graph



Εφαρμογές: Συστήματα Συστάσεων (2)

- Για ένα κόμβο πελάτη i όρισε ομοιότητα ως:
  - h(i,j)
  - c(i,j)
  - Ή cosine similarity

- Επομένως το ερώτημα είναι να υπολογιστούν γρήγορα τα παραπάνω σε πολύ μεγάλους γράφους
  - Fast iterative techniques (Brand 2005)
  - Fast Random Walk with Restart (Tong, Faloutsos 2006)
  - Finding nearest neighbors in graphs (Sarkar, Moore 2007)

Εφαρμογή: Αλγόριθμοι Κατάταξης στο Web

- HITS (Kleinberg, 1998) & Pagerank (Page & Brin, 1998)

- Θα εστιάσουμε στο Pagerank
- Μια ιστοσελίδα είναι σημαντική, αν άλλες σημαντικές ιστοσελίδες δείχνουν (έχουν σύνδεσμο) προς αυτή

- Διαισθητικά: $v(i) = \sum_{j \to i} \frac{v(j)}{\deg^{out}(j)}$

- **v** είναι τελικά η stationary distribution της αλυσίδας Markov που αντιστοιχεί στο web

Pagerank & Perron-Frobenius

- Το θεώρημα Perron-Frobenius ισχύει μόνο αν ο γράφος είναι irreducible και aperiodic
- Πως μπορούμε όμως να το εξασφαλίσουμε για τον web graph?
  - Το κάνουμε με μια μικρή πιθανότητα επανεκκίνησης c
- Σε κάθε βήμα, ο τυχαίος περίπατος:
  - Κάνει άλμα (teleport) σε οποιοδήποτε άλλο κόμβο με πιθανότητα c
  - Κάνει άλμα σε ένα απευθείας γείτονα με πιθανότητα $1 - c$

$$\tilde{\mathbf{P}} = (1-c)\mathbf{P} + c\mathbf{U}$$

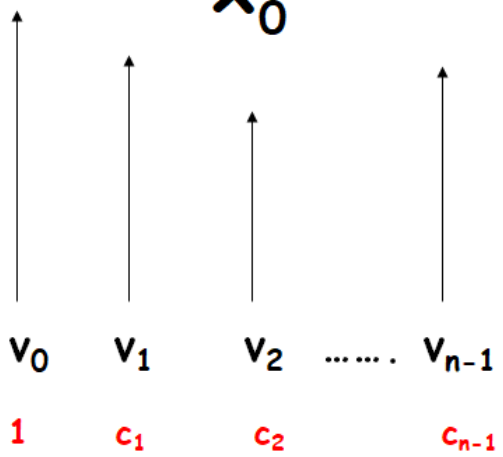
$$\mathbf{U}_{ij} = \frac{1}{n} \forall i, j$$

Power Iteration

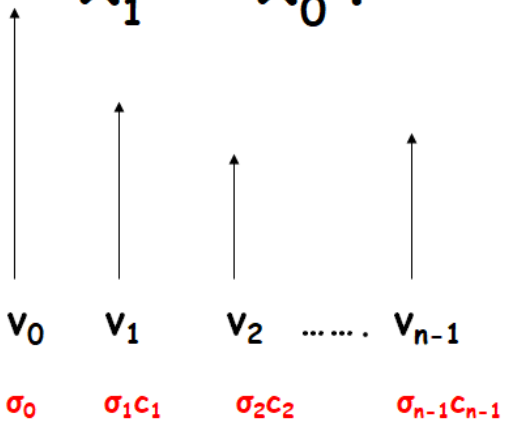
- Power Iteration είναι ένας αλγόριθμος για τον υπολογισμό της stationary distribution
  - Ξεκινά με οποιαδήποτε αρχική κατανομή x_0
  - Υπολογίζει τη $x_{t+1} = x_t P$
  - Σταματά όταν x_{t+1} και x_t είναι σχεδόν ίδιες
- Γιατί δουλεύει?
  - Γράφοντας το x_0 ως γραμμικό συνδυασμό των αριστερών ιδιοδιανυσμάτων $\{v_0, v_1, \dots, v_{n-1}\}$ του P
  - Λαμβάνοντας υπόψη ότι το v_0 είναι η στάσιμη κατανομή
  - $x_0 = c_0 v_0 + c_1 v_1 + c_2 v_2 + \dots + c_{n-1} v_{n-1}$
  - $c_0 = 1$

Power Iteration – Εφαρμογή

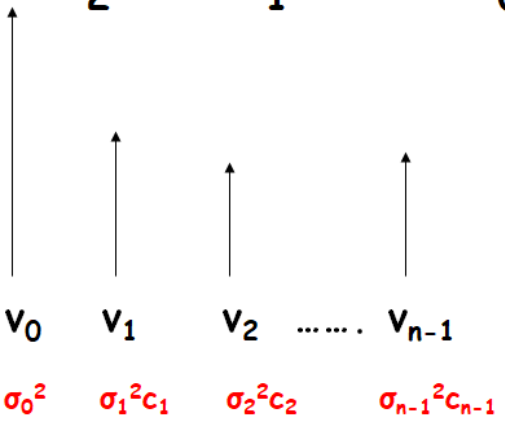
x_0



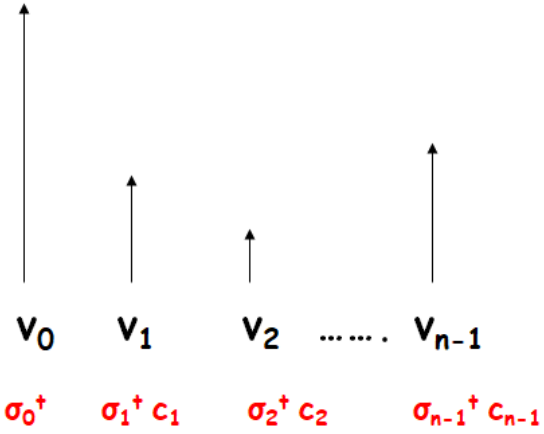
$x_1 = x_0 \tilde{P}$



$x_2 = x_1 \tilde{P} = x_0 \tilde{P}^2$

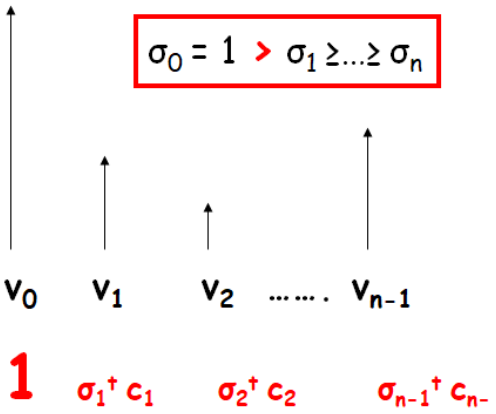


$x_t = x_0 \tilde{P}^t$



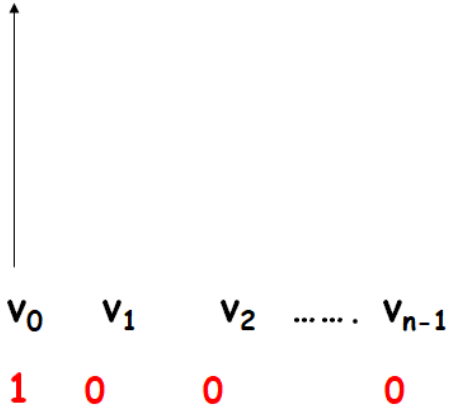
$x_t = x_0 \tilde{P}^t$

$\sigma_0 = 1 > \sigma_1 \geq \ldots \geq \sigma_n$



x_∞

$\sigma_0 = 1 > \sigma_1 \geq \ldots \geq \sigma_n$



Σύγκλιση

- Τυπικά $||x_0 P^t - v_0|| \leq |\lambda|^t$
  - λ είναι η ιδιοτιμή με το δεύτερο μεγαλύτερο πλάτος
- Όσο μικρότερη η δεύτερη μεγαλύτερη ιδιοτιμή (σε πλάτος), τόσο γρηγορότερη η ανάμιξη
- Για $\lambda < 1$ υπάρχει μια μοναδική stationary distribution, δηλαδή το αριστερό ιδιοδιάνυσμα του πίνακα μεταβάσεων
- Ο πίνακας μεταβάσεων που χρησιμοποιεί το Pagerank είναι στην πραγματικότητα $\tilde{P} = (1 - c)P + cU$
- Αποδεικνύεται ότι η δεύτερη μεγαλύτερη ιδιοτιμή του πίνακα αυτού είναι $\leq (1 - c)$
- Σημαίνει ότι ο υπολογισμός του Pagerank συγκλίνει γρήγορα!

Pagerank

- Ψάχνουμε για διάνυσμα $\boldsymbol{v}$ έτσι ώστε

$$\mathbf{v} = (1 - c)\mathbf{v}\mathbf{P} + c\boldsymbol{r}$$

- $\boldsymbol{r}$ είναι η κατανομή στις web-pages

- Αν το $\boldsymbol{r}$ είναι η ομοιόμορφη κατανομή λαμβάνουμε το pagerank!

- Τι συμβαίνει αν το $\boldsymbol{r}$ είναι non-uniform?
  - Personalization

Personalized Pagerank (1)

- Η μόνη διαφορά είναι ότι χρησιμοποιούμε non-uniform teleportation distribution, σε κάθε βήμα άλμα σε ένα σύνολο webpages

- Αναζητούμε το διάνυσμα $\boldsymbol{v}$ έτσι ώστε

$$\mathbf{v} = (1 - c)\mathbf{v}\mathbf{P} + c\mathbf{r}$$

- $\boldsymbol{r}$ είναι το διάνυσμα non-uniform προτιμήσεων που σχετίζεται με ένα χρήστη

- $\boldsymbol{v}$ δίνει “personalized views” του web

Personalized Pagerank (2)

- Pre-computation: Το $\boldsymbol{r}$ δεν είναι γνωστό εκ των προτέρων
- Υπολογισμός του τη στιγμή που γίνεται το query δεν είναι εφικτός
- Μια σημαντική παρατήρηση είναι ότι το διάνυσμα του personalized Pagerank είναι γραμμικό σε σχέση με το $\boldsymbol{r}$

$$\mathbf{r} = \begin{pmatrix} \alpha \\ 0 \\ 1-\alpha \end{pmatrix} \Rightarrow \boldsymbol{v}(\boldsymbol{r}) = \alpha \boldsymbol{v}(\boldsymbol{r}_0) + (1-\alpha)\boldsymbol{v}(\boldsymbol{r}_2)$$

$$\mathbf{r}_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{r}_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Topic-sensitive Pagerank (Haveliwala'01)

- Χώρισε τις webpages σε 16 μεγάλες κατηγορίες
- Για κάθε κατηγορία υπολόγισε το biased personalized Pagerank διάνυσμα κάνοντας ομοιόμορφα άλματα σε websites σε αυτή την κατηγορία
- Τη στιγμή του query, υπολογίζεται η πιθανότητα το query να είναι από οποιαδήποτε από τις παραπάνω κατηγορίες, και το τελικό διάνυσμα page-rank vector υπολογίζεται ως γραμμικός συνδυασμός των biased Pagerank διανυσμάτων που υπολογίστηκαν offline

Personalized Pagerank: Άλλες Προσεγγίσεις

- Scaling Personalized Web Search (Jeh & Widom ’03)
- Towards scaling fully personalized Pagerank: algorithms, lower bounds and experiments (Fogaras et al, 2004)
- Dynamic personalized Pagerank in entity-relation graphs. (Soumen Chakrabarti, 2007)

Τέλος Θεωρίας

Σας ευχαριστώ πολύ για το ενδιαφέρον!

Καλή επιτυχία στις εξετάσεις