Consiglio Nazionale delle Ricerche
Istituto di Elettronica e di Ingegneria dell'Informazione e delle Telecomunicazioni

The PageRank Computation in Google: Randomization and Ergodicity

Roberto Tempo

CNR-IEIIT
Consiglio Nazionale delle Ricerche
Politecnico di Torino
roberto.tempo@polito.it

CNR-IEIIT

Randomization over Networks - 1

- We study three specific applications
- Opinion dynamics in social networks
- Clock synchronization of sensor networks
- PageRank computation in Google

Randomization over Networks - 2

- Common features between these approaches:
- Development of distributed randomized algorithms
- Time averaging

PageRank Problem

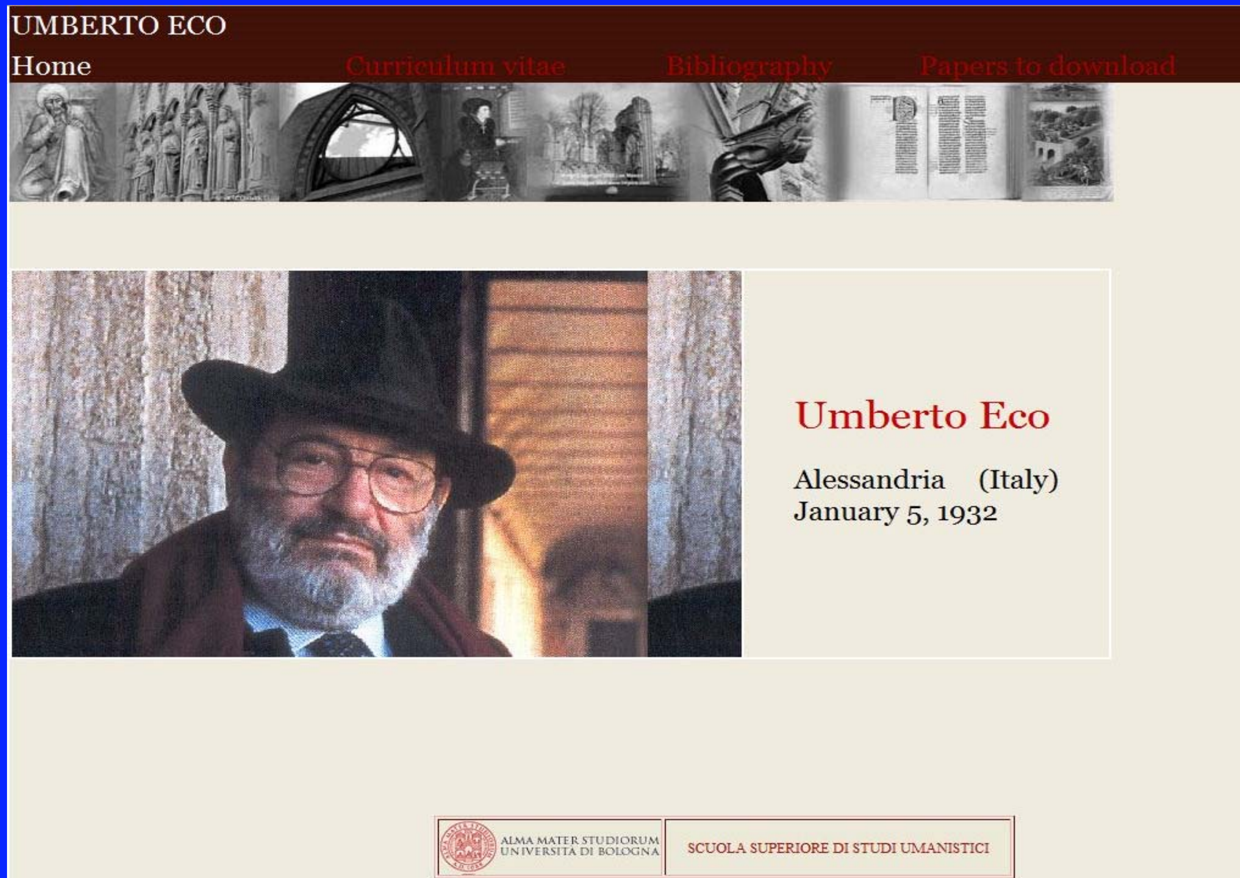
CNR-IEIIT

Umberto Eco

Author of

“The Name of the Rose” translated in more than 40 languages

30 million copies



Translated in Russian by

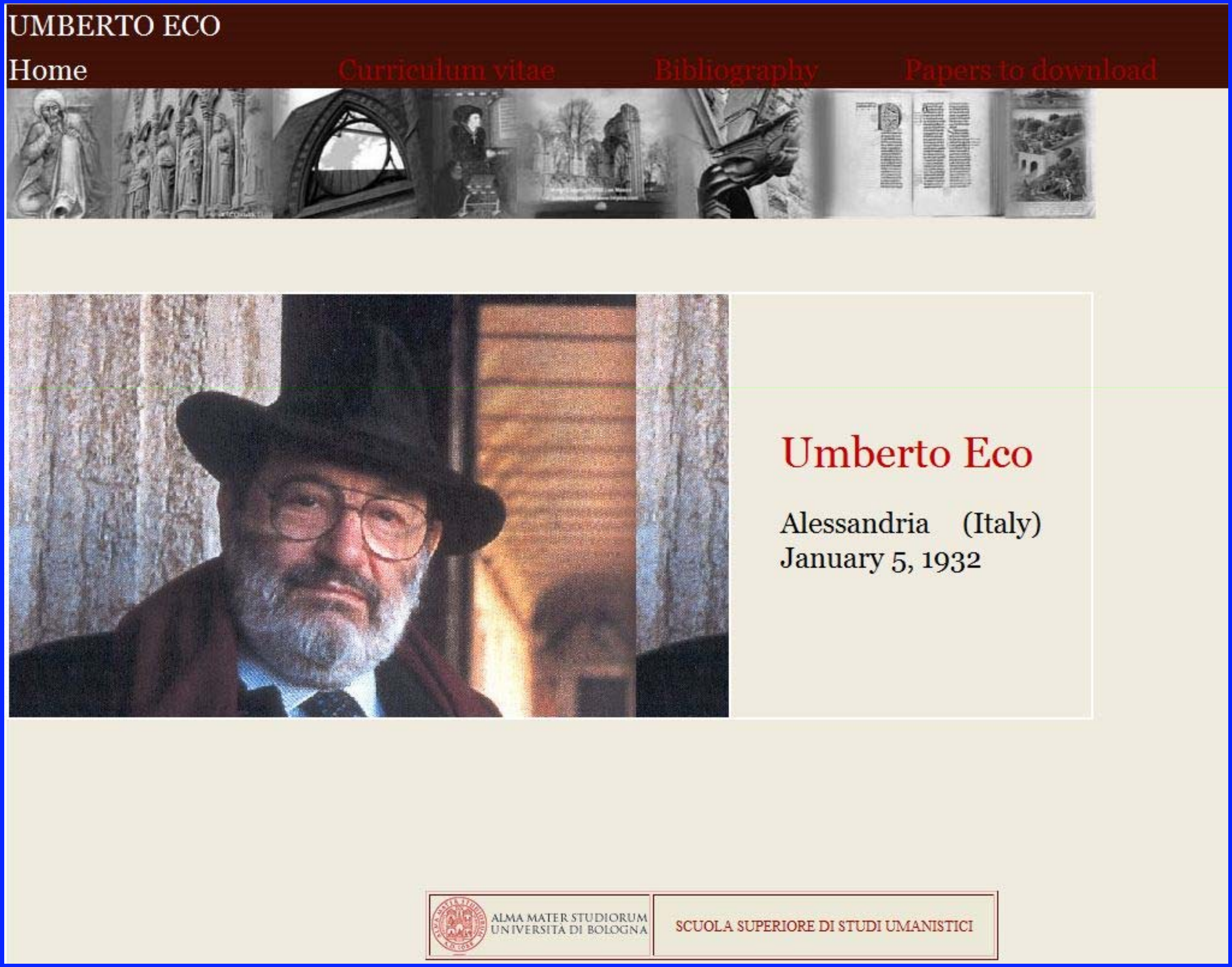
Elena Kostioukovitch

CNR-IEIIT

Website of Umberto Eco

Author of
"The Name of the Rose" translated
in more than
40 languages

Looking for
"Umberto Eco"
in Google we
find the website



PageRank for Umberto Eco

Using a PageRank checker we compute

"PageRank is Google's view of the importance of this page"

PageRank is a numerical value in the interval [0,1] which indicates the importance of the page you are visiting

PageRank

- PageRank is used in search engines to indicate the *importance* of the page currently visited
- PageRank has broader utility than search engines
- Used in various areas for ranking other "objects"
    - scientific journals (Eigenfactor)
    - economics
    - cancer biology and protein detection
    - ranking top sport players
    - …

CNR-IEIIT

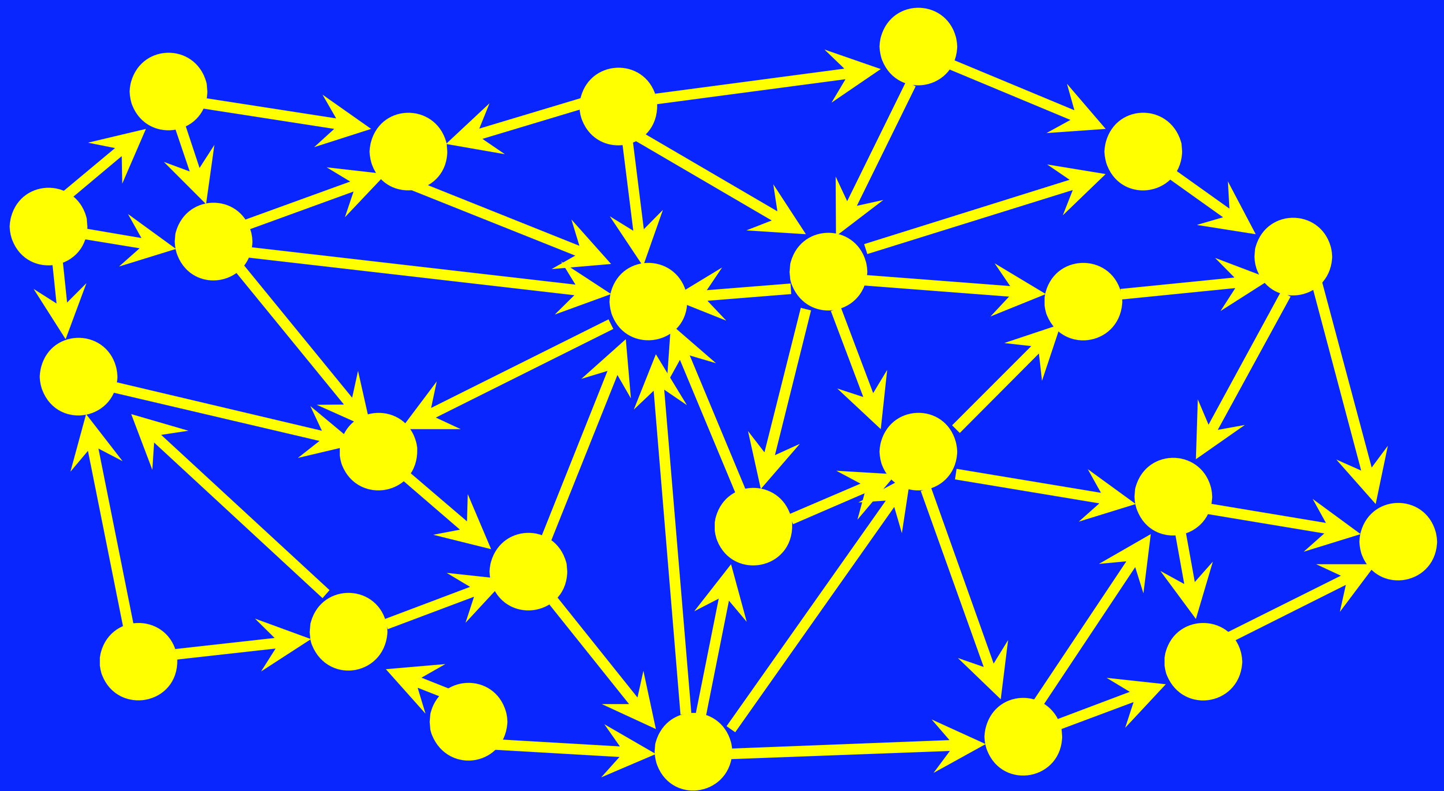
Random Surfer Model

- Network consisting of servers (nodes) connected by directed communication links
- Web surfer moves along randomly following the hyperlink structure
- When arriving at a page with several outgoing links, one is chosen at random, then the random surfer moves to a new page, and so on…

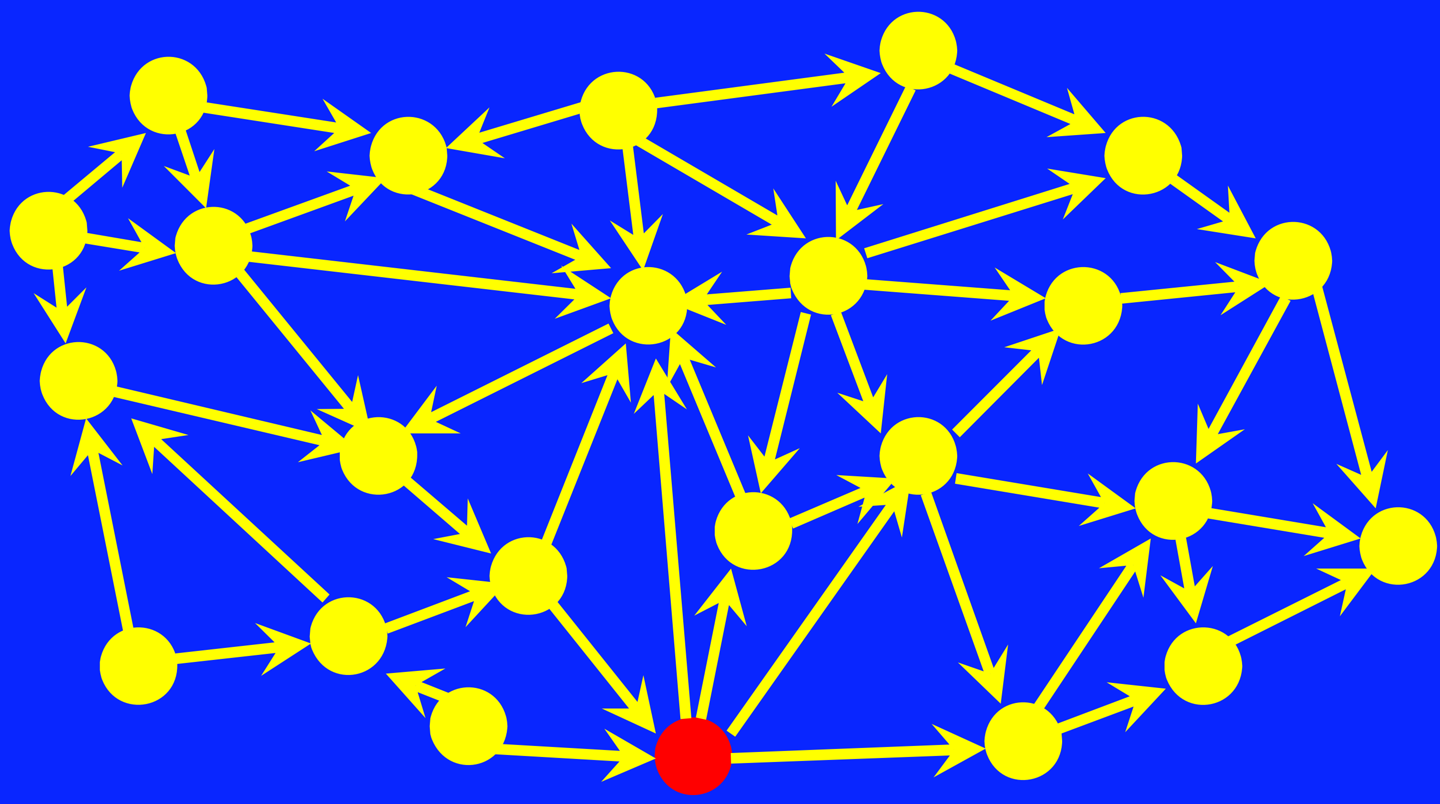
Random Surfer Model

CNR-IEIIT

- Web representation with incoming and outgoing links

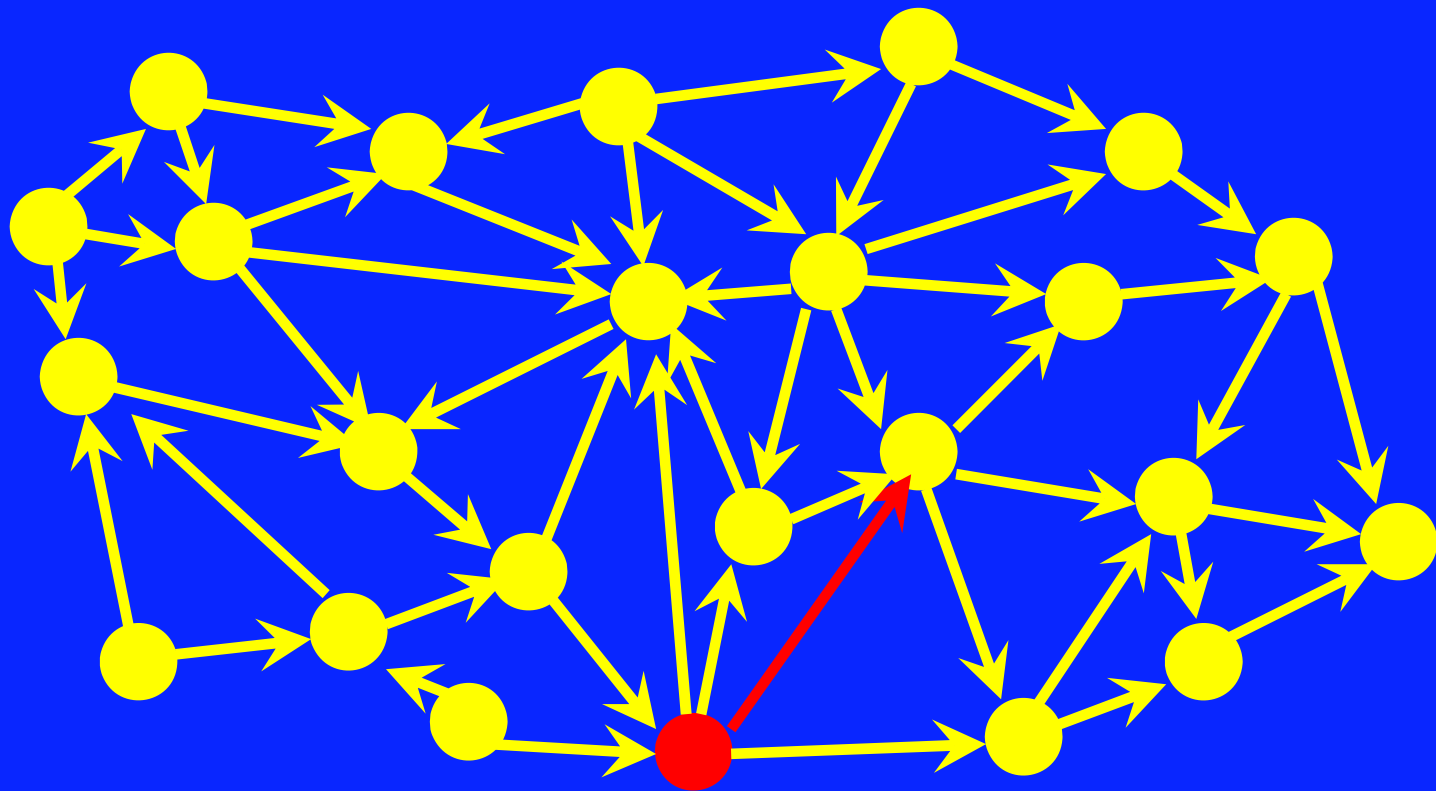


Random Surfer Model



Random Surfer Model

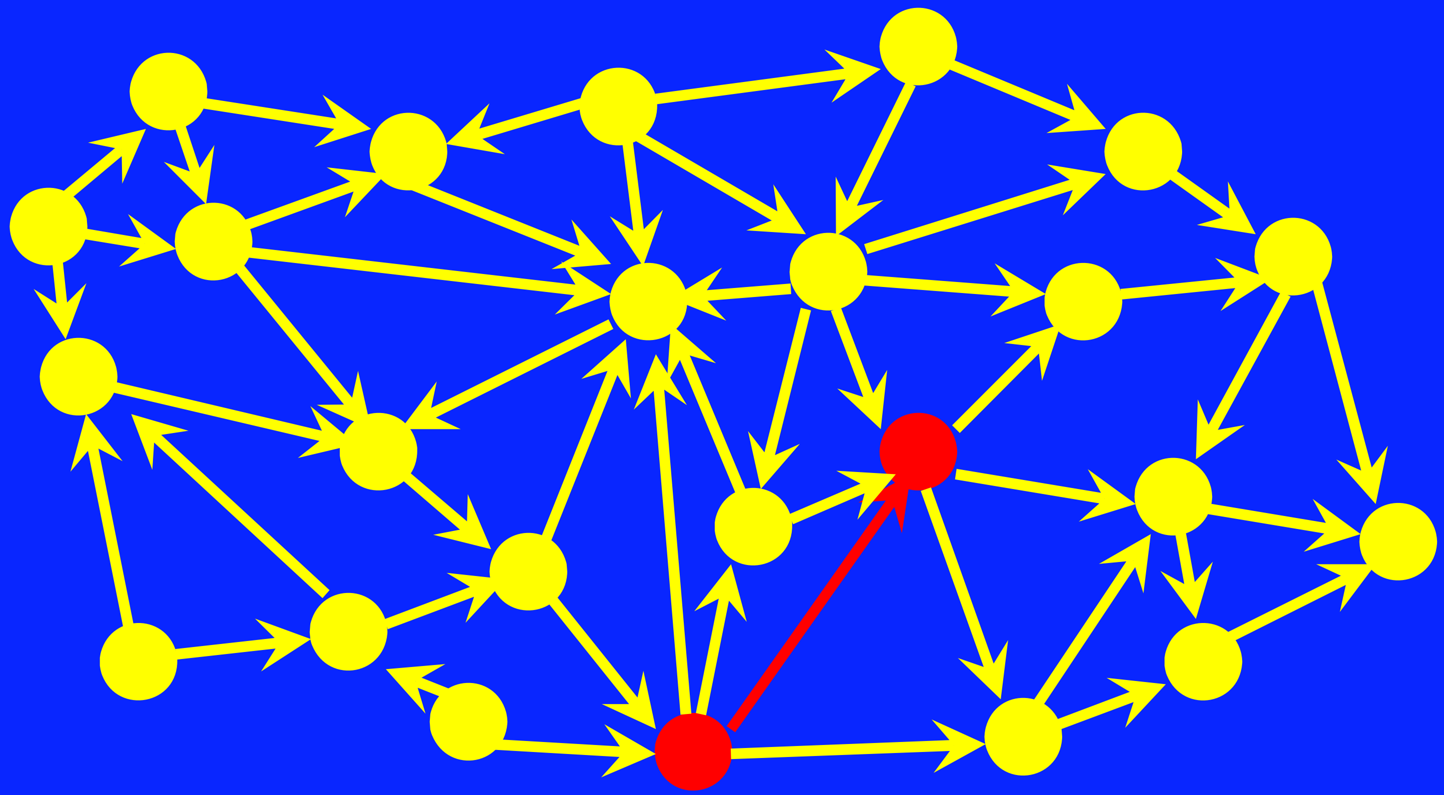
- Pick an outgoing link at random



CNR-IEIIT

Random Surfer Model

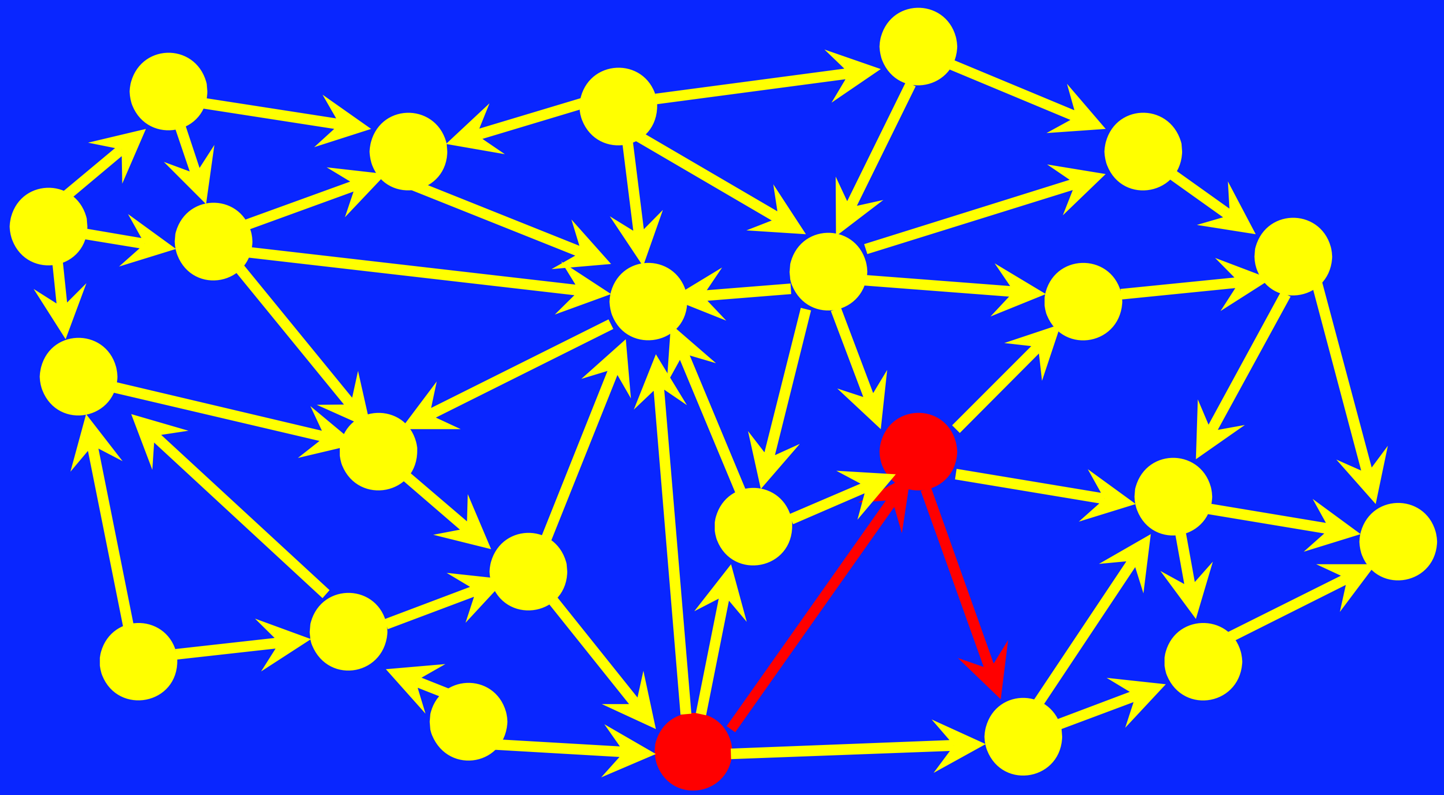
- Arriving at a new web page



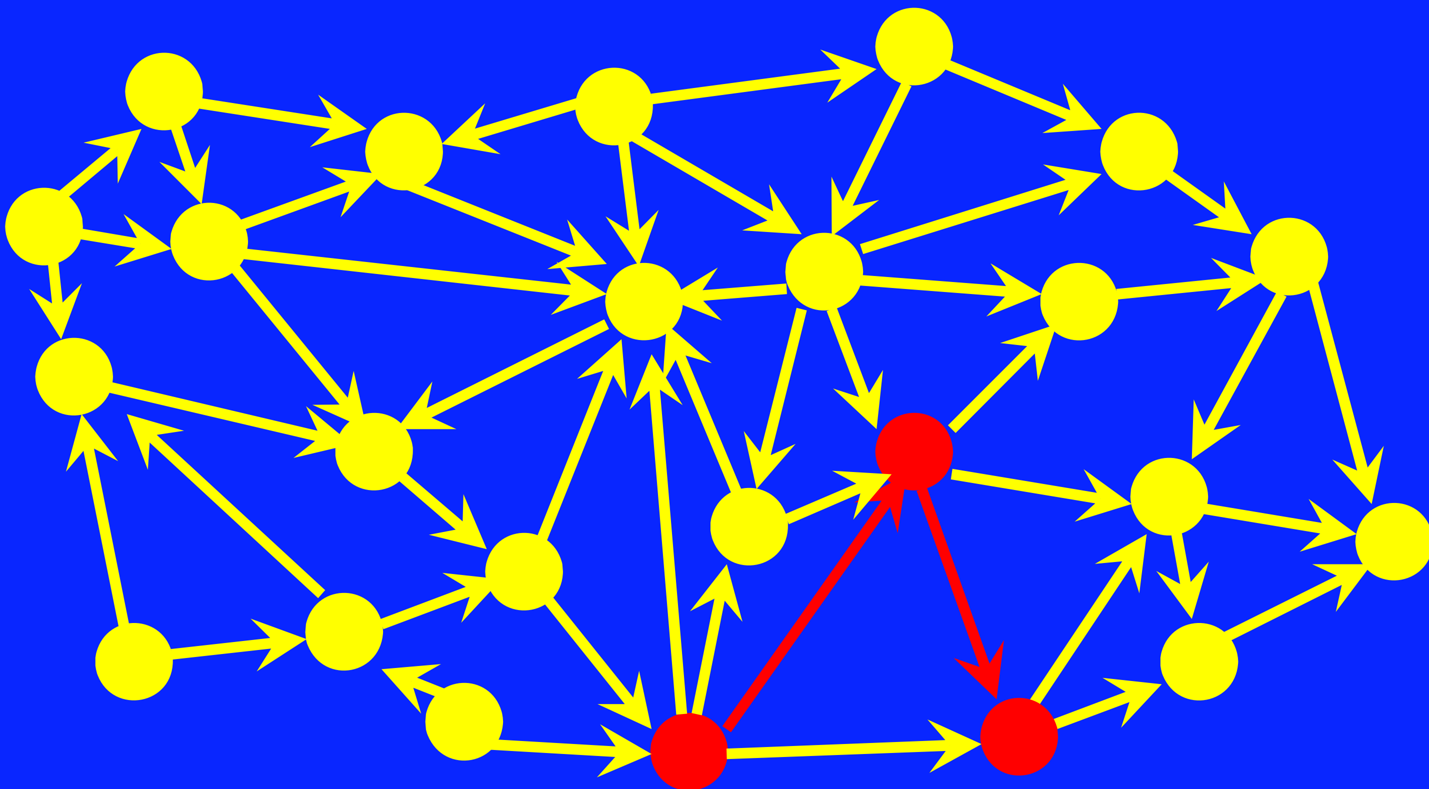
CNR-IEIIT

Random Surfer Model

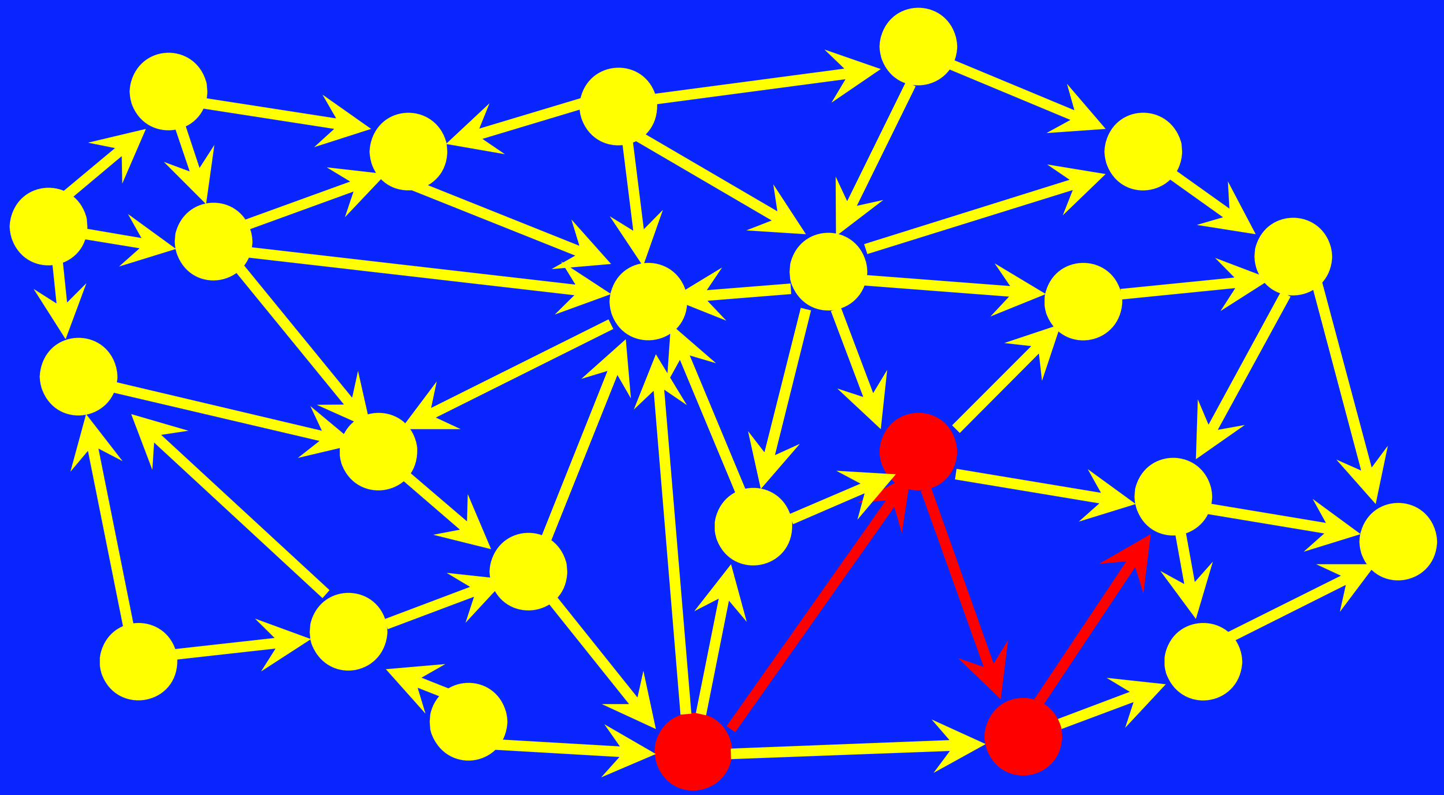
- Pick another outgoing link at random



Random Surfer Model

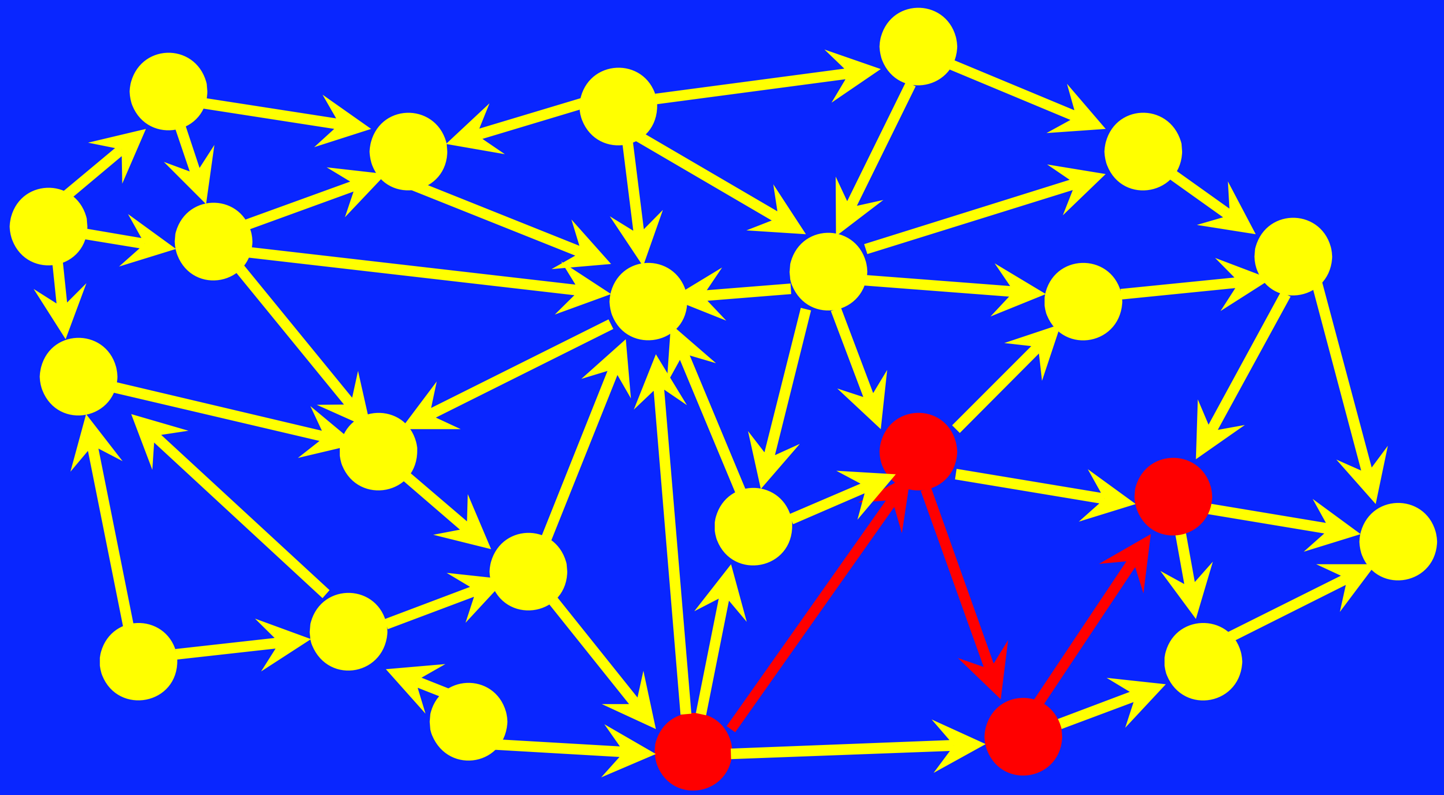


Random Surfer Model



CNR-IEIIT

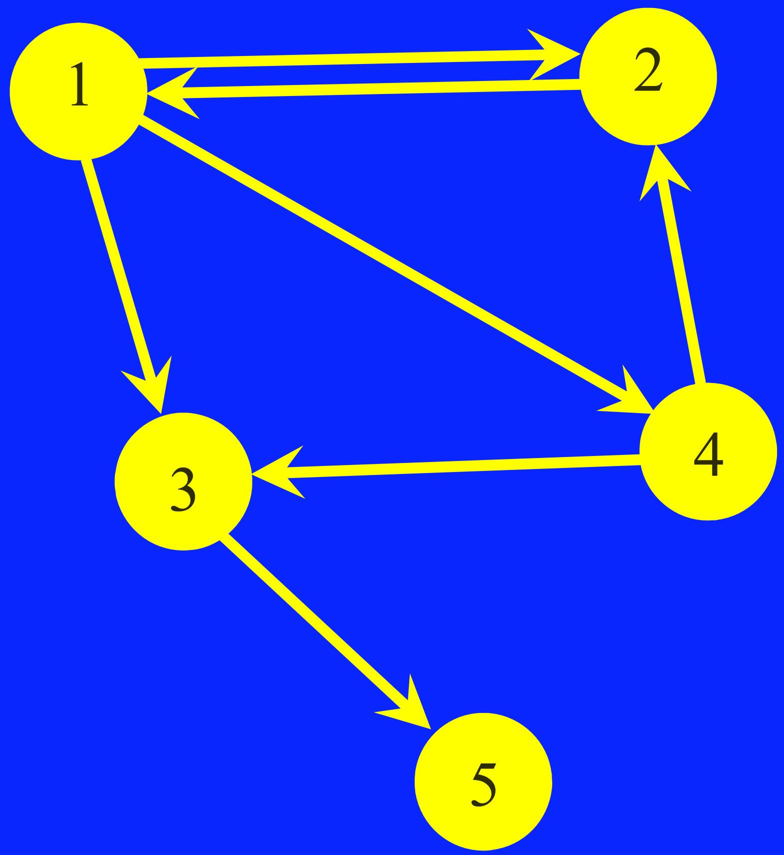
Random Surfer Model



Random Surfer Model

- If a page is *important* then it is visited more often...
- The time the random surfer spends on a page is a measure of its importance
- If important pages point to your page, then your page becomes important because it is visited often
- What is the *probability* that your page is visited?
- Need to rank the pages in order of importance for facilitating the web search

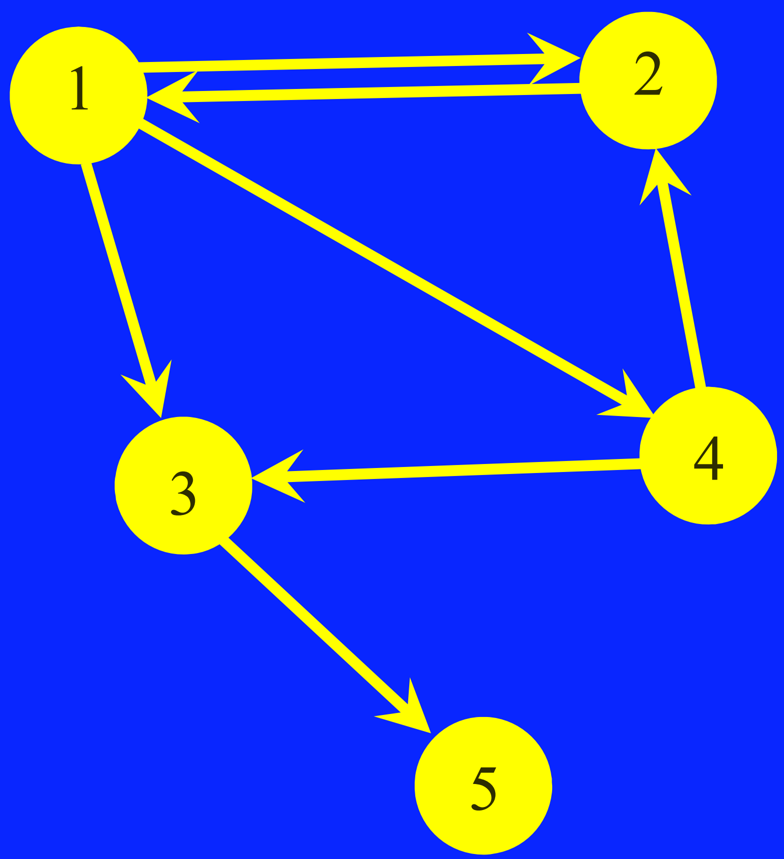
Graph Representation



- Directed graph with nodes (pages) and links representing the web
- Graph is not necessarily strongly connected (from 5 you cannot reach other pages)
- Graph is constructed using *crawlers and spiders* moving continuously along the web

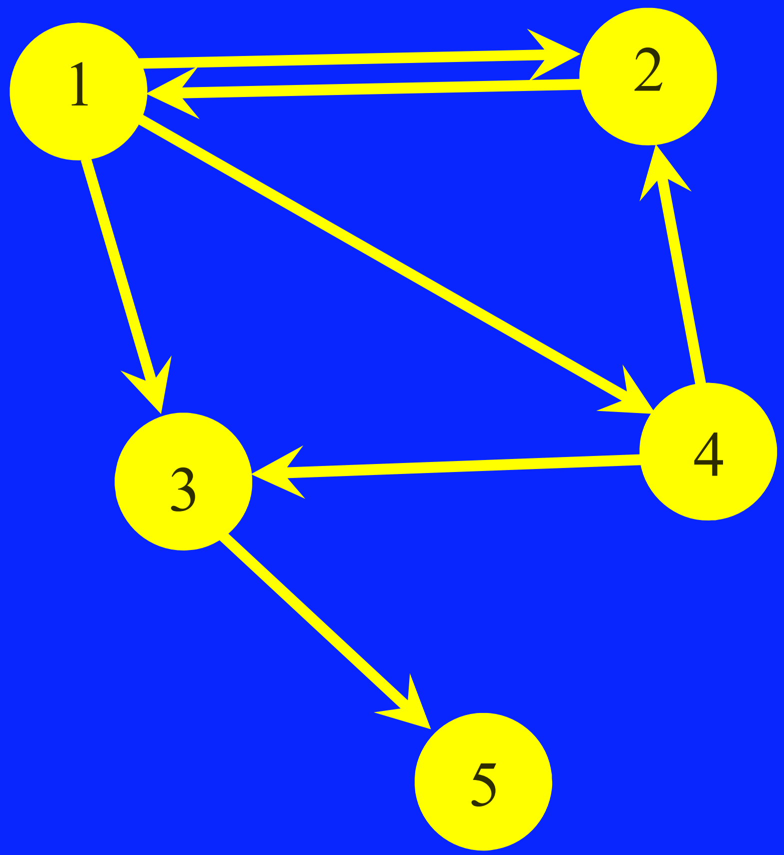
CNR-IEIIT

Hyperlink Matrix



- For each node we count the number of outgoing links and normalize their sum to 1
- Hyperlink matrix is a nonnegative (column) substochastic matrix

Hyperlink Matrix



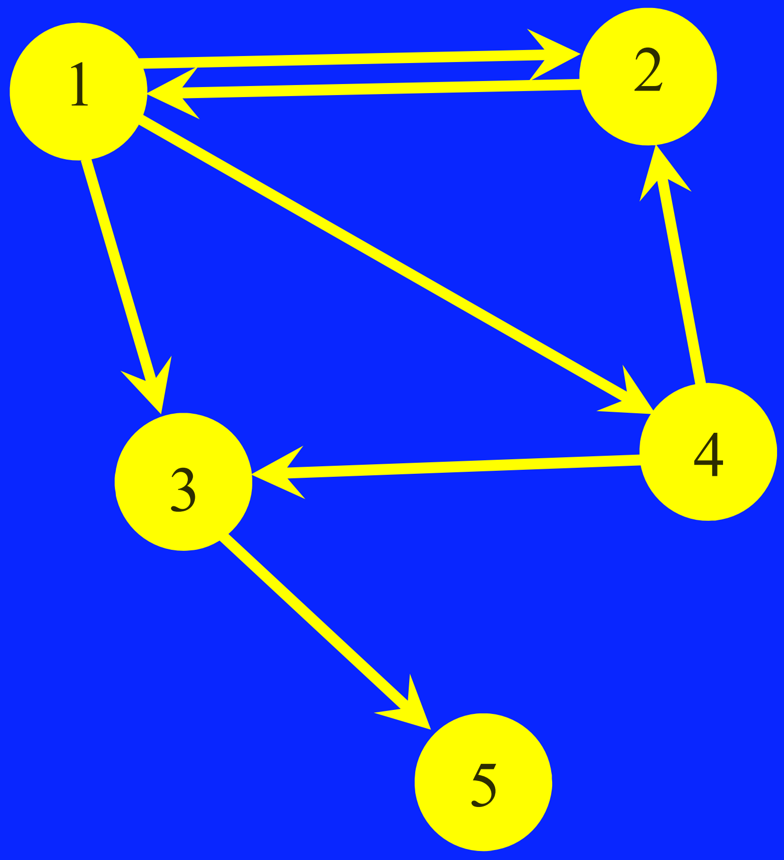
$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 1/2 & 0 \\ 1/3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

CNR-IEIIT

Issue of Black Holes

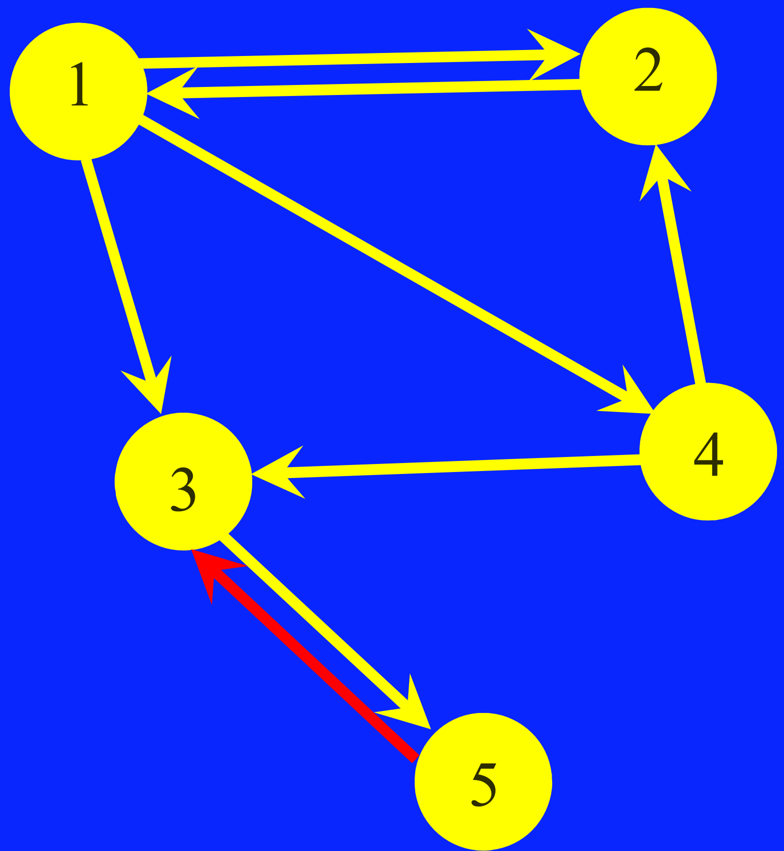
- We have *black holes* (pages having no outgoing link)
- Random surfer gets stuck when visiting a pdf file
- The problem is not well-posed: the probability associated to page 5 increases and increases… the probabilities of all the other pages decrease and decrease…
- At the end, we have p5 = 1 and p1 = p2 = p3 = p4 = 0
- In this case the “back button” of the browser is used
- Mathematically, the hyperlink matrix is nonnegative and (column) substochastic

Page 5 is a Black Hole



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

Easy Fix: Add New Link



- We add a new outgoing link from page 5 to page 3

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

Assumption: No Black Holes

- This is a web modeling problem
- Assume that there are *no black holes*
- This implies that A is a nonnegative stochastic matrix (instead of substochastic)

CNR-IEIIT

Random Surfer Model and Markov Chains

- Random surfer model is represented as a Markov chain

$$x(k+1) = Ax(k)$$

where $x(k)$ is a probability vector

- $x(k) \in [0,1]^n$ and $\sum_i x_i(k) = 1$
- $x_i(k)$ represents the *importance* of the page i at time k

CNR-IEIIT

Convergence of the Markov Chain

- Question: Does the Markov chain converge to a stationary value

$$x(k) \to x^* \quad \text{for} \quad k \to \infty$$

representing the probability that the pages are visited?

- Answer: No

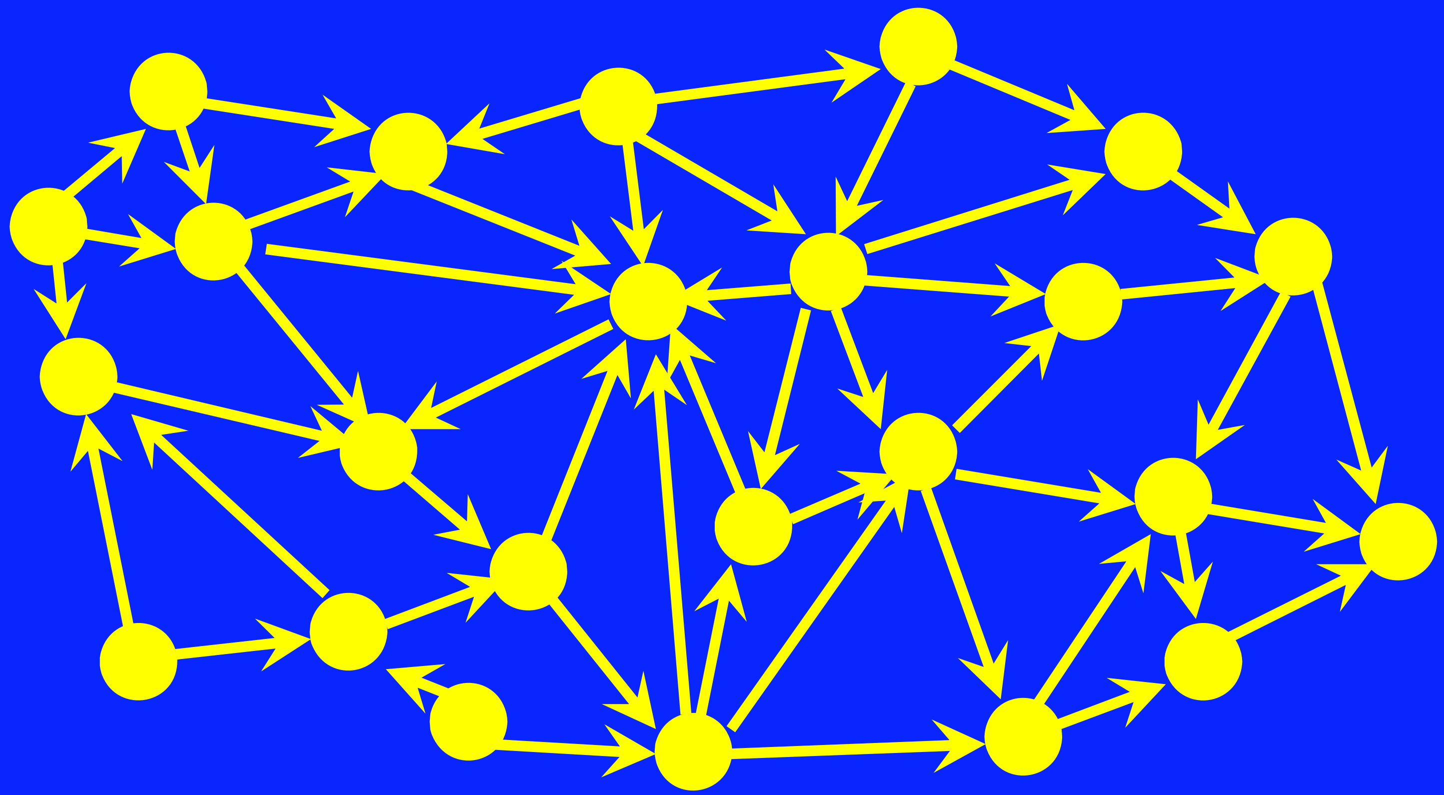
- Recall that the matrix A is a nonnegative stochastic matrix
- We introduce a different model

Teleportation Model

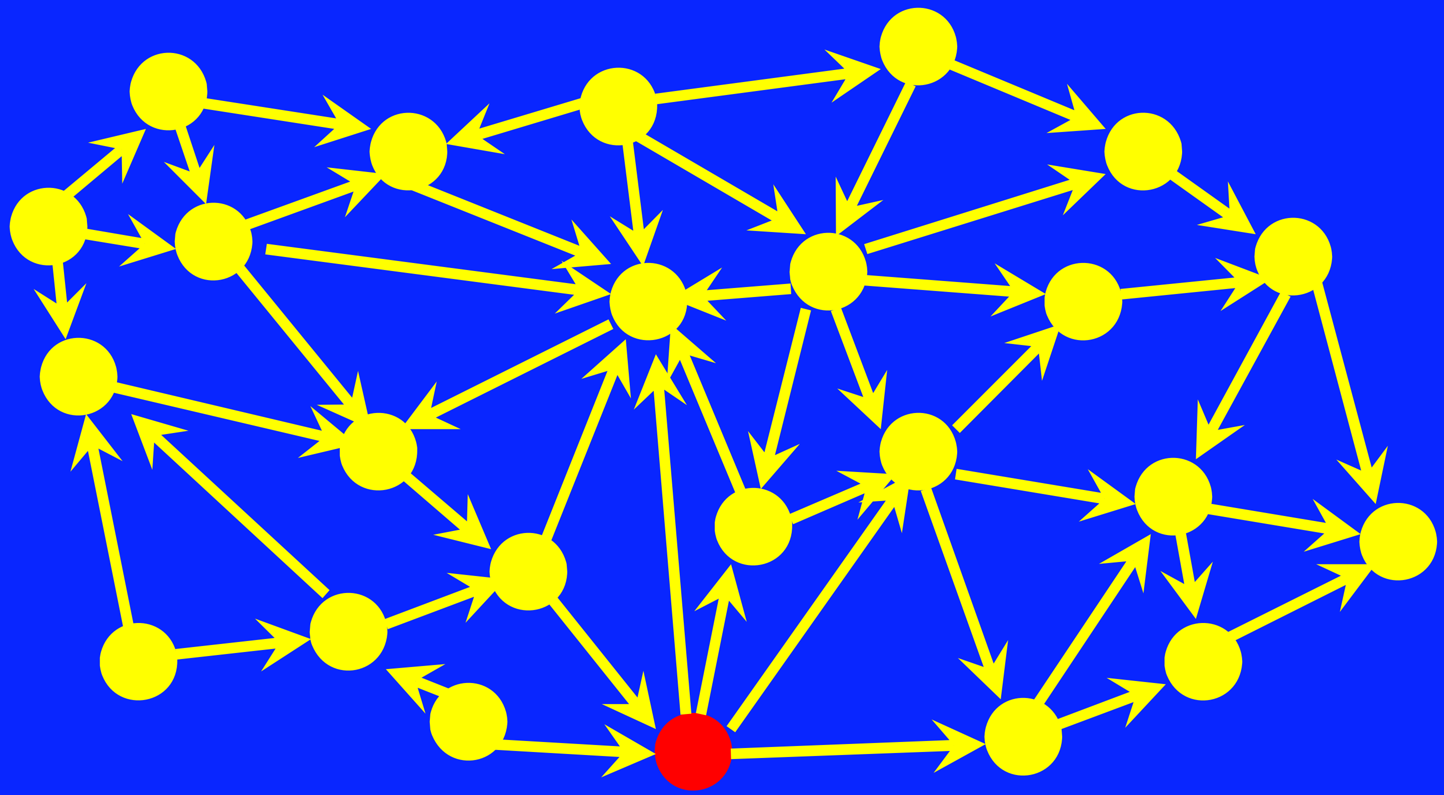
- Teleportation: After a while the random surfer gets bored and decides to "jump" to another page not directly connected to that currently visited
- New page may be geographically or content-based located far away

Recall the Random Surfer Model

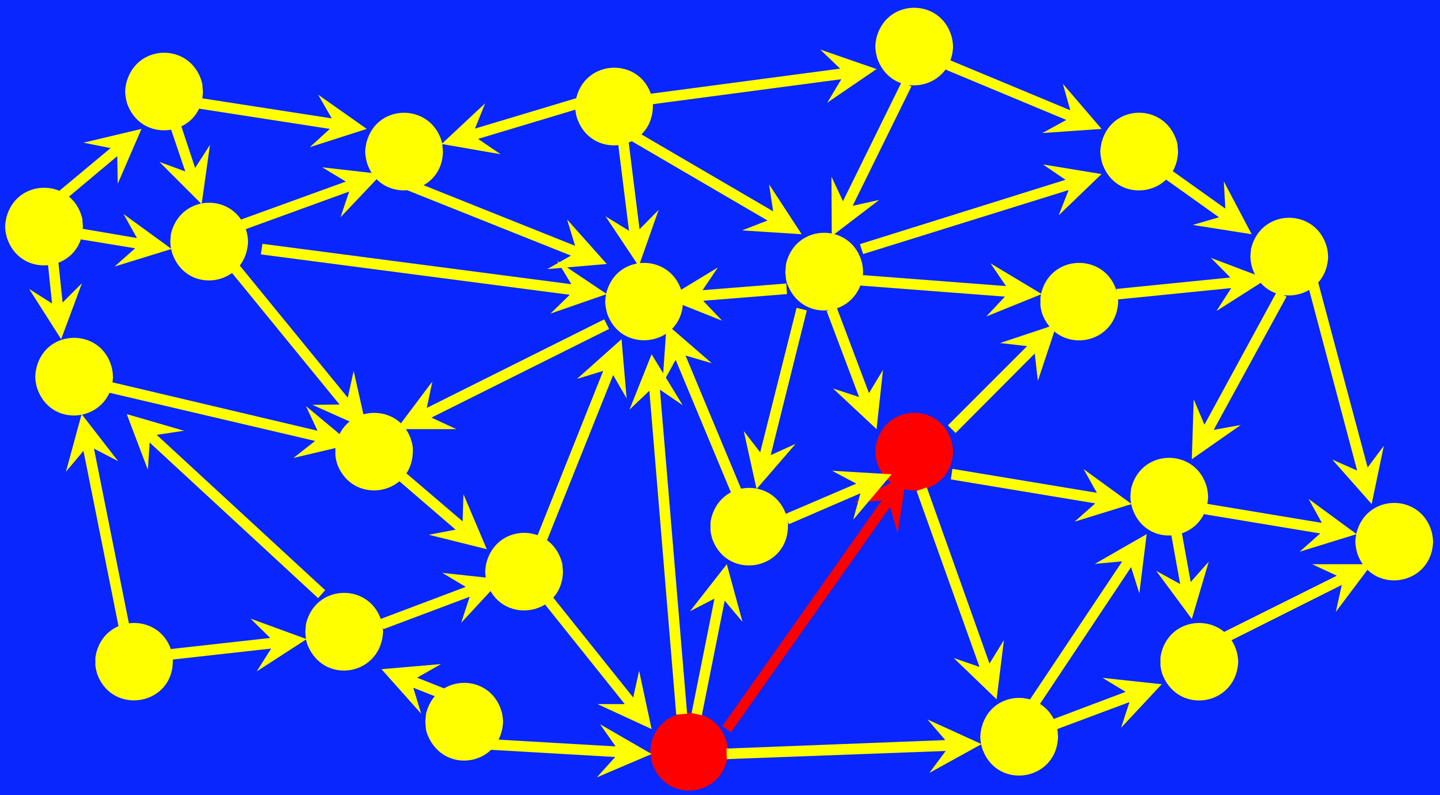
- Web representation with incoming and outgoing links



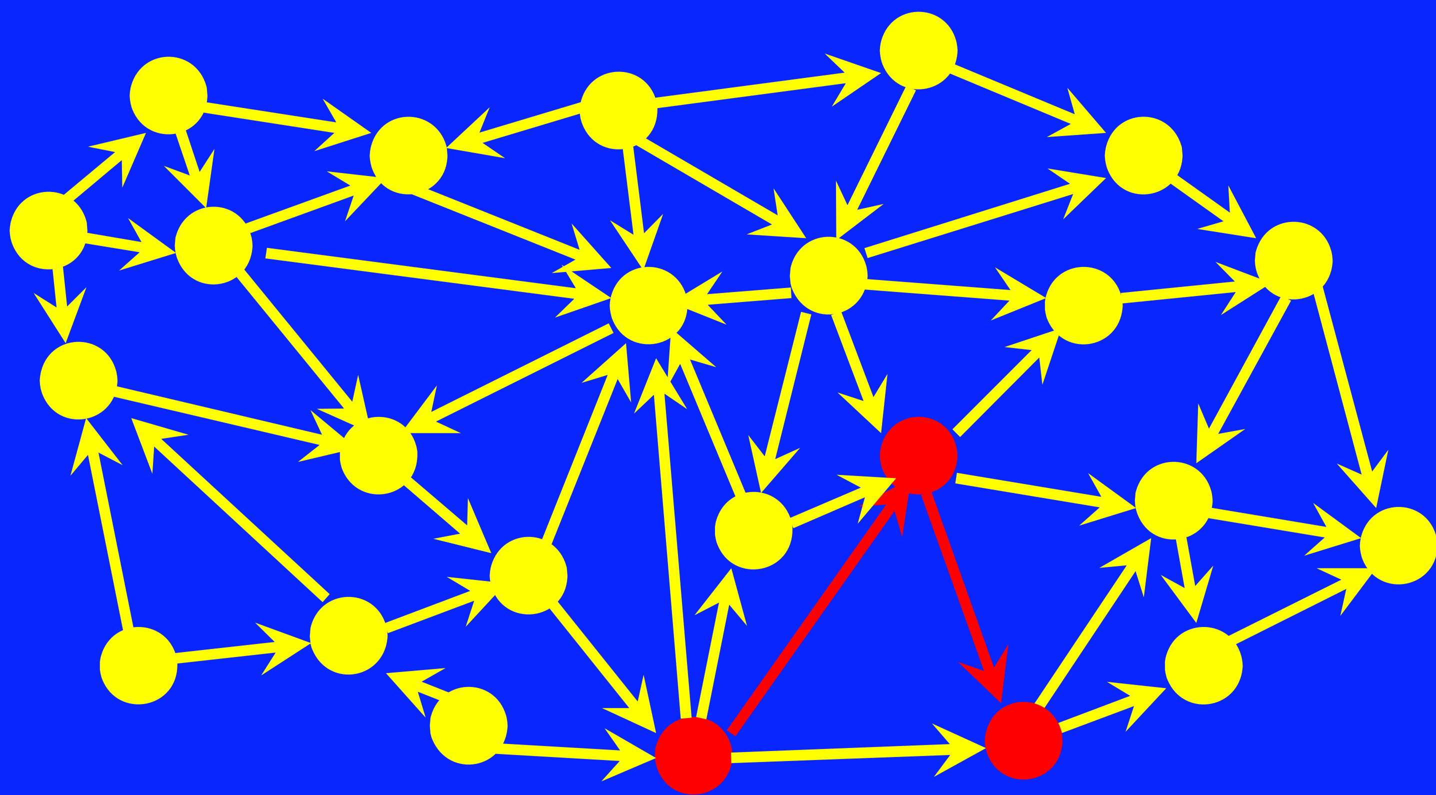
Recall the Random Surfer Model



Recall the Random Surfer Model

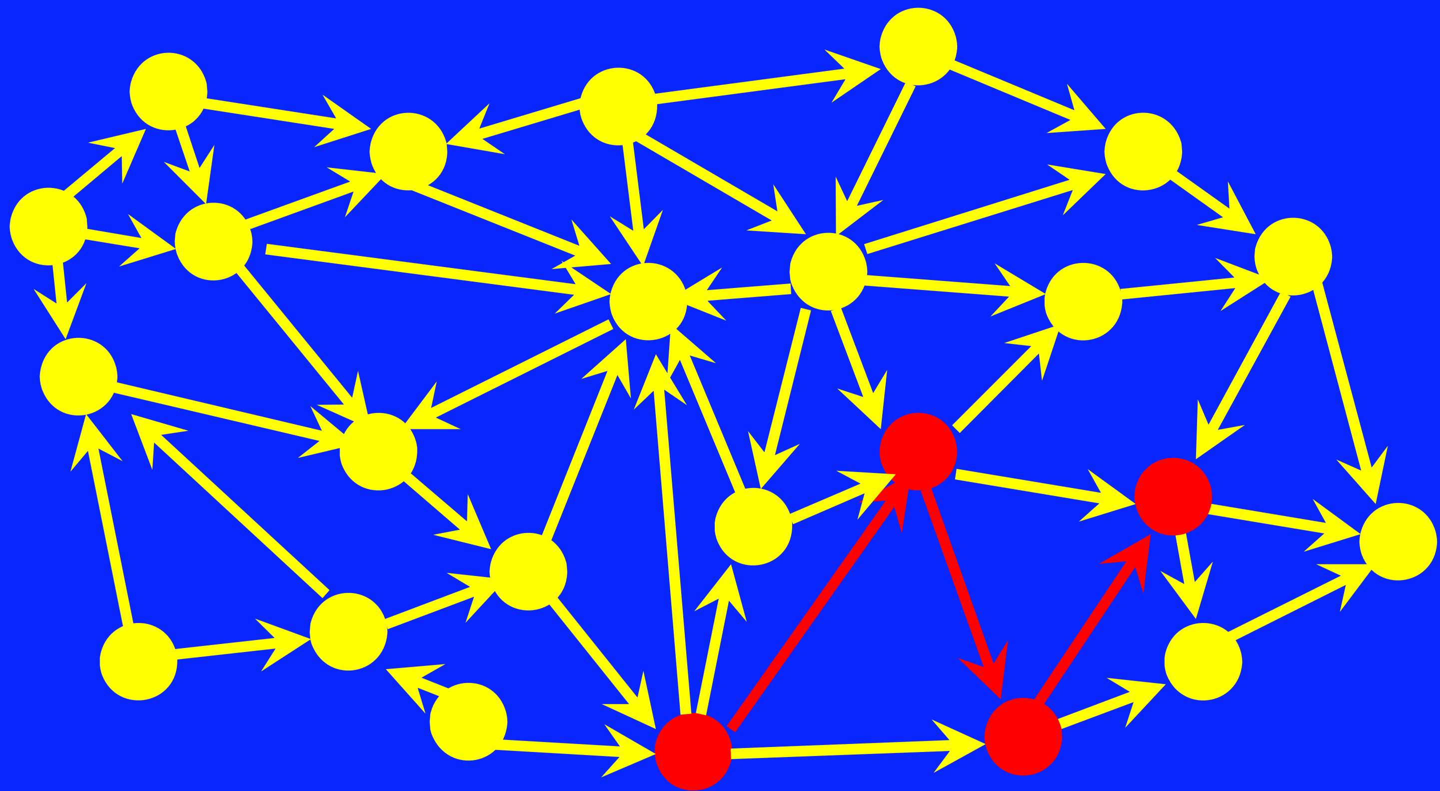


Recall the Random Surfer Model



CNR-IEIIT

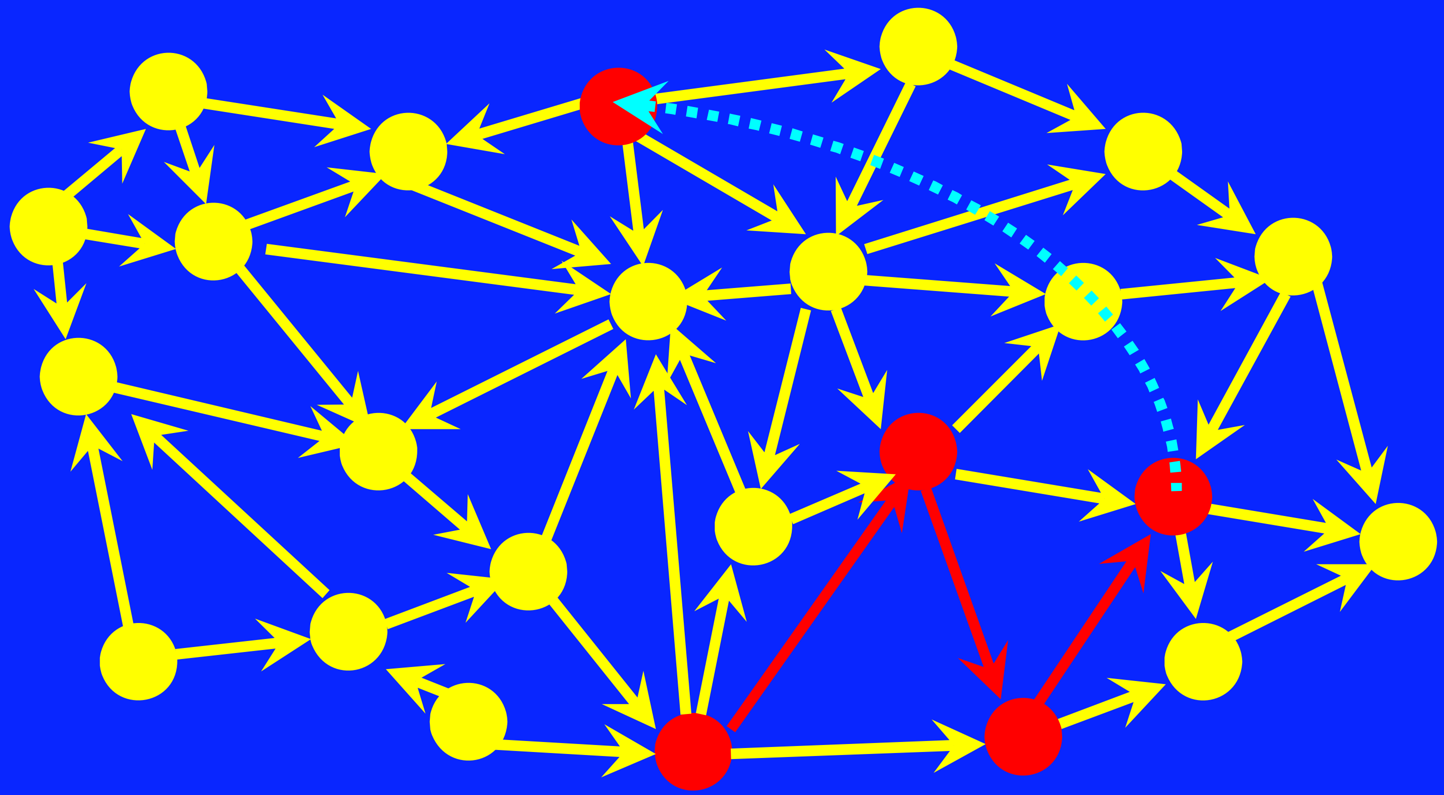
Recall the Random Surfer Model



CNR-IEIIT

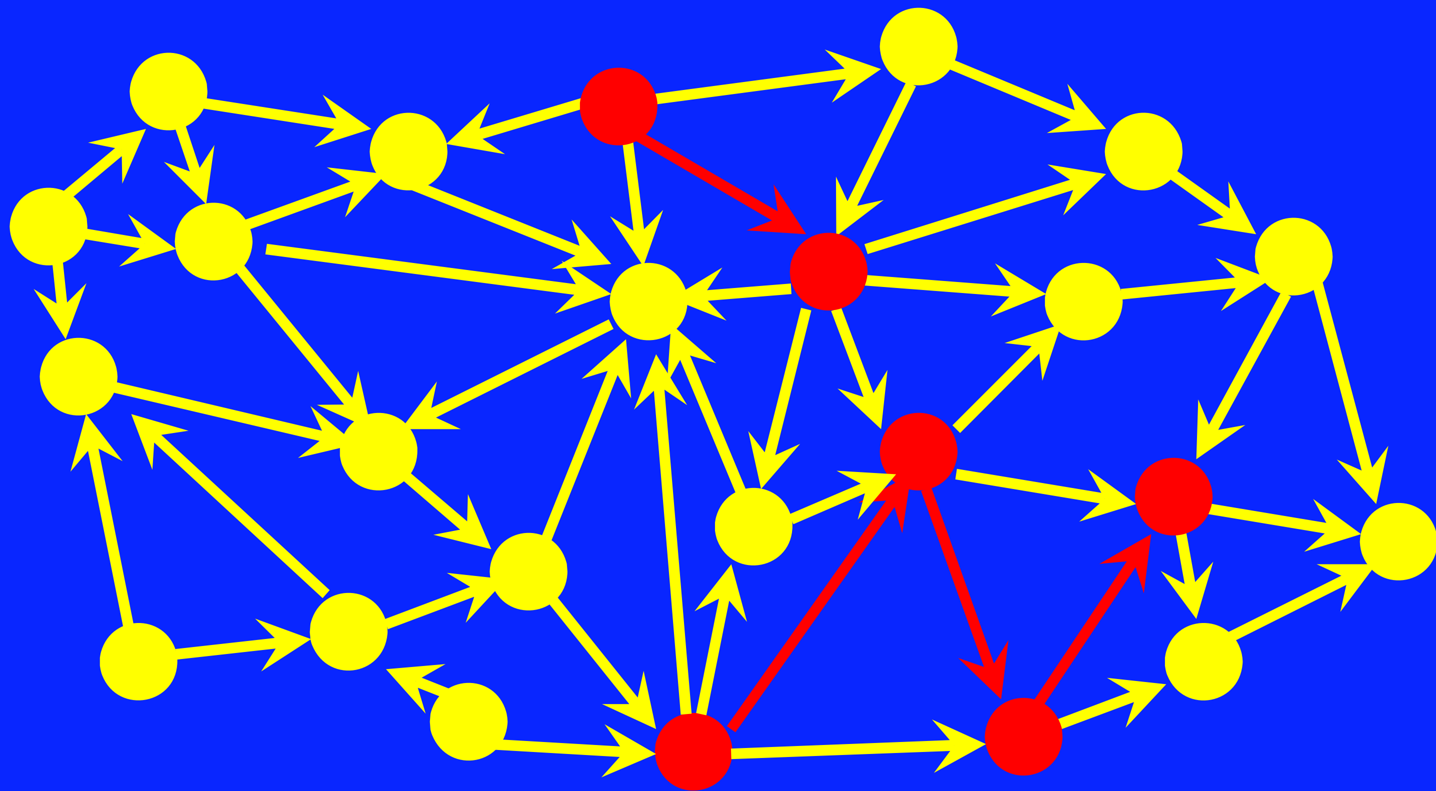
Teleportation Model

- We are "teleported" to a web page located far away



Random Surfer Model Again

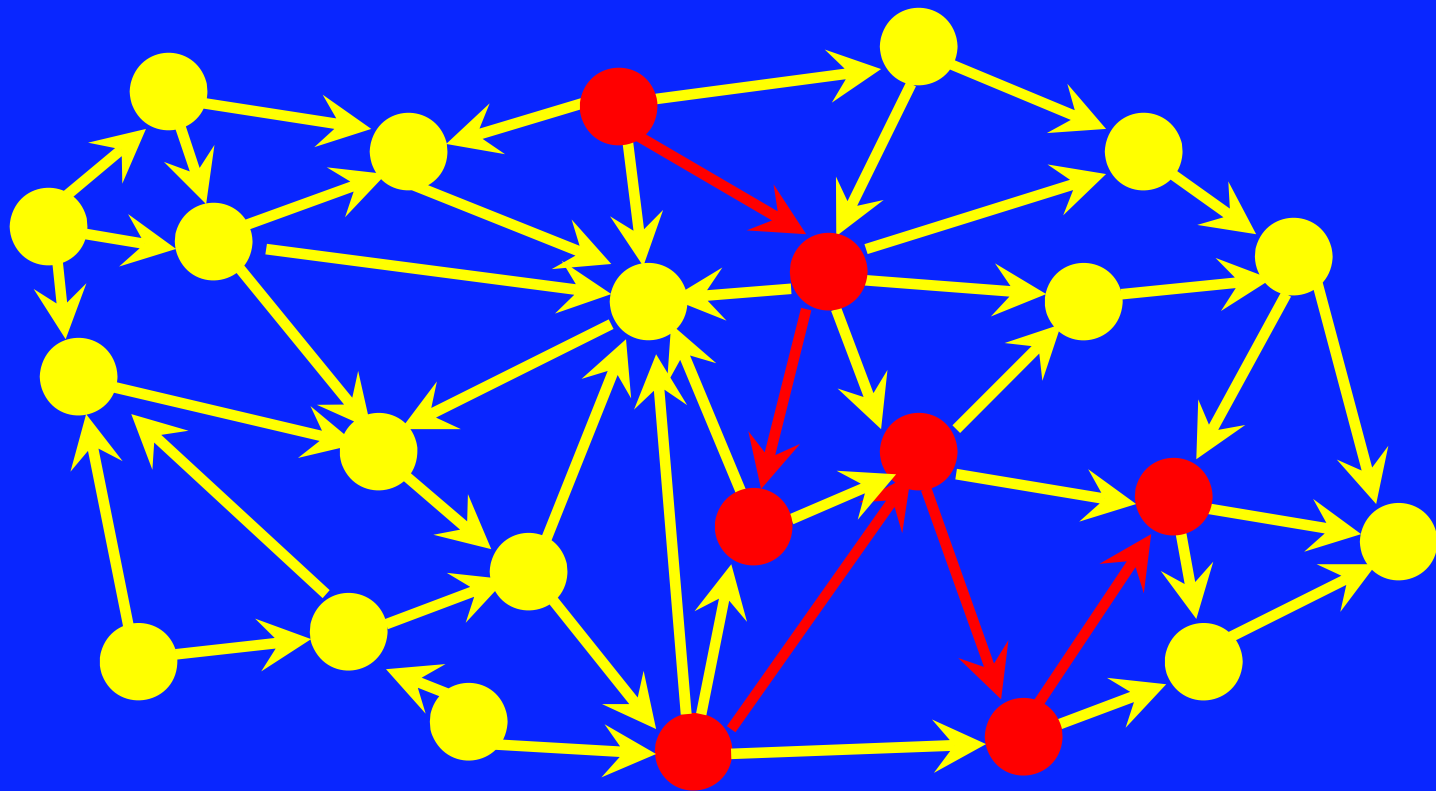
- Pick another outgoing link at random



CNR-IEIIT

Random Surfer Model Again

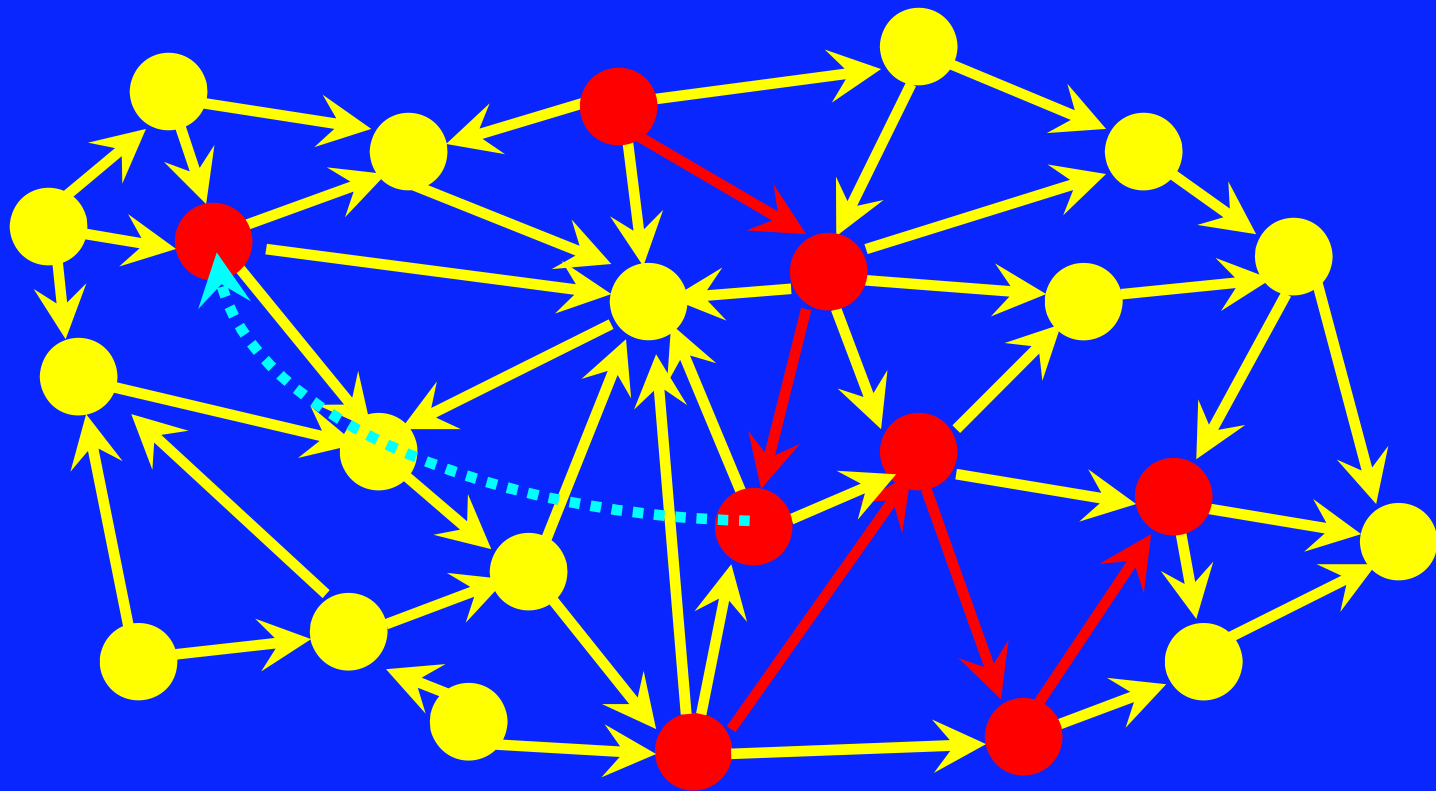
- Pick another outgoing link at random



Teleportation Model Again

CNR-IEIIT

- We are teleported to another web page located far away



CNR-IEIIT

Convex Combination of Matrices

- Teleportation model is represented as a convex combination of matrices A and S/n
- $S = \mathbf{1}\,\mathbf{1}^T$ is a rank-one matrix
- $\mathbf{1}$ vector with all entries equal to one

$$S = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

- Consider a matrix M defined as

$$M = (1 - m)\, A + m/n\, S \qquad m \in (0,1)$$

where n is the number of pages

- The value $m = 0.15$ is used at Google

Matrix M

- M is a convex combination of two nonnegative stochastic matrices and $m \in (0,1)$
- M is a strictly *positive* stochastic matrix

Convergence of the Markov Chain

- Consider the Markov chain

$$x(k+1) = M\, x(k)$$

where M is a strictly positive stochastic matrix

- If $\sum_i x_i(0) = 1$ convergence is guaranteed by Perron Theorem

$$x(k) \to x^* \quad \text{for} \quad k \to \infty$$

$$x^* = M\, x^* = \left[(1 - m)\, A + m/n\, S\right] x^* \qquad m \in (0,1)$$

- Corresponding graph is strongly connected

CNR-IEIIT

PageRank: Bringing Order to the Web

- Rank n web pages in order of importance
- Ranking is provided by x^*
- PageRank x^* of the hyperlink matrix M is defined as

$$x^* = M x^* \qquad \text{where} \quad x^* \in [0,1]^n \ \text{ and } \sum_i x_i^* = 1$$

S. Brin, L. Page (1998)

S. Brin, L. Page, R. Motwani, T. Winograd (1999)

CNR-IEIIT

PageRank: Bringing Order to the Web

- Rank n web pages in order of importance
- Ranking is provided by x^*
- PageRank x^* of the hyperlink matrix M is defined as

$$x^* = M x^* \qquad \text{where} \qquad x^* \in [0,1]^n \ \text{ and } \sum_i x_i^* = 1$$

- x^* is the stationary distribution of the Markov Chain (steady-state probability that pages are visited is x^*)
- x^* is a nonnegative unit eigenvector corresponding to the eigenvalue 1 of M

Stochastic Matrices and Perron Theorem

- The eigenvalue 1 of M is a simple eigenvalue and any other eigenvalue (possibly complex) is strictly smaller than 1 in absolute value
- There exists an eigenvector x^* of M with eigenvalue 1 such that all components of x^* are positive
- M is irreducible and the corresponding graph is strongly connected (for any two vertices i,j there is a sequence of edges which connects i to j)
- Stationary probability vector x^* is the eigenvector of M
- x^* exists and it is unique

PageRank Computation

PageRank Computation with Power Method

CNR-IEIIT

- PageRank is computed with the power method

$$x(k+1) = M\,x(k)$$

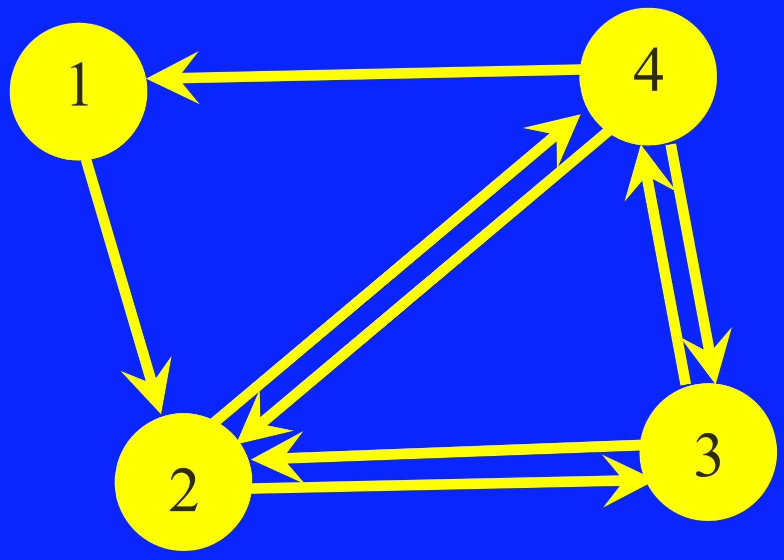
- If $\sum_i x_i(0) = 1$ convergence is guaranteed because M is a strictly positive matrix (Perron Theorem)

$$x(k) \to x^* \quad \text{for} \quad k \to \infty$$

- PageRank computation requires 50-100 iterations
- This computation takes about a week and it is performed centrally at Google once a month

CNR-IEIIT

PageRank Computation with Power Method



$$A = \begin{bmatrix} 0 & 0 & 0 & 1/3 \\ 1 & 0 & 1/2 & 1/3 \\ 0 & 1/2 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \qquad m = 0.15$$

$$M = \begin{bmatrix} 0.038 & 0.037 & 0.037 & 0.321 \\ 0.887 & 0.037 & 0.462 & 0.321 \\ 0.037 & 0.462 & 0.037 & 0.321 \\ 0.037 & 0.462 & 0.462 & 0.037 \end{bmatrix}$$

$$x^* = \begin{bmatrix} 0.12 & 0.33 & 0.26 & 0.29 \end{bmatrix}^{\mathrm{T}}$$

Size of the Web

- The size of M is more than 8 billion (and it is increasing)!

Why $m = 0.15$?

- Asymptotic rate of convergence of power method is *exponential* and depends on the ratio

$$| \lambda_2(M) / \lambda_1(M) |$$

- We have

$$\lambda_1(M) = 1 \qquad | \lambda_2(M) | \leq 1 - m = 0.85$$

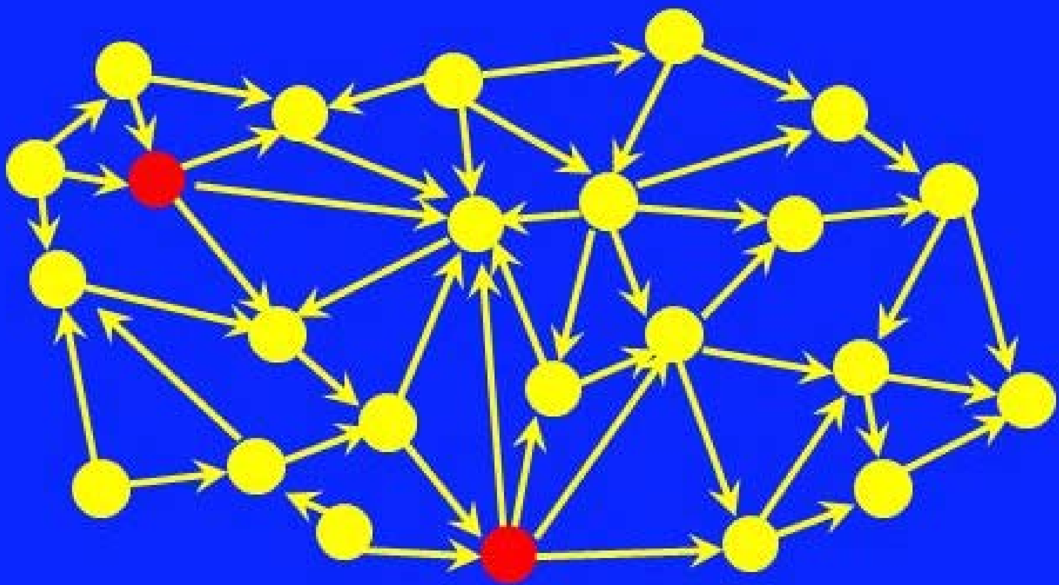
- For $N_{it} = 50$ we have $0.85^{50} \approx 2.95 \; 10^{-4}$
- For $N_{it} = 100$ we have $0.85^{100} \approx 8.74 \; 10^{-8}$
- Larger m implies faster convergence

CNR-IEIIT

Columbia River, The Dalles, Oregon

Distributed Viewpoint

- More and more computing power is needed…
- … the distributed viewpoint because *global* information about the network is difficult to obtain



Randomized Decentralized Algorithms

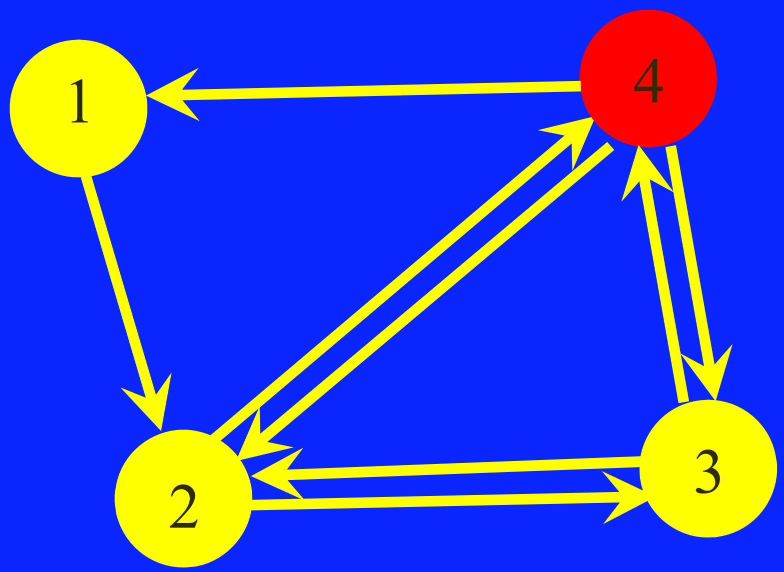
CNR-IEIIT

Randomized Decentralized Algorithms for PageRank Computation

- **Main idea:** Develop Las Vegas randomized decentralized algorithms for computing PageRank

CNR-IEIIT

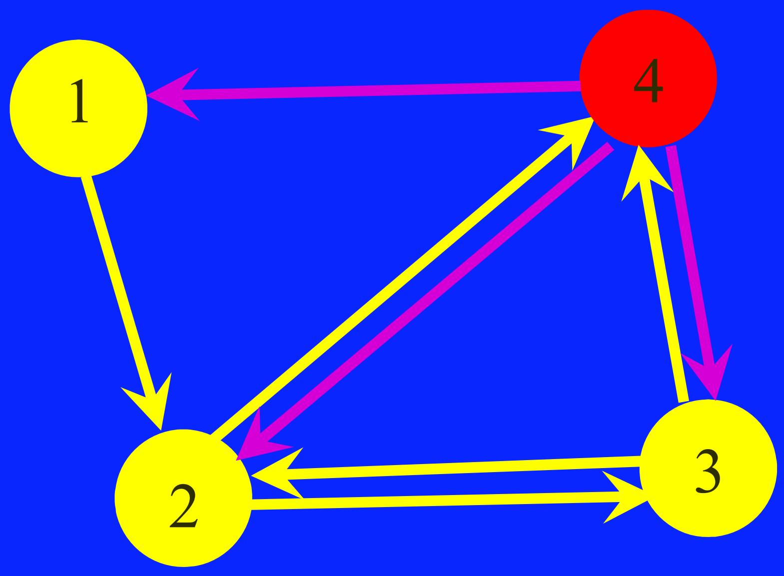
Decentralized Communication Protocol



Gossip communication protocol:

1. at time k randomly select page i (i=4) for PageRank update

Decentralized Communication Protocol



Gossip communication protocol:

1. at time k randomly select page i (i=4) for PageRank update

2. send PageRank value of page i to the outgoing pages that are linked to page i

Global Clock

- The pages taking action are determined via a stochastic process $\theta(k) \in \{1, \ldots, n\}$
- $\theta(k)$ is assumed to be i.i.d. with uniform probability

$$\text{Prob}\{\theta(k)=i\} = 1/n$$

- If at time k $\theta(k) = i$ then page i initiates PageRank update
- $\theta(k)$ plays the role of a clock, which is a global clock known to all the pages

Randomized Update Scheme

- Consider the randomized update scheme

$$x(k+1) = A_{\theta(k)}\, x(k)$$

where $A_{\theta(k)}$ are the *decentralized link matrices* related to the matrix A

- This is a Markovian jump system

Decentralized Link Matrices - 1

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

$$A_4 = \begin{bmatrix} & & & 1/3 \\ & & & 1/3 \\ & & & 1/3 \\ & & & 0 \end{bmatrix}$$

CNR-IEIIT

Decentralized Link Matrices - 2

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

$$A_4 = \begin{bmatrix} & 0 & 0 & 1/3 \\ 0 & & 0 & 1/3 \\ 0 & 0 & & 1/3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

CNR-IEIIT

Decentralized Link Matrices - 3

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

$$A_4 = \begin{bmatrix} 1 & 0 & 0 & 1/3 \\ 0 & 1 & 0 & 1/3 \\ 0 & 0 & 1 & 1/3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

CNR-IEIIT

Decentralized Link Matrices - 4

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

$$A_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1/2 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1 \end{bmatrix}$$

$$A_4 = \begin{bmatrix} 1 & 0 & 0 & 1/3 \\ 0 & 1 & 0 & 1/3 \\ 0 & 0 & 1 & 1/3 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

CNR-IEIIT

Average Matrix A_{ave}

- Average matrix $A_{\mathrm{ave}} = \mathrm{E}[\, A_{\theta(k)} \,]$
- Taking uniform distribution in the stochastic process $\theta(k)$ we have

$$A_{\mathrm{ave}} = 1/n \sum_i A_i$$

- **Lemma** (properties of the average matrix A_{ave})

  (i) $$A_{\mathrm{ave}} = 2/n\, A + (1\text{-}2/n)\, I$$

  (ii) Matrices A and A_{ave} have the same eigenvector corresponding to the eigenvalue 1

CNR-IEIIT

Modified Randomized Update Scheme

- Remark: Need to work with the teleportation model (positive stochastic matrix *M*)
- Consider the modified randomized update scheme

$$x(k+1) = M_{\theta(k)}\ x(k)$$

where $M_{\theta(k)}$ are the *modified decentralized link matrices* computed as

$$M_i = (1- r)\, A_i + r/n\, S \quad i = 1, 2, \ldots, n$$

and $r \in (0,1)$ is a design parameter

$$r = 2m/(n - mn + 2m)$$

CNR-IEIIT

Average Matrix M_{ave}

- Average matrix $M_{ave} = \mathrm{E}[\, M_{\theta(k)} \,]$
- Define $r = 2m/(n - mn + 2m)$

- **Lemma** (properties of the average matrix M_{ave})

(i) $r \in (0,1)$ and $r < m = 0.15$

(ii) $M_{ave} = r/m\, M + (1 - r/m)\, I$

(iii) Eigenvalue 1 for M_{ave} is the unique eigenvalue of maximum modulus and PageRank is the corresponding eigenvector

Averaging

- Define the time average

$$\boxed{y(k) = \frac{1}{k+1}\sum_{i=0}^{k} x(i)}$$

CNR-IEIIT

Convergence Properties

- **Theorem** (convergence properties)
- The time average $y(k)$ of the modified randomized update scheme converges to PageRank x^* in MSE

$$\boxed{E[\ \|y(k) - x^*\|^2\ ] \to 0 \quad \text{for} \ \ k \to \infty}$$

provided that $\sum_i x_i(0) = 1$

- **Proof:** Based on the theory of ergodic matrices

H. Ishii, R. Tempo (2010)

Comments

- Time average $y(k)$ can be computed recursively as a function of $y(k-1)$
- Because of averaging convergence rate is $1/k$
- Sparsity of the matrix A is preserved because

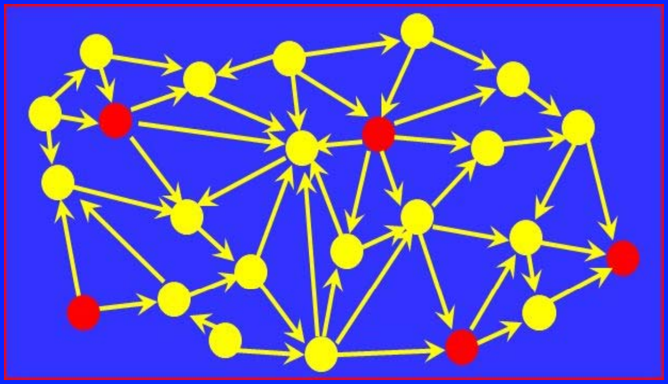
$$x(k+1) = M\,x(k) = (1-r)\,A\,x(k) + r/n\,\mathbf{1}$$

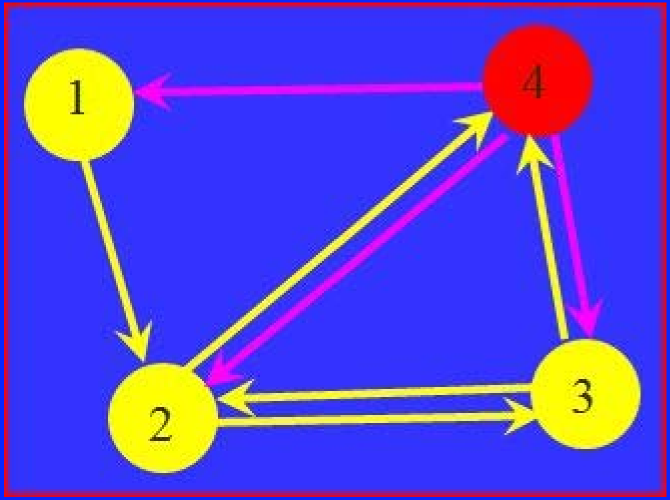
where $\mathbf{1}$ is a vector with all entries equal to one

$$A = \begin{bmatrix} 0 & 0 & 0 & 1/3 \\ 1 & 0 & 1/2 & 1/3 \\ 0 & 1/2 & 0 & 1/3 \\ 0 & 1/2 & 1/2 & 0 \end{bmatrix} \qquad M = \begin{bmatrix} 0.038 & 0.037 & 0.037 & 0.321 \\ 0.887 & 0.037 & 0.462 & 0.321 \\ 0.037 & 0.462 & 0.037 & 0.321 \\ 0.037 & 0.462 & 0.462 & 0.037 \end{bmatrix}$$

Extensions

- Stopping criteria to compute approximately PageRank
- Extensions to simultaneous update of multiple web pages
- Robustness for fragile links
- All web pages need to know the global clock θ(*k*)
- Almost-sure convergence properties of the algorithm hold





Ranking Control Journals

CNR-IEIIT

ISI Web of Knowledge

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME HELP

2010 JCR Science Edition

Journal Summary List

Journal Title Changes

Journals from: subject categories AUTOMATION & CONTROL SYSTEMS VIEW CATEGORY SUMMARY LIST

Sorted by: Journal Title SORT AGAIN

Journals 1 - 20 (of 60)

[1 | 2 | 3]

Page 1 of 3

MARK ALL UPDATE MARKED LIST

Ranking is based on your journal and sort selections.

Mark	Rank	Abbreviated Journal Title *(linked to journal information)*	ISSN	JCR Data						Eigenfactor™ Metrics	
				Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	Articles	Cited Half-life	*Eigenfactor™* Score	*Article Influence™* Score
☐	1	ANNU REV CONTROL	1367-5788	410	1.884	2.273	0.042	24	5.4	0.00160	0.750
☐	2	ASIAN J CONTROL	1561-8625	289	0.578	0.524	0.085	71	5.0	0.00131	0.218

CNR-IEIIT

Ranking Journals: Impact Factor

- Impact Factor IF

$$\textbf{IF } 2012 = \frac{\textit{number citations in 2012 of articles published in } 2010-2011}{\textit{number of articles published in } 2010-2011}$$

- Census period (2012) of one year and a window period (2010-2011) of two years
- Remark: Impact Factor is a “flat criterion” (it does not take into account where the citations come from)

CNR-IEIIT

ISI Web of Knowledge

ISI Web of Knowledge℠

Journal Citation Reports®

WELCOME HELP

2010 JCR Science Edition

Journal Summary List

Journal Title Changes

Journals from: subject categories AUTOMATION & CONTROL SYSTEMS VIEW CATEGORY SUMMARY LIST

Sorted by: Journal Title SORT AGAIN

Journals 1 - 20 (of 60) [1 | 2 | 3] **Page 1 of 3**

MARK ALL UPDATE MARKED LIST

Ranking is based on your journal and sort selections.

Mark	Rank	Abbreviated Journal Title *(linked to journal information)*	ISSN	JCR Data						Eigenfactor™ Metrics	
				Total Cites	Impact Factor	5-Year Impact Factor	Immediacy Index	Articles	Cited Half-life	*Eigenfactor™* Score	*Article Influence™* Score
☐	1	ANNU REV CONTROL	1367-5788	410	1.884	2.273	0.042	24	5.4	0.00160	0.750
☐	2	ASIAN J CONTROL	1561-8625	289	0.578	0.524	0.085	71	5.0	0.00131	0.218

CNR-IEIIT

Ranking Journals: Eigenfactor

- Eigenfactor **EF**

- Ranking journals using ideas from PageRank computation in Google
- In Eigenfactor journals are considered influential if they are cited often by other influential journals
- What is the *probability* that a journal is cited?

C. T. Bergstrom (2007)

Random Reader Model

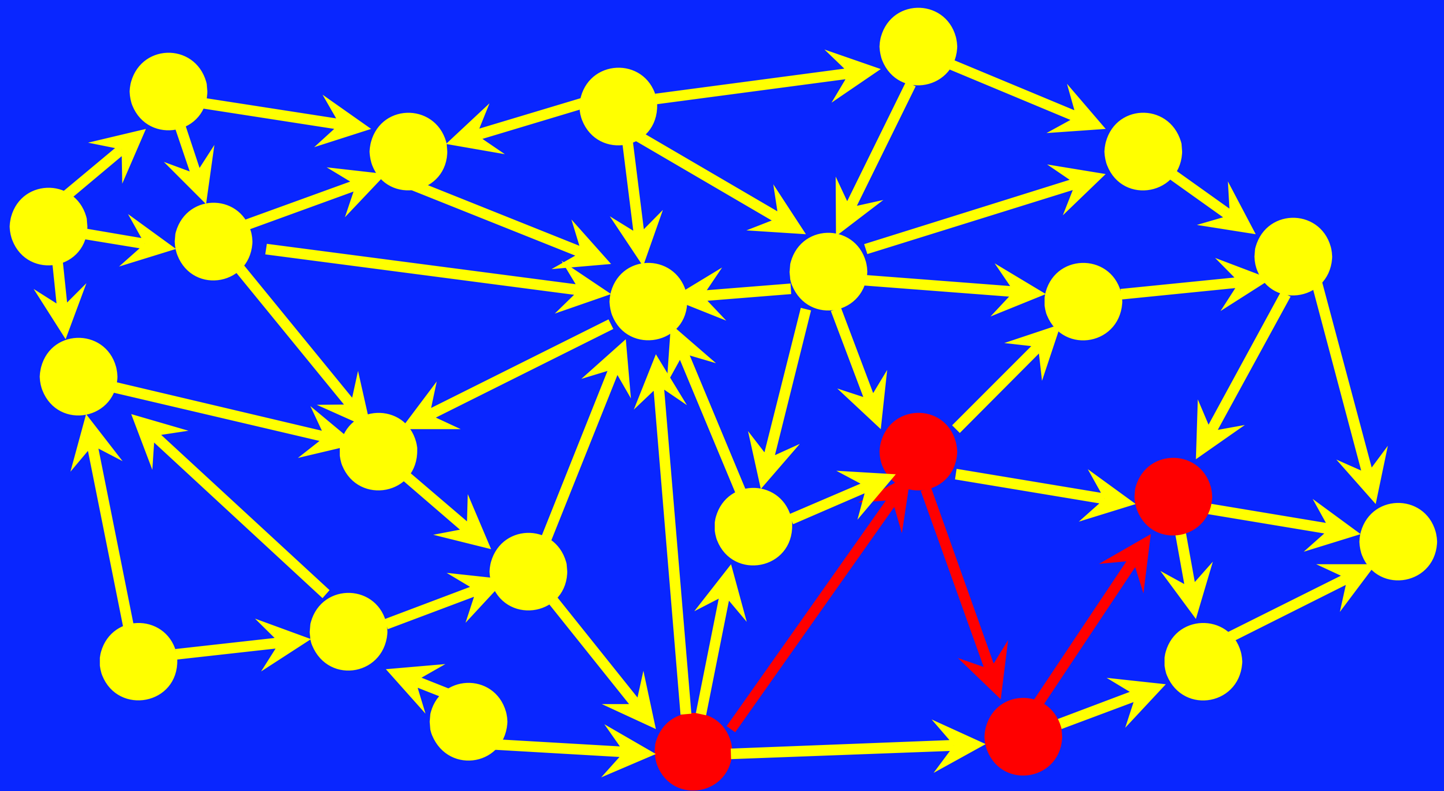
"Imagine that a researcher is to spend all eternity in the library randomly following citations within scientific periodicals. The researcher begins by picking a random journal in the library. From this volume a random citation is selected. The researcher then walks over to the journal referenced by this citation. The probability of doing so corresponds to the fraction of citations linking the referenced journal to the referring one, conditioned on the fact that the researcher starts from the referring journal. From this new volume the researcher now selects another random citation and proceeds to that journal. This process is repeated ad infinitum."

A. D. West, T. C. Bergstrom, C. T. Bergstrom (2010)

Random Reader Model: Graph Representation

CNR-IEIIT

❖ Directed graph with journals (nodes) and citations (links)



Cross-citation Matrix - 1

$$a_{ij} = number\ citations\ in\ 2012\ from\ journal\ \ j\ to\ articles\ published\ in\ journal\ \ i\ in\ 2007 - 2011$$

- Self-citations are omitted

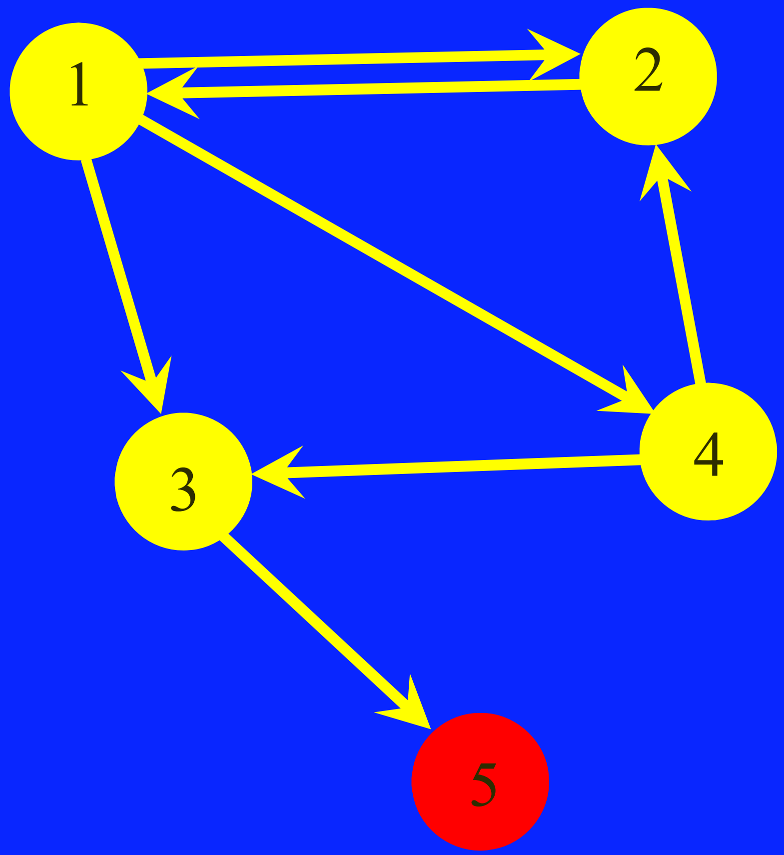
$$A = \begin{bmatrix} 0 & \cdots & \\ \vdots & 0 & \vdots \\ & \cdots & 0 \end{bmatrix}$$

- Normalization to obtain a column substochastic matrix

$$a_{ij} = \frac{a_{ij}}{\sum_k a_{kj}}$$

(note: if 0/0 we set $a_{ij}= 0$)

Black Holes



- Black holes: journals that do not cite any other journal
- Columns with entries equal to zero
- A substochastic matrix

$$A = \begin{bmatrix} \vdots & \boxed{0} & \ddots & \vdots \\ \vdots & 0 & \ddots & \vdots \\ \vdots & 0 & \ddots & \vdots \\ \vdots & 0 & \ddots & \vdots \end{bmatrix}$$

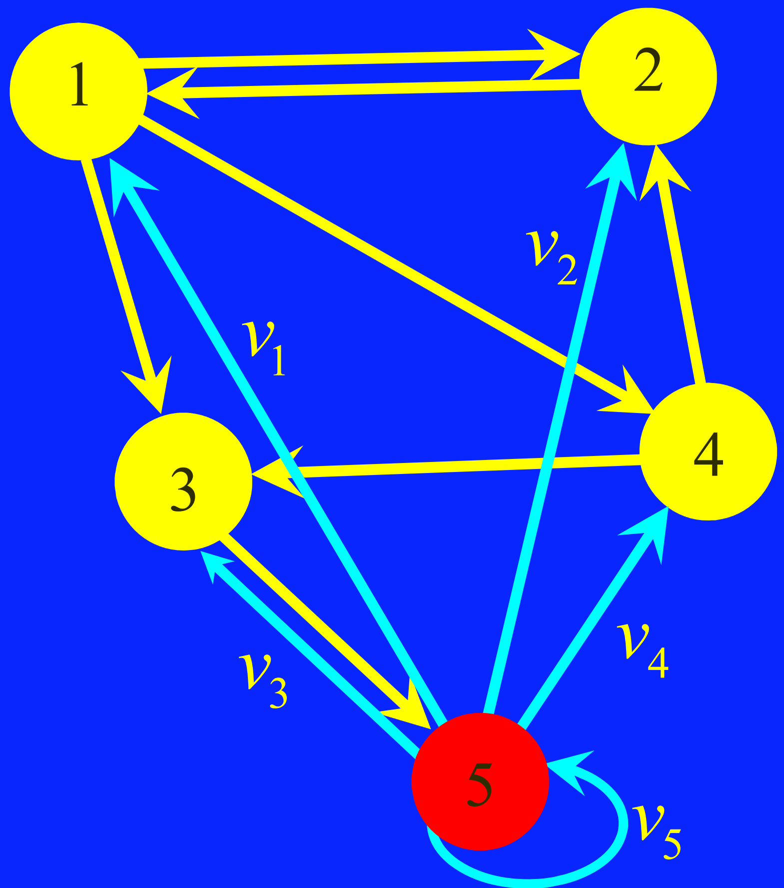
Article Vector

- Article vector v

$$v_i = \frac{number\ articles\ published\ by\ journal\ i\ in\ 2007 - 2011}{number\ articles\ published\ by\ all\ journals\ in\ 2007 - 2011}$$

- v_i is the fraction of all published articles coming from journal i during the window period 2007-2011
- Article vector v is a stochastic vector
- Total number of journals n = 8336

Cross-citation Matrix – 2



- Replace A with stochastic matrix $\tilde{A}$ introducing the article vector v

$$A = \begin{bmatrix} * & * & * & * & \boxed{0} \\ * & * & * & * & \boxed{0} \\ * & * & * & * & \boxed{0} \\ * & * & * & * & \boxed{0} \\ * & * & * & * & \boxed{0} \end{bmatrix}$$

$$\tilde{A} = \begin{bmatrix} * & * & * & * & \boxed{v_1} \\ * & * & * & * & \boxed{v_2} \\ * & * & * & * & \boxed{v_3} \\ * & * & * & * & \boxed{v_4} \\ * & * & * & * & \boxed{v_5} \end{bmatrix}$$

CNR-IEIIT

Cross-citation Matrix - 2

❖ Replace substochastic matrix A with stochastic matrix $\tilde{A}$ introducing the article vector v

$$A = \begin{bmatrix} * & * & * & * & 0 \\ * & * & * & * & 0 \\ * & * & * & * & 0 \\ * & * & * & * & 0 \\ * & * & * & * & 0 \end{bmatrix} \Rightarrow \tilde{A} = \begin{bmatrix} * & * & * & * & v_1 \\ * & * & * & * & v_2 \\ * & * & * & * & v_3 \\ * & * & * & * & v_4 \\ * & * & * & * & v_5 \end{bmatrix}$$

Random Reader Model and Markov Chains

CNR-IEIIT

- Rank $n = 8336$ journals in order of importance
- Random reader model represented as a Markov chain

$$x(k+1) = \tilde{A}x(k)$$

where $x(k) \in [0,1]^n$ and $\sum_i x_i(k) = 1$ is the journal influence vector at step k

- $x_i(k)$ represents the importance of the journal i

- Question: Does the Markov chain converge to a stationary value representing the probability of citation?

CNR-IEIIT

Convergence of the Markov Chain - 1

- Answer: No
- Need to replace $\tilde{A}$ with the matrix M obtaining the Eigenfactor equation

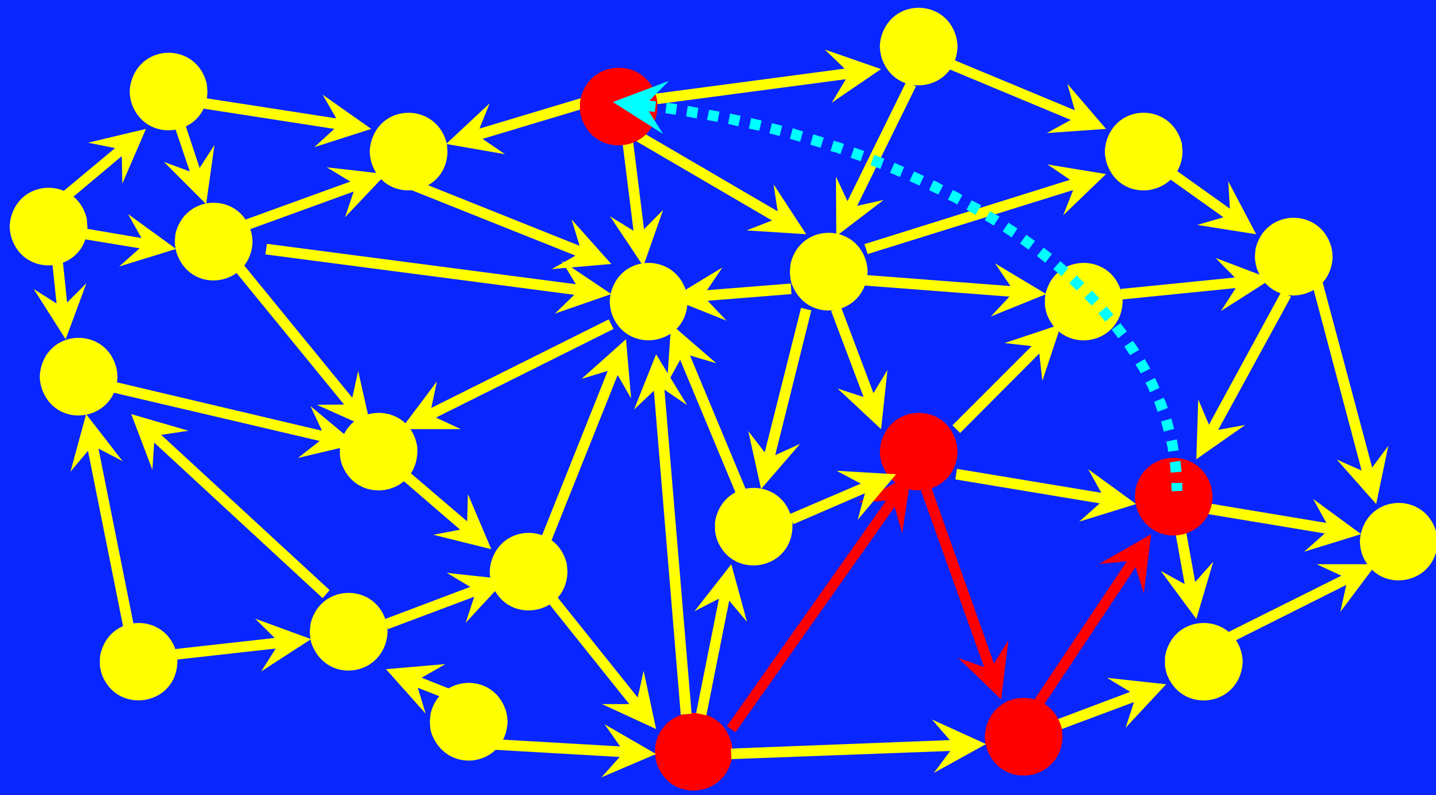
$$M = (1 - m)\tilde{A} + m\, v\mathbf{1}^{\mathrm{T}} \qquad m = 0.15$$

where $v\,\mathbf{1}^{\mathrm{T}} = \begin{bmatrix} v_1 & \cdots & v_1 \\ \vdots & \ddots & \vdots \\ v_n & \cdots & v_n \end{bmatrix}$ is a weighting matrix

CNR-IEIIT

Teleportation Model

- This avoids being trapped in a small set of journals



Convergence of the Markov Chain - 2

- Consider the Markov chain

$$x(k+1) = M\, x(k)$$

- If $\sum_i x_i(0) = 1$ convergence is guaranteed by Perron Theorem because M is a strictly positive matrix

$$x(k) \rightarrow x^* \quad \text{for} \quad k \rightarrow \infty$$

- The corresponding graph is strongly connected
- We have $\qquad x^* = M\, x^*$
- Journal influence vector x^* exists and it is unique
- Eigenvector corresponding to the simple eigenvalue 1

CNR-IEIIT

PageRank Iteration

- To preserve sparsity of A we write the journal influence vector iteration using power method

$$x(k+1) = (1-m)Ax(k) + [(1-m)b^{\mathrm{T}}x(k)+m]v$$

where b is the black holes vector having 1 for entries corresponding to black holes and 0 elsewhere

$$A = \begin{bmatrix} \vdots & \boxed{\begin{matrix} 0 \\ 0 \\ 0 \\ 0 \end{matrix}} & \ddots & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & & \ddots & \vdots \\ \vdots & & \ddots & \vdots \end{bmatrix} \qquad b = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

Eigenfactor EF

- Journal influence vector x^* is the steady-state fraction of time spent reading each journal
- x^* provides weights to the citation values
- Eigenfactor $\mathbf{EF}_j$ of journal j is the percentage of the total weighted citations that journal j receives from all 8336 journals

$$\mathbf{EF}_j = \frac{100\,(Ax^*)_j}{\sum_i (Ax^*)_i}$$

- Article influence score $\mathbf{AI}_j$

$$\mathbf{AI}_j = \frac{0.01\,\mathbf{EF}_j}{v_j}$$

2012 IMPACT FACTOR (IF) AND 2012 EIGENFACTOR SCORE (EF)

IF	Journal	Ranking	Journal	EF
2.919	Automatica	1	IEEE Transactions Automatic Control	0.04492
2.718	IEEE Transactions Automatic Control	2	Automatica	0.04478
2.372	IEEE Control Systems Magazine	3	SIAM J. Control & Optimization	0.01479
2.000	IEEE Transactions Contr. Sys. Tech.	4	IEEE Transactions Contr. Sys. Tech.	0.01196
1.900	Int. J. Robust Nonlinear Control	5	Systems & Control Letters	0.01087
1.805	Journal Process Control	6	Int. Journal Control	0.00859
1.669	Control Engineering Practice	7	Int. J. Robust Nonlinear Control	0.00854
1.667	Systems & Control Letters	8	Control Engineering Practice	0.00696
1.379	SIAM J. Control & Optimization	9	Journal Process Control	0.00622
1.250	European J. Control	10	IEEE Control Systems Magazine	0.00554

Consensus and PageRank

Distributed Consensus

- **Problem:** Consider a set of n agents each having a numerical value transmitted to the neighboring agents iteratively
- Different approaches depending on the range of node values: Real, integer (quantized) and binary values
- **Consensus:** All agents eventually reach a common value

A. Jadbabaie, J. Lin, A. S. Morse (2003)

CNR-IEIIT

Graph of Agents

- Consider a graph of agents which communicate using a *decentralized communication pattern* (similar to that for PageRank)
- The value of agent i at time k is x_i
- Values of agents may be updated using a LVRA

$$x(k+1) = A_{\theta(k)}\ x(k)$$

CNR-IEIIT

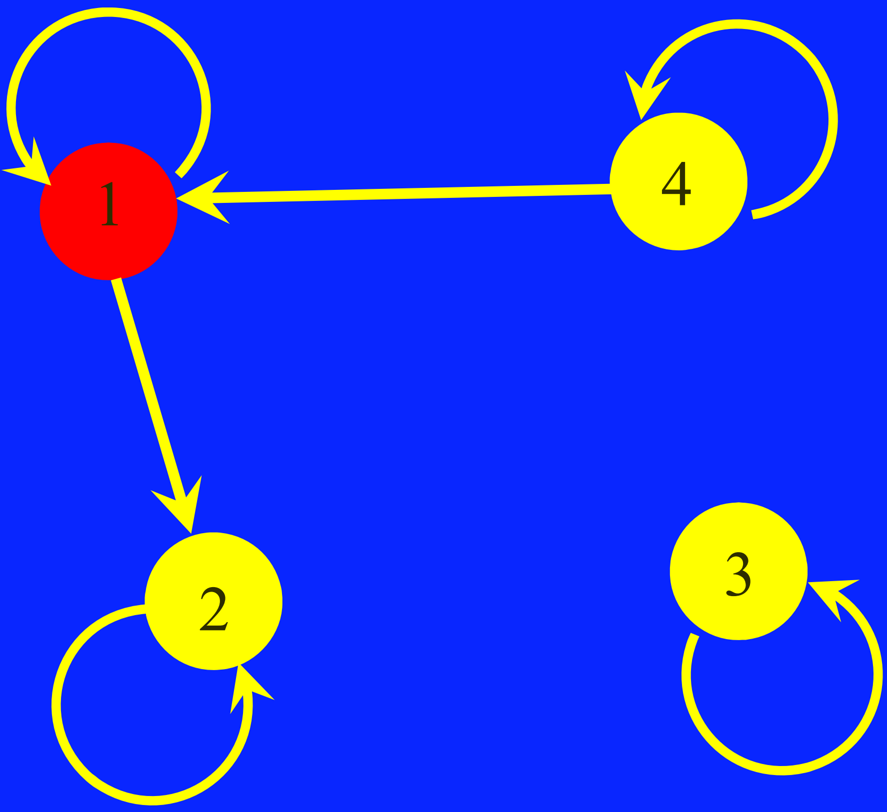
Decentralized Communication Pattern

- Stochastic process $\theta(k) \in \{1, \ldots, d\}$ where d is the number of communication patterns
- $\theta(k)$ is assumed to be i.i.d. with uniform probability

$$\text{Prob}\{\theta(k)=i\} = 1/d$$

- If at time k $\theta(k) = i$ then agent i initiates update
- Technical differences with communication protocol for PageRank but the ideas are similar

Communication Pattern for Agent 1



$$A_1 = \begin{bmatrix} * & 0 & 0 & * \\ * & * & 0 & 0 \\ 0 & 0 & * & 0 \\ 0 & 0 & 0 & * \end{bmatrix}$$

$$A_1 = \begin{bmatrix} 1/2 & 0 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

A_1 is row stochastic

CNR-IEIIT

Consensus

- We say that consensus w.p.1 is achieved if for any initial condition $x(0)$ we have

$$\text{Prob}\left\{\lim_{k\to\infty}\left|x_i(k)-x_j(k)\right|=0\right\}=1$$

for all agents i, j

CNR-IEIIT

Convergence Properties

- **Lemma** (convergence properties)
- Assuming that the graph is strongly connected, the scheme

$$x(k+1) = A_{\theta(k)}\ x(k)$$

achieves consensus w.p.1

$$\mathrm{Prob}\left\{\lim_{k\to\infty}\left|x_i(k)-x_j(k)\right|=0\right\}=1$$

for all agents i, j and for any initial condition $x(0)$

Y. Hatano, M. Mesbahi (2005); C. W. Wu (2006)
A. Tahbaz-Salehi, A. Jadbabaie (2008)

CNR-IEIIT

PageRank and Consensus

Consensus	PageRank
all agent values become equal	page values converge to constant
strongly connected agent graph	web is not strongly connected
A_i are row stochastic matrices	A_i and M_i are column stochastic
self-loops in the agent graph	no self-loops in the web
convergence for any $x(0)$	$x(0)$ stochastic vector
time averaging $y(k)$ not necessary	averaging $y(k)$ required

Economic Complexity

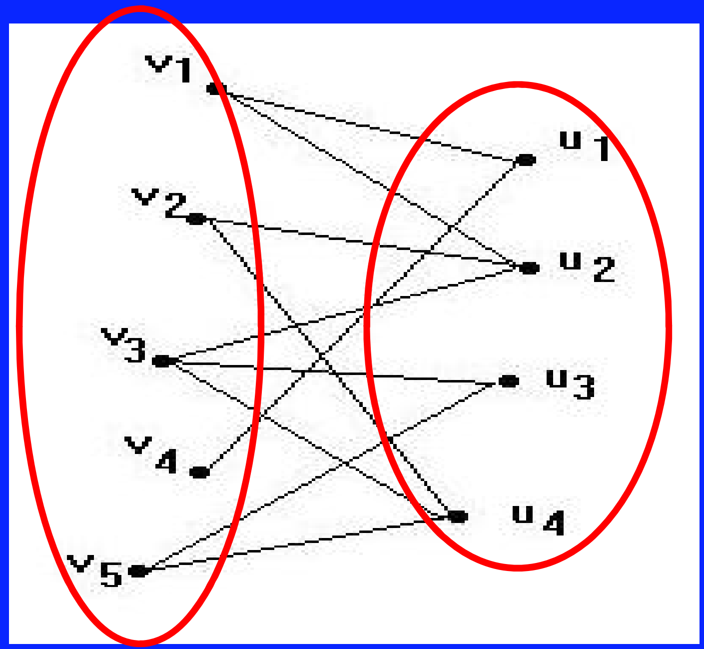
CNR-IEIIT

Bipartite Graphs - 1

- Bipartite graph representing countries and products

countries v

$x_v=\{v1,v2,v3,v4,v5\}$



products

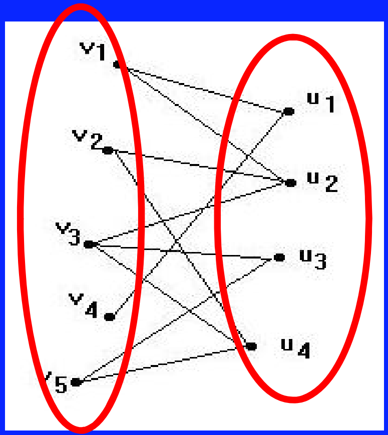
$x_u=\{u1,u2,u3,u4\}$

Bipartite Graphs - 2

- Bipartite graph representing countries and products

countries v

$x_v=\{v1,v2,v3,v4,v5\}$



products

$x_u=\{u1,u2,u3,u4\}$

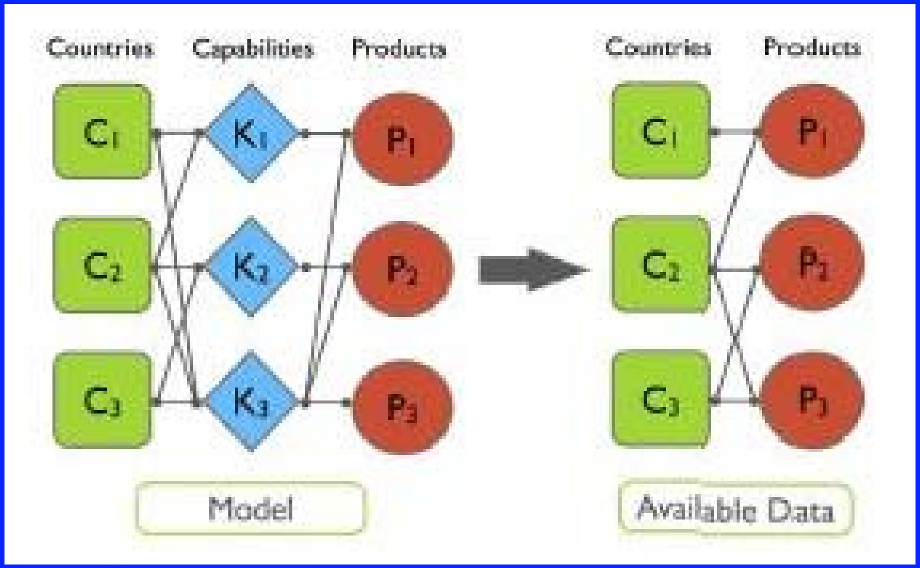
- Construct a binary cross-product matrix M
- Compute PageRank x_v* and x_u* representing the fitness of countries and the quality of products

CNR-IEIIT

Economic Complexity

❖ Tripartite graph representing countries, capabilities and products

theory of hidden capabilities



This leads to a nonlinear system $x(k+1) = f(x(k))$!

No guarantee of convergence to PageRank x^*!

Distributed Algorithms for Web Aggregation

CNR-IEIIT

Aggregation, PageRank and Complexity

- Aggregation is a tool to break complexity

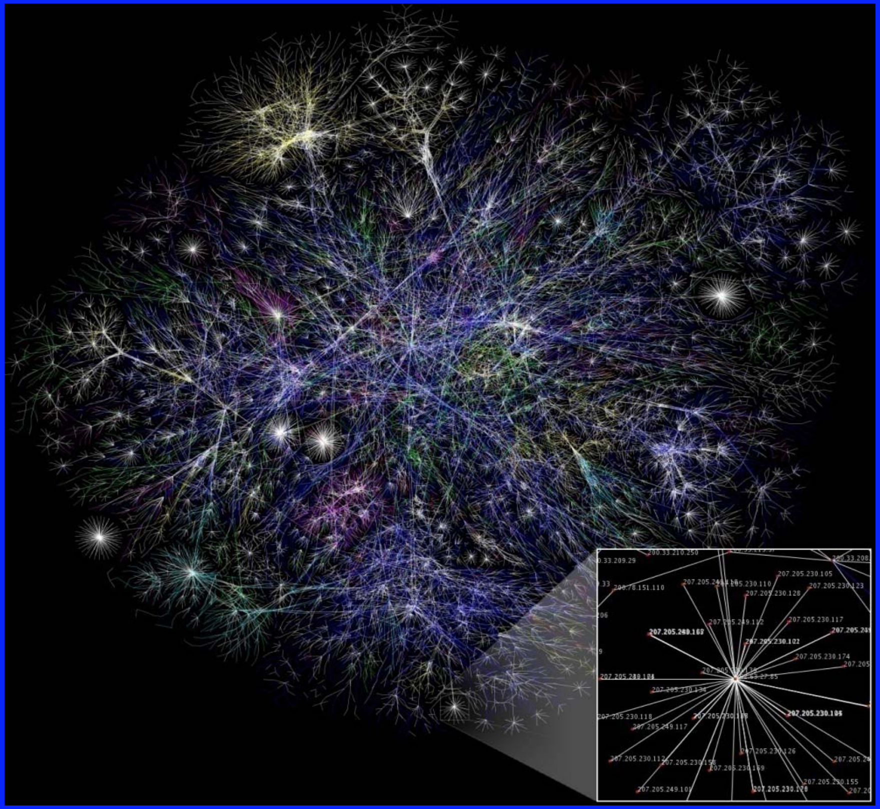
- Applications:

    - web
    - power grid
    - biology
    - social networks
    - journals
    - e-commerce

CNR-IEIIT

PageRank and Web Aggregation

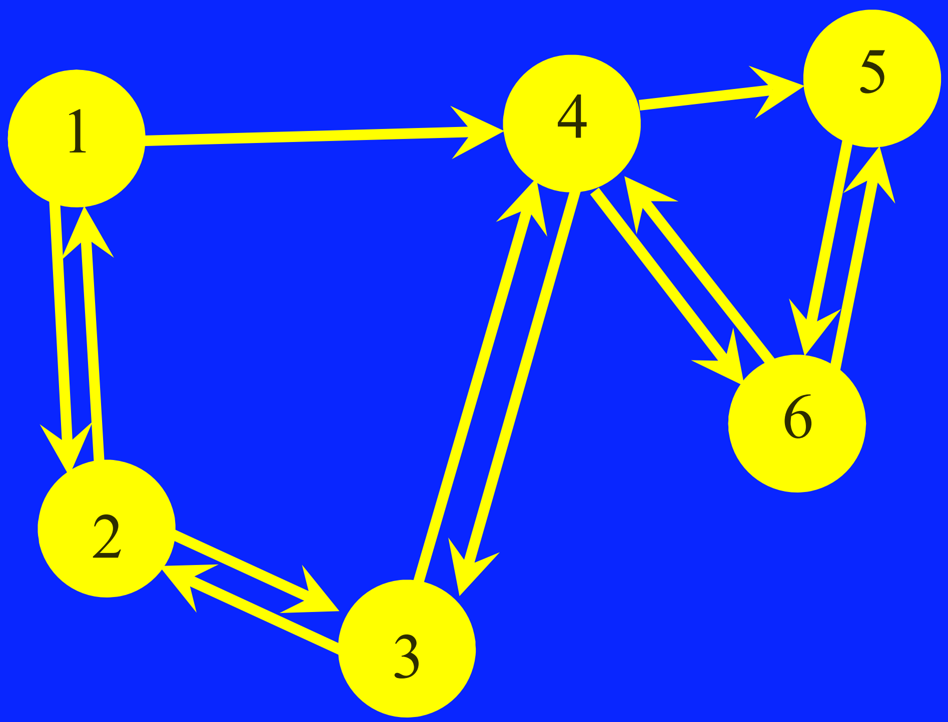
- View the web as a network of agents connected via links
- Web aggregation
- Approximate computation of PageRank to reduce computation cost and communication at servers
- Guaranteed bounds on the approximation error



H. Ishii, R. Tempo, E.-W. Bai (2012)

CNR-IEIIT

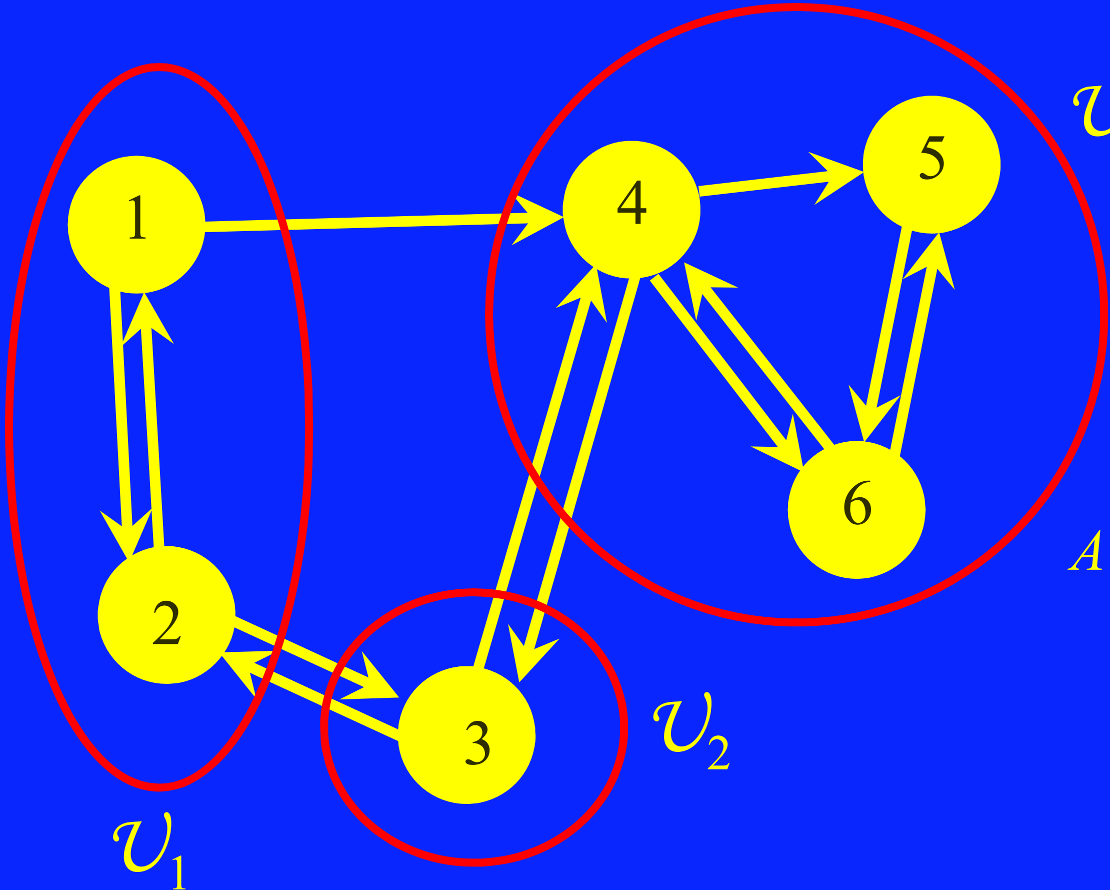
Graph Representation



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

$$x^* = \begin{bmatrix} 0.0614 & 0.0857 & 0.122 & 0.214 & 0.214 & 0.302 \end{bmatrix}^T$$

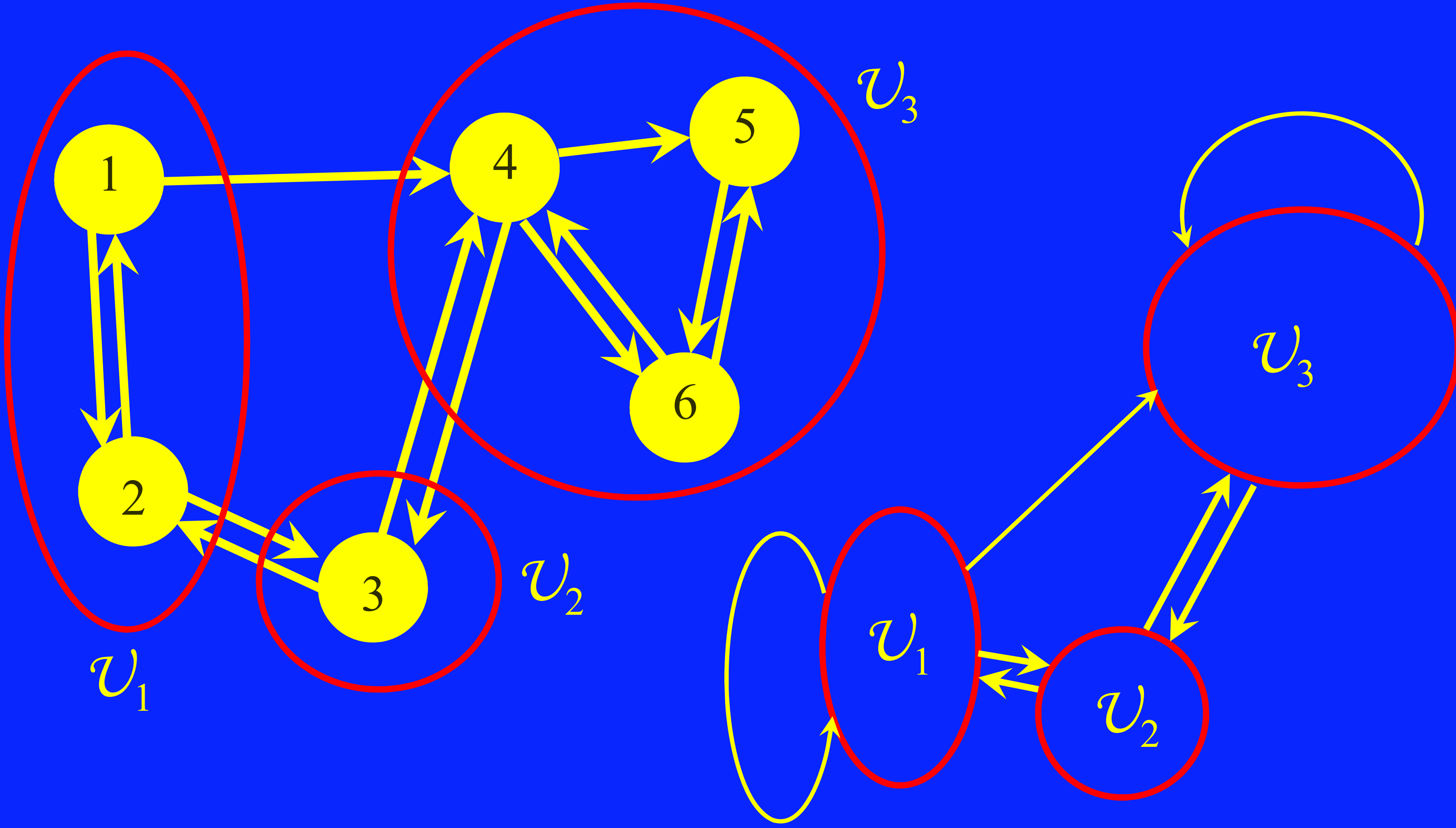
Web Aggregation



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

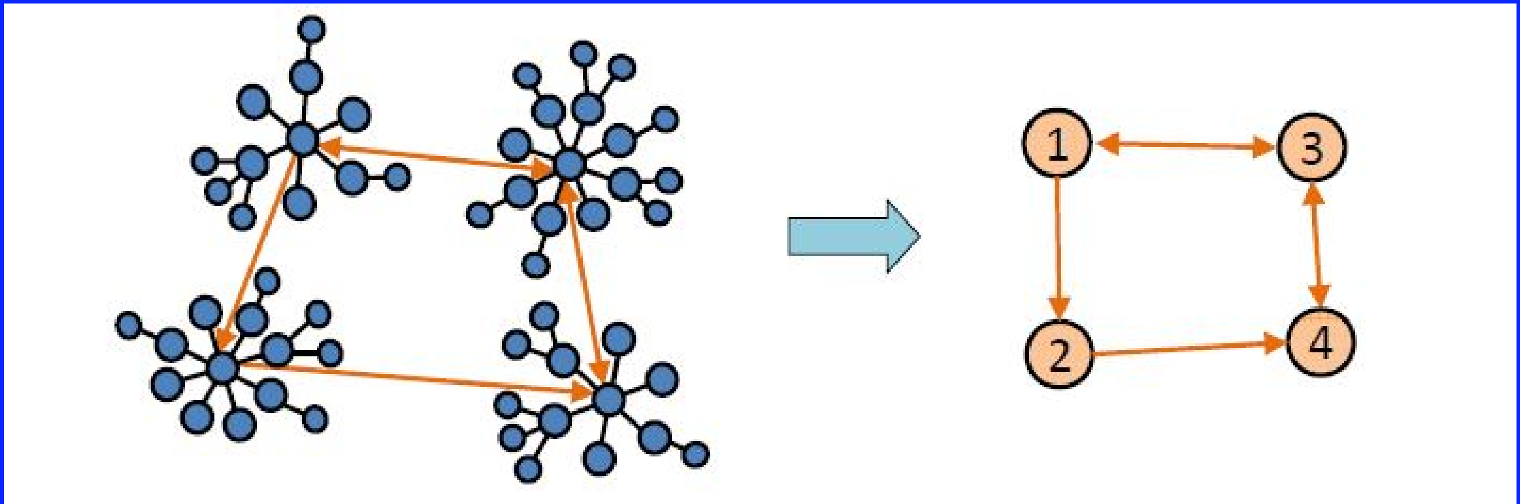
$$x^* = \begin{bmatrix} 0.0614 & 0.0857 & 0.122 & 0.214 & 0.214 & 0.302 \end{bmatrix}^T$$

Aggregation of Agents



CNR-IEIIT

Web Aggregation

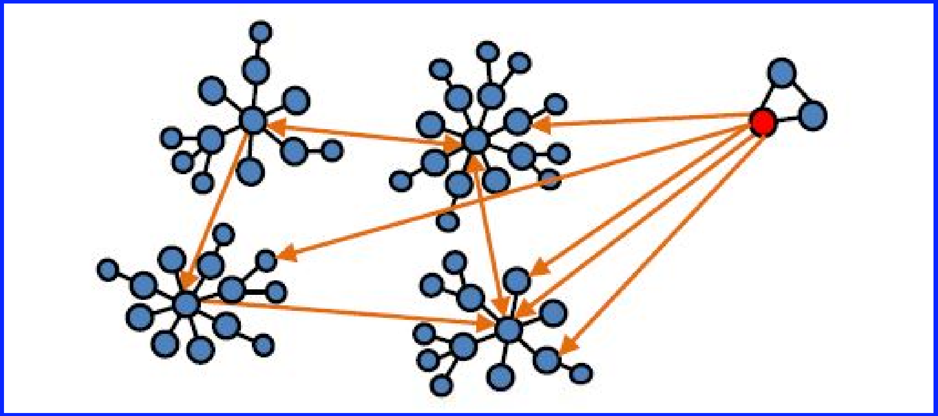


- Most links are intra-host ones
- Two-step approach:
  1. Global step: Compute total PageRank for each group
  2. Local step: Distribute value among group members

CNR-IEIIT

Web Aggregation

- Singular perturbation methods for PageRank computation



$$\frac{\text{number external outlinks}}{\text{number outlinks}} \leq \delta$$

- Web pages having many external outlinks are considered as single groups

J.H. Chow, P.V. Kokotovic (1985); E. Biyik, M. Arcak (2008)

CNR-IEIIT

Coordinate Change for PageRank

$$\tilde{x}^* = Vx^* \qquad \tilde{x}^* = \begin{bmatrix} \tilde{x}_1^* \\ \tilde{x}_2^* \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} x^*$$

where $\tilde{x}_1^*$ is the aggregated PageRank (based on group value), V_1 and V_2 are block diagonal matrices (after reordering)

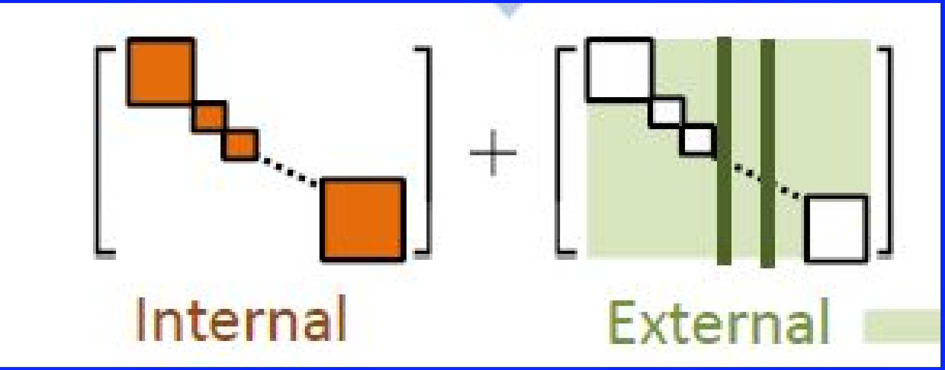
- Objective: design a distributed randomized algorithm for computing approximately PageRank x^* based on the computation of $\tilde{x}_1^*$

PageRank Equation: Internal/External Communications

CNR-IEIIT

- PageRank equation

$$x^* = (1 - m) A x^* + m/n \, \mathbf{1}$$



internal communications in the same group

external communications inter-groups

- Internal: block diag stochastic
- External: block diag norm bounded by 2δ

CNR-IEIIT

Approximate PageRank Equation

- New PageRank equation

$$\begin{bmatrix} \tilde{x}_1^* \\ \tilde{x}_2^* \end{bmatrix} = (1-m)\begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}\begin{bmatrix} \tilde{x}_1^* \\ \tilde{x}_2^* \end{bmatrix} + \frac{m}{n}\begin{bmatrix} u \\ 0 \end{bmatrix}$$

where $u = V_1 \mathbf{1}$

- $\tilde{x}^{\mathrm{A}}$ is approximate PageRank

$$\begin{bmatrix} \tilde{x}_1^{\mathrm{A}} \\ \tilde{x}_2^{\mathrm{A}} \end{bmatrix} = (1-m)\begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22}^{\mathrm{I}} \end{bmatrix}\begin{bmatrix} \tilde{x}_1^{\mathrm{A}} \\ \tilde{x}_2^{\mathrm{A}} \end{bmatrix} + \frac{m}{n}\begin{bmatrix} u \\ 0 \end{bmatrix}$$

CNR-IEIIT

Aggregated PageRank Algorithm - 1

1. Group values: reduced order computation

$$\tilde{x}_1^{\mathrm{A}}(k+1) = (1-m)\tilde{A}_{11}\tilde{x}_1^{\mathrm{A}}(k) + \frac{m}{n}\begin{bmatrix} u \\ 0 \end{bmatrix}$$

$\tilde{A}_{11}\tilde{x}_1^{\mathrm{A}}(k)$
communications
inter-groups

2. Iteration

$$\tilde{x}_2^{\mathrm{A}}(k+1) = \tilde{x}_2^{\mathrm{A}}(k) = (1-m)[I\text{-}(1-m)\tilde{A}_{22}^{\mathrm{A}}]^{-1}\tilde{A}_{21}\tilde{x}_1^{\mathrm{A}}(k)$$

$\tilde{A}_{22}^{\mathrm{A}}$
communications
in the same group

$\tilde{A}_{21}\tilde{x}_1^{\mathrm{A}}(k)$
communications
inter-groups

CNR-IEIIT

Aggregated PageRank Algorithm - 2

3. Inverse state transformation

$$x^{A}(k) = \boxed{V^{-1}}\tilde{x}^{A}(k)$$

V^{-1}
block diagonal matrix

- Theorem (convergence properties):

$x^{A}(k) \rightarrow x^{A}$ for $k \rightarrow \infty$ in MSE sense provided that

$x^{A}(0)$ is a stochastic vector

CNR-IEIIT

Approximation Error

$$\begin{bmatrix} \tilde{x}_1^{\mathrm{A}} \\ \tilde{x}_2^{\mathrm{A}} \end{bmatrix} = (1-m) \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{21} & \tilde{A}_{22}^{\mathrm{I}} \end{bmatrix} \begin{bmatrix} \tilde{x}_1^{\mathrm{A}} \\ \tilde{x}_2^{\mathrm{A}} \end{bmatrix} + \frac{m}{n} \begin{bmatrix} u \\ 0 \end{bmatrix}$$

- Approximate PageRank $x^{\mathrm{A}} = V^{-1}\tilde{x}^{\mathrm{A}}$
- Theorem (approximation error):

  Given ε > 0, take

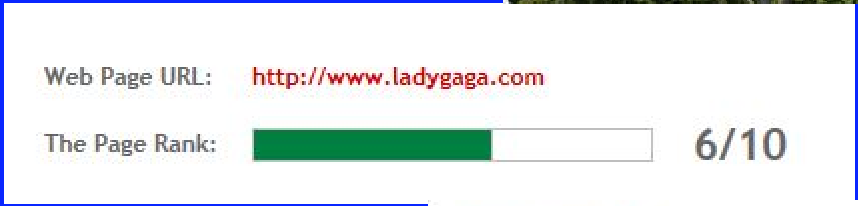
$$\frac{\text{number external outlinks}}{\text{number outlinks}} \leq \delta \leq \frac{4\varepsilon}{4(1-m)(1+\varepsilon)}$$

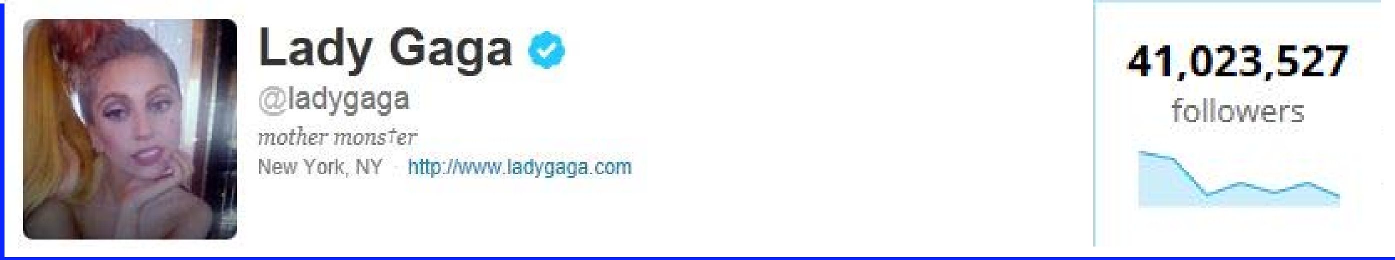
then

$$\| x^* - x^{\mathrm{A}} \|_1 \leq \varepsilon$$

CNR-IEIIT

PageRank and Twitter Followers





References

CNR-IEIIT

PageRank Main References

- H. Ishii and R. Tempo, "Distributed Randomized Algorithms for the PageRank Computation," IEEE TAC, 2010
- H. Ishii, R. Tempo and E. W. Bai, "A Web Aggregation Approach for Distributed Randomized PageRank Algorithms," IEEE TAC, 2012
- H. Ishii, R. Tempo and E. W. Bai, "PageRank Computation via a Distributed Randomized Approach with Lossy Communication," Systems & Control Letters, 2012
- H. Ishii and R. Tempo, "The PageRank Problem, Multi-Agent Consensus and Web Aggregation,," IEEE CSM, 2014

CNR-IEIIT

PageRank (Other References)

- O. Fercoq, M. Akian, M. Bouhtou, S. Gaubert, “PageRank, Markov Chains and Ergodic Control,” IEEE TAC, 2013
- W.-X. Zhao, H.-F. Chen, H.T. Fang, “Almost-Sure Convergence of the Randomized Scheme,” IEEE TAC, 2013
- A.V. Nazin and B.T. Polyak. “Randomized Algorithm to Determine the Eigenvector of a Stochastic Matrix,” Automation and Remote Control, 2011

CNR-IEIIT

Eigenfactor References

- J.D. West, T.C. Bergstrom, C.T. Bergstrom, "The Eigenfactor Metrics: A Network Approach to Assessing Scholarly Journals," College of Research Libraries, 2010
- C.T. Bergstrom, "Eigenfactor: Measuring the Value and Prestige of Scholarly Journals," C&RL News, 2007
- L. Leydesdorff, "How are New Citation-Based Journal Indicators Adding to the Bibliometric Toolbox?," Journal of the American Society for Information Science and Technology, 2008

Conclusions (robustness)

Uncertainty

"The use of equalizing structures to compensate for the variation in the phase and attenuation characteristics of transmission lines and other pieces of apparatus is well known in the communication art… the characteristics demanded of the equalizer cannot be prescribed in advance, either because… are not known with sufficient precision, or because they vary with time… transmission lines the exact lengths of which are unknown, or the characteristics of which may be affected by changes in temperature and humidity…. and since the daily cycle of temperature changes may be large…"

Variable Equalizers

- The quote is taken from the paper titled “Variable Equalizers” by Hendrik W. Bode published in 1938 in *Bell System Technical Journal*

- Bode fully recognized the importance to control a system subject to uncertainty

CNR-IEIIT

Automatica 50th Anniversary

- I.R. Petersen and R. Tempo, “Robust Control of Uncertain Systems: Classical Results and Recent Developments,” Automatica, May 2014

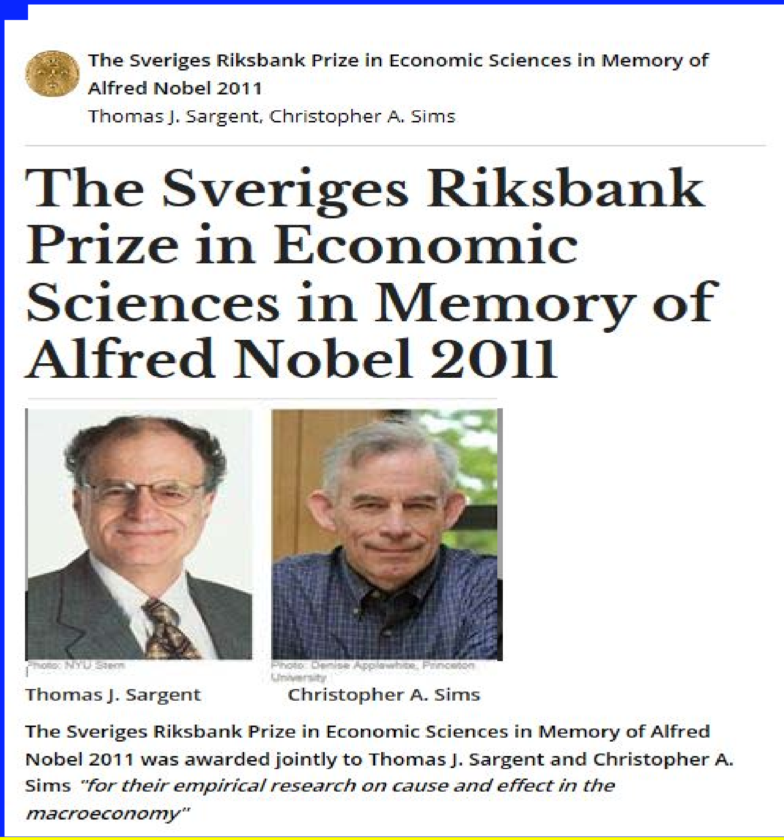
Successes of Robustness

Keywords: *Robust economics*

CNR-IEIIT

Robust Economics

- Thomas J. Sargent Nobel Prize in Economics in 2011

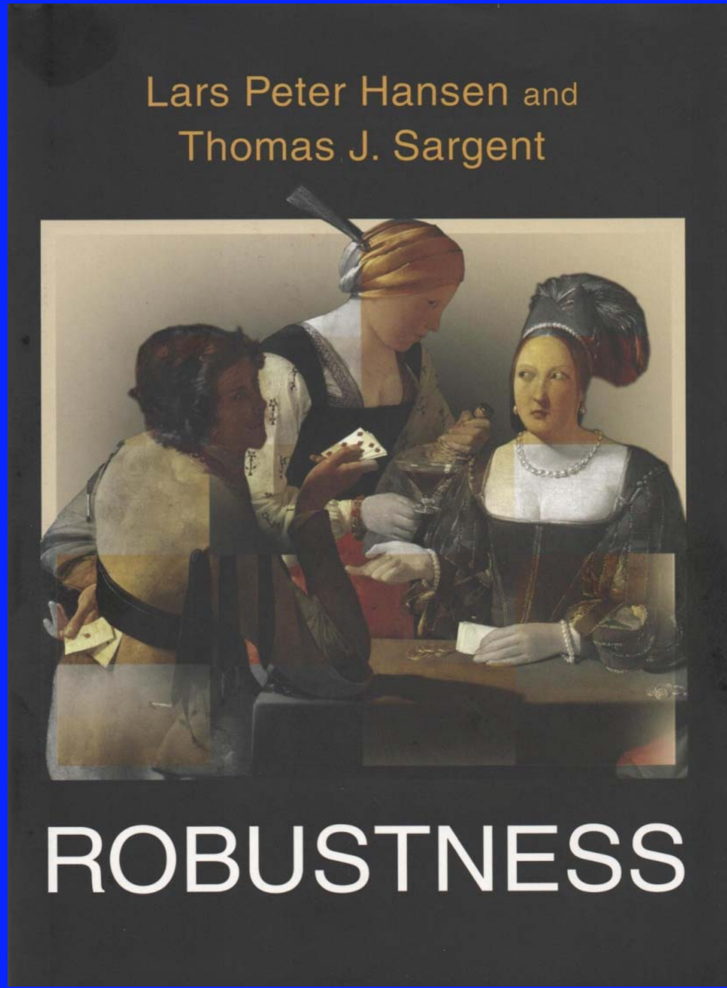

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2011
Thomas J. Sargent, Christopher A. Sims

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2011

Photo: NYU Stern
Photo: Denise Applewhite, Princeton University

Thomas J. Sargent | Christopher A. Sims

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2011 was awarded jointly to Thomas J. Sargent and Christopher A. Sims *"for their empirical research on cause and effect in the macroeconomy"*



CNR-IEIIT

Robust Economics

- Lars Peter Hansen Nobel Prize in Economics in 2013

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2013
Eugene F. Fama, Lars Peter Hansen, Robert J. Shiller

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2013

Photo: A. Mahmoud
Eugene F. Fama

Photo: A. Mahmoud
Lars Peter Hansen

Photo: A. Mahmoud
Robert J. Shiller

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2013 was awarded jointly to Eugene F. Fama, Lars Peter Hansen and Robert J. Shiller *"for their empirical analysis of asset prices"*.