UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES
Laboratoire d'informatique CNRS - UMR 8144 PRiSM
MAISON DE LA SIMULATION

Big data, Page Ranking and Application

Nahid Emad

Maison de la Simulation / PR*i*SM laboratory
University of Versailles

ORAP Forum

April 10, 2014

With contribution of

- *S. Ben Amor (PRiSM/UVSQ)*
- *A. Bui (PRiSM/UVSQ)*
- *Michel Lamure (University of Lyon 1, France)*
- *Z. Liu, PhD Student (MDLS, PRiSM/UVSQ)*
- *J.-M. Batto (INRA/MGP)*
- *N.A. Gaye, PhD Student (INRA/MGP, PRiSM/UVSQ)*

In collaboration with the *Pharmaco-Epidemiology and Infectious Diseases* laboratory of UVSQ and the *Pasteur Institute*.

Outline

1. Big Data & HPC
2. PageRank approach
3. Epidemic modeling
4. Computational algorithms and experiments
5. Concluding remarks

Big Data & HPC

Some characteristics:

- Telescopic scale rather than microscopic;
- The possibility to do things on a large scale that can not be done at small scale ;
- Once the data used, they are not outdated;
- Making “talk" data by focusing on ***what*** rather than ***why*** ;

To study the huge amounts of data, new methods/tools/ models/… are needed.

✓ Technique “out of core” of *Google MapReduce* which has been widely used in parallel computing .

Big Data & HPC

The heart of big data is the ***prediction*** :

apply mathematics to big data to derive probabilities.

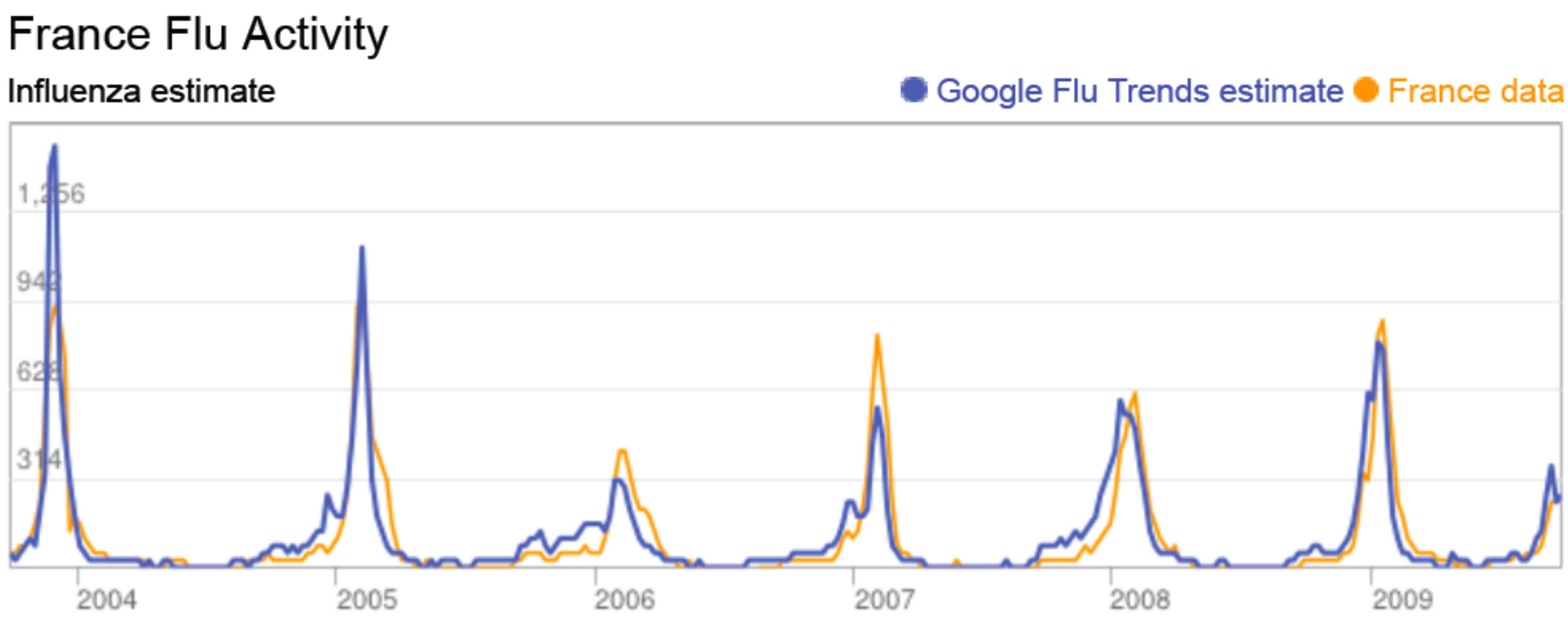
Methodology: search of ***correlation***

- Spam email detection
- Correct spelling of a word detection
- Automatic translation
- …

The research in HPC and more particularly in *Exascale Computing*, is more than ever necessary

Big Data: flu epidemic

Aggregation of Google search data to estimate current flu activity in near real-time

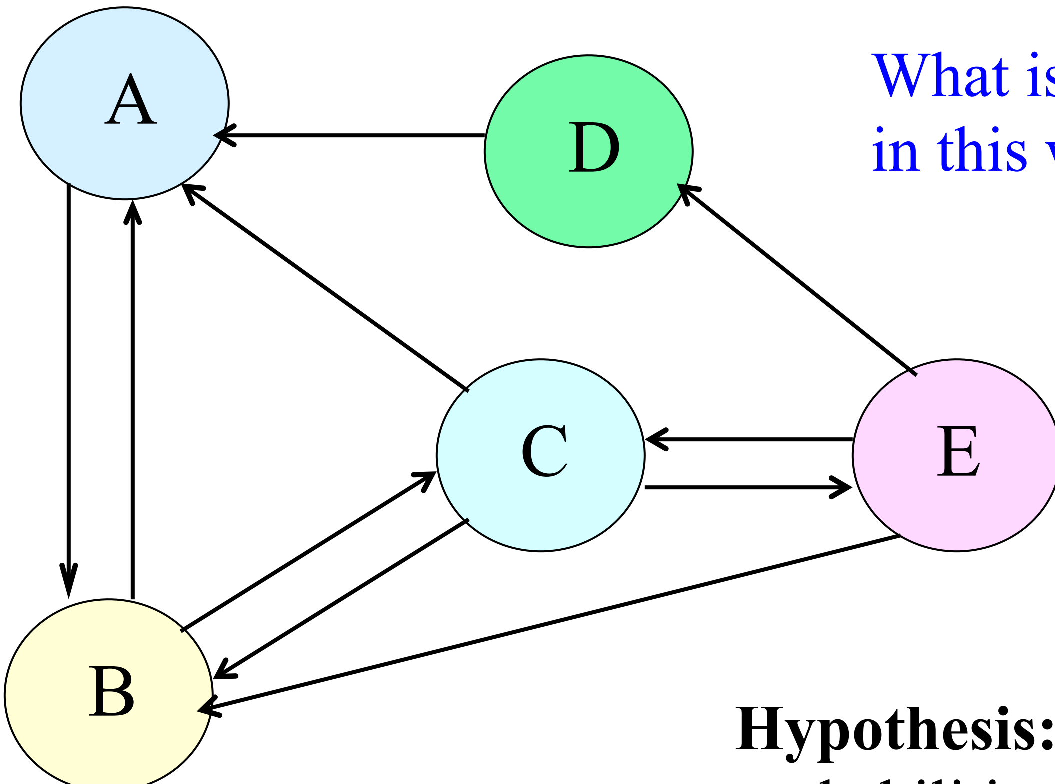


France: Influenza-like illness (ILI) data provided publicly by the Sentinelles network, INSERM, UPMC.

PageRanking

Random walk in web of 5 pages

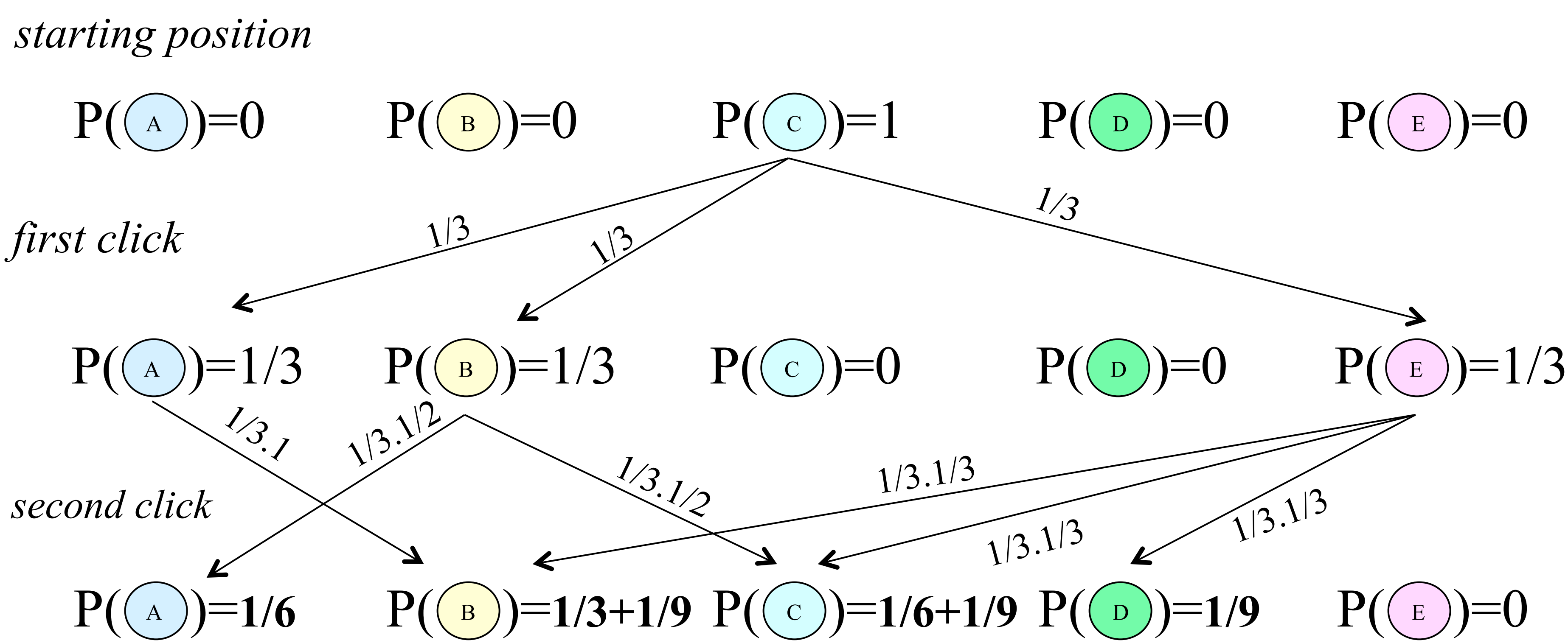
PageRank Google considers links to a page as the recommendation for this page; the recommendation of an important page counts more than the recommendation of a less important page.



What is the most important page in this web of five pages?

Hypothesis: Walk with uniform probabilities on the possible pages

What is the probability of being in a given page after a "long" walk?



The position of the walker after the t^{th} click depends only on the its position on $(t\text{-}1)^{th}$ click

Notations

- V a set of n pages (positions, stats)

  Ex: $V=\{A, B, C, D, E\}$ avec $n=5$

- $X_t \in V$ the position of the walker at time t for $t=0,1,2,\ldots$

- $P(I|J)$ the probability that I occurs if J occurred

  Ex: $P(X_1=A|X_0=C)$ the probability that the walker be on the page A starting from page C

Markov Chain

- $\{X_t,\ t=0,\ 1,\ 2,\ldots\}$ a random process taking its values in V
- Si $P(X_t=i)$ for $i \in V$ only depends to X_{t-1} and doesn't depend to X_{t-2}, X_{t-3}, X_{t-4}, …, then ***$\{X_t\}$ is a Markov Chain***.
- It is characterized by *its initial state and a transition matrix* given by:

$P_{j,i}=P(x_t=j|x_{t-1}=i)$ with $P_{j,i} \in [0,\ 1]$ for all $i,j \in V$ and $\sum_{i\in V} P_{j,i}=1$

The position of the walker after the t^{th} click depends only on the its position on $(t\text{-}1)^{th}$ click

The transition matrix of the web of 5 pages

$$P = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/3 & 1 & 0 \\ 1 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 1/3 & 0 & 0 \end{pmatrix} \end{array}$$

The columns represent the possible destinations (from the page C, the walker can only go to pages A, B and E). Non-zero elements on the lines indicate the origin (we can be on C if we come from B or E).

Stating Point: The walker is on the page C.

Let P_0 be the vector of probability representing this condition.

$$P_0 = \begin{pmatrix} P(x_0=A) \\ P(x_0=B) \\ P(x_0=C) \\ P(x_0=D) \\ P(x_0=E) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}$$

After the first click:

$$P_1 = P.\ P_0$$

$$P_1 = \begin{pmatrix} P(x_1=A) \\ P(x_1=B) \\ P(x_1=C) \\ P(x_1=D) \\ P(x_1=E) \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 1/3 & 1 & 0 \\ 1 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 1/3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 1/3 \\ 0 \\ 0 \\ 1/3 \end{pmatrix}$$

After the 2^{th} click:

$$P_2 = P.\, P_1 = P.\, (P.\, P_0) = P^2.\, P_0$$

$$P_1 = \begin{pmatrix} P(x_2=A) \\ P(x_2=B) \\ P(x_2=C) \\ P(x_2=D) \\ P(x_2=E) \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 1/3 & 1 & 0 \\ 1 & 0 & 1/3 & 0 & 1/3 \\ 0 & 1/2 & 0 & 0 & 1/3 \\ 0 & 0 & 0 & 0 & 1/3 \\ 0 & 0 & 1/3 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1/3 \\ 1/3 \\ 0 \\ 0 \\ 1/3 \end{pmatrix} = \begin{pmatrix} 1/6 \\ 4/9 \\ 5/18 \\ 1/9 \\ 0 \end{pmatrix}$$

After the t^{th} click:

$$P_t = P.\, P_{t-1} = P.\, (P.\, P_{n-2}) = \ldots = P^t.\, P_0$$

After an infinitely long walk?

La transposé de la matrice de transition

$$P^T = \begin{pmatrix} & A & B & C & D & E \\ A & 0 & 1 & 0 & 0 & 0 \\ B & 1/2 & 0 & 1/2 & 0 & 0 \\ C & 1/3 & 1/3 & 0 & 0 & 1/3 \\ D & 1 & 0 & 0 & 0 & 0 \\ E & 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix} \qquad u = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\sum_{j \in V} P^T_{i,j} = 1 \qquad \sum_{i \in V} P^T_{i,j} . u_i = \sum_{i \in V} P^T_{i,j} . 1 = 1$$

$P^T u = u$: $\lambda=1$ is an eigenvalue of P^T, u is its associated eigenvector and $\lambda=1$ is an eigenvalue of P

$\forall P_0^i = P(X_0 = i),\ i \in V$ avec $\sum_{j \in V} P_0^i = 1$, the probability distribution $P^t = P(X_t = i),\ i \in V$ converges to the a stationary state π when $t \rightarrow \infty$:

$$P^t = P(X_t = i)_{t \rightarrow \infty} \rightarrow \pi \text{ pour } i \in V$$

$$P^t = P(X_t = i)_{t\to\infty} \to \pi \text{ for } i \in V$$

The eigenvalues of the transition matrix P of our example are:

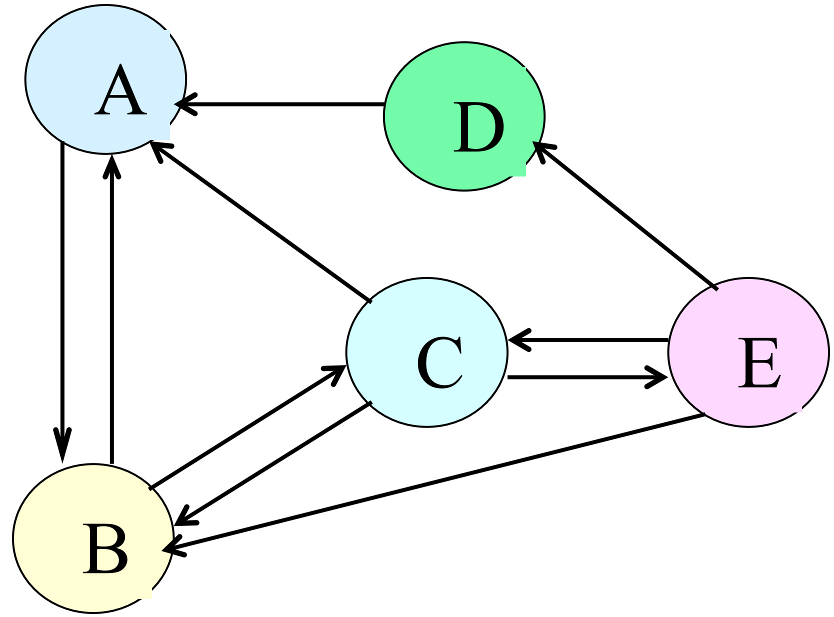
$$1 = \lambda_1 > |\lambda_2| = |\lambda_3| = 0.70228 > |\lambda_4| = |\lambda_{N=5}| = 0.33563$$

$P.\pi = \pi$ with

$$\pi = \begin{pmatrix} 12 \\ 16 \\ 9 \\ 1 \\ 3 \end{pmatrix} \qquad \pi / \|\pi\| = \begin{pmatrix} 12 \\ 16 \\ 9 \\ 1 \\ 3 \end{pmatrix} /41$$

During an infinitely long walk, walker will visit often the page B and less often the page D

$$\pi = \begin{pmatrix} 12 \\ 16 \\ 9 \\ 1 \\ 3 \end{pmatrix}$$



Each page inherits its rank as those that link to it.

$$rank(B) = 1/3\ rank(C) + 1/3\ rank(E) + rank(A)$$
$$= (1/3).9 + (1/3).3 + 12 = 16$$

Epidemic Modeling

Goal: to predict which individuals or groups of individuals most likely to spread an epidemic ?

Goal: Quick response and effective control of infectious disease propagation in order to help the vaccination campaigns in the actions carried out by healthcare organizations.

Homogeneous epidemiological models

- Each individual has equal contact to any other individual
- Rate of infection is determined by the density of the infected population

> ✧ These models allow to predict the epidemic threshold
> ✧ Good approximation of virus propagation where the contacts are sufficiently homogeneous

But the real network are not homogeneous

Our objective

Epidemiological models with any particular propagation topology

A model predicting the *epidemic threshold* with a good accuracy for arbitrary network is proposed by Wang and al. The threshold is related to the largest eigenvalue of the adjacency matrix of considered network

Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint Y. Wang, D. Charkrabarti, C. Wang, C. Faloutsos

Notations

λ_c minimum infectiousness of a virus for invading a network
v rate of infection of an individual in network
δ rate of curing an infected individual

$\lambda = v/\delta$ effective spreading rate
If $\lambda \geq \lambda_c$ the infection becomes persistent
if $\lambda < \lambda_c$ it dies out fast

Proposed Pagerank-like model

Pagerank-like model	Pagerank model
An individual in a social graph	A webpage in a web graphe
A virus	A walker
Propagation of the virus	Promenade of the walker
Pagerank of an individual is the probability to be infected by the virus in the course of epidemic	Pagerank of a specific page is the probability of the presence of the walker on the page

Mathematical formalism

$G=(V,E)$ directed graph where

V set of individuals

E set of outlinks between individuals (if $i \rightarrow j$, $j \rightarrow i$ is not necessarily true)

n number of individuals in G.

d_j number of links of individual j to other individuals

$d=(d_1, \ldots, d_n)$ degree of graph

A virus on individual i at step time t moves to individual j with the probability:

$P_{j,i}=P[s_{t+1}=j \mid s_t=i]$ is $1/d_i$ if $i \rightarrow j$ and is 0 otherwise

where s_t *the state of the virus at step time* t.

$\{s_t\}$ *is a Markov chain characterized by its initial state and a transition matrix* P *given by* $P_{j,i}=P[s_t=j \mid s_{t-1}=i]$ *with* $P_{j,i} \in [0,1]$ for $i,j \in V$ and $\sum_{i \in V} P_{j,i}=1$.

Mathematical formalism

Frobenius theorem → $\lambda=1$ is the largest eigenvalue of the matrix P.

Then, there is a stationary distribution for the final state of epidemic spread: $Px=x$.

x_i the probability that individual i be infected during epidemic

$x = (x_1, x_2, \ldots, x_n)$ the stationary distribution (infection vector) for the whole population is independent of starting distribution and verifies $Px=x$.

The impact of infection vector x in social graph is similar to that of pagerank vector in web graph.

Problem	Solution
Dangling individual	add a loop to itself
Small world non-uniqueness of ranking vector	add a jumping vector to the random virus propagation process

Computational algorithms

$$A = \alpha P + (1\text{-}\alpha) v z^T$$

A is disease transition matrix
v is the teleportation vector
z is the vector $(1, \ldots, 1)^T$
α (<1) damping factor
$1\text{-}\alpha$ jumping rate; the probability for the virus to jump from any individual to any other individual in a social graph.

Computational algorithms

Input:

$A(n \times n)$: the disease transition matrix with each column sum as 1,

w_0: the starting vector,

m: the size of subspace,

r: the number of shifts and $m = r + k$.

Output:

x: the dominant eigenvector associated with eigenvalue 1.

$w_1 = w_0 / \parallel w_0 \parallel$;

compute the m-step Arnoldi factorization: $AW_m = W_m H_m + f_m e_m^*$;

while *not converge* **do**

  compute the spectrum of H_m $(\sigma(H_m))$ and select r shifts $\mu_1, \mu_2, ..., \mu_r$;

  $Q = I_m$;

  for $j = 1, 2, ...r$ **do**

    QR factorization: $Q_j R_j = H_m - \mu_j I$;

    $H_m = Q_j^* H_m Q_j$;

    $Q = QQ_j$;

  end

  $\beta_k = H_m(k+1, k); \sigma_k = Q(m, k)$;

  $f_k = w_{k+1}\beta_k + f_m \sigma_k$;

  $W_k = W_m Q(:, 1:k); H_k = H_m(1:k, 1:k)$;

  begin with the k step Arnoldi factorization $AW_k = W_k H_k + f_k e_k^*$, apply r additional steps of the Arnoldi procedure to obtain a new m-step Arnoldi factorization $AW_m = W_m H_m + f_m e_m^*$

end

Experiments

Parallel programming model

- Distributed computation
- Message passing MPI

Grid5000 platform

- Cluster "Taurus": 16 nodes2 cpus per node6 cores per cpu=192 cores
- Cluster "Graphene": 144 nodes1 cpus per node4 cores per cpu=576 cores

Name of Cluster	CPU	Network	Memory
Taurus	Intel Xeon	Gigabit Ethernet	32 GB
Graphene	Intel Xeon X3440	Gigabit Ethernet	16 GB

Experiments

Graphs/matrices tests

ba a real network graph collected at the Oregon router views
stanford Graph representing pages (nodes) from Stanford University (stanford.edu) and directed edges represent hyperlinks between them.
twitter Graph collected from 467 million Twitter posts from 20 million users covering a 7 month period from June 1 2009 to December 31 2009.
yahoo This dataset contains URLs and hyperlinks for over 1.4 billion public web pages indexed by the Yahoo! AltaVista search engine in 2002. The dataset encodes the graph or map of links among web pages, where nodes in the graph are URLs.

Name	n	nnz	$maxDegree$	Storage
ba	7010	13985	148	117 KB
stanford	281,903	2,321,669	255	30 MB
twitter	41,652,230	1,469,914,131	2997469	25 GB
yahoo	1,413,511,394	8,050,112,173	2514	78 GB

Experiments

Stochastic simulation using the infection vector

Initialization

- Introduction of *x%* randomly infected individuals in social graph
- If (vaccination) *x%* randomly individuals in social graph

Iterate

1. Individual infects each of its neighbors with $v = 0.2$ probability
2. If (individual is infected) then it tries to infect a non-neighbor individual with $(1-\alpha) = 0.2$
3. probability
4. Each infected individual is cured with $\delta = 0.24$ probability
5. Go to 1

Pagerank-like Model

Initialization

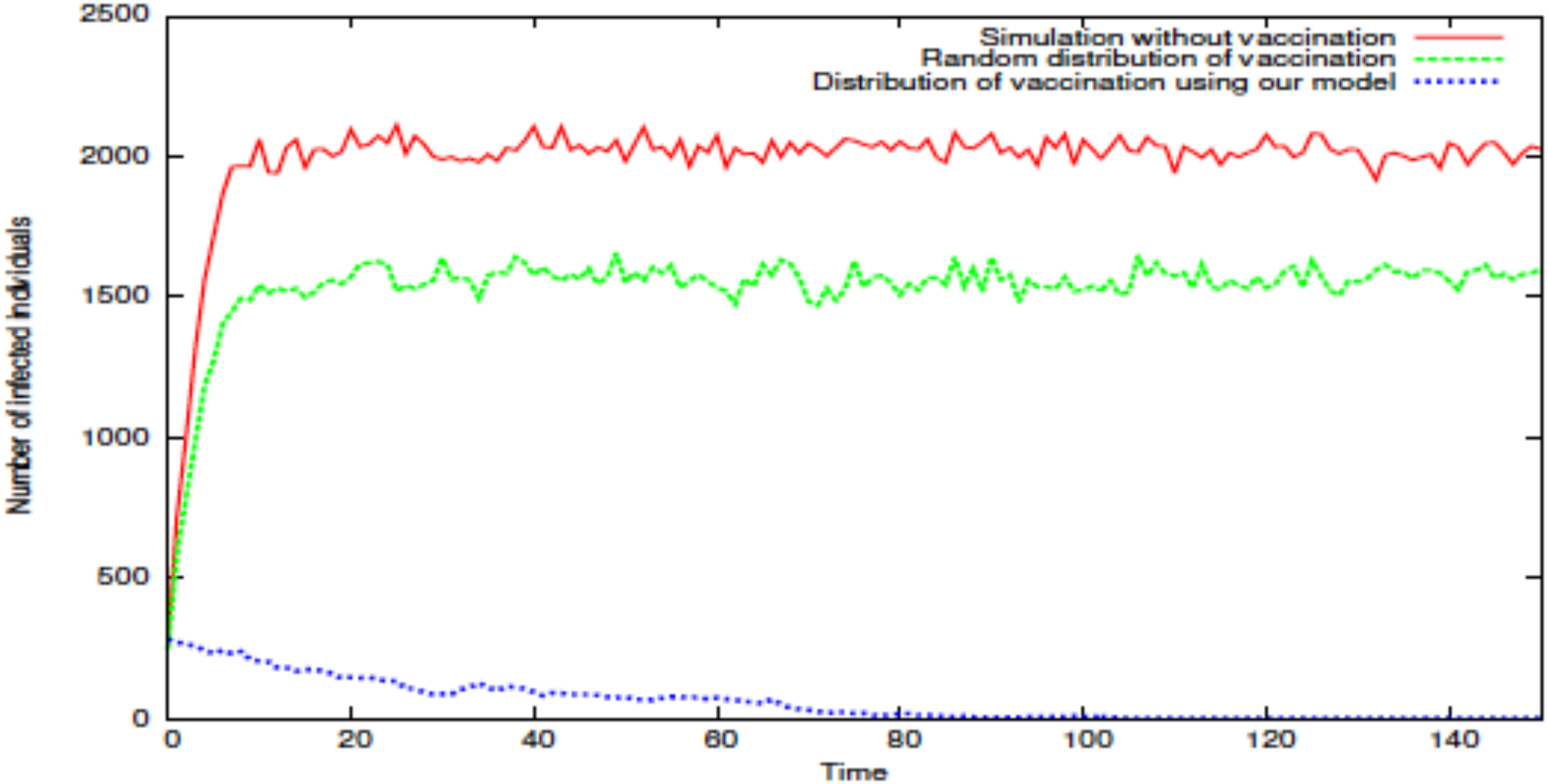
- Introduction of *x%* randomly infected individuals in social graph
- *x%* of most "important" individuals in infection vector is vaccinated

Iterate

1. Individual infects each of its neighbors with $v = 0.2$ probability
2. If (individual is infected) then it tries to infect a non-neighbor individual with $(1-\alpha) = 0.2$ probability
3. Each infected individual is cured with $\delta = 0.24$ probability
4. Go to 1

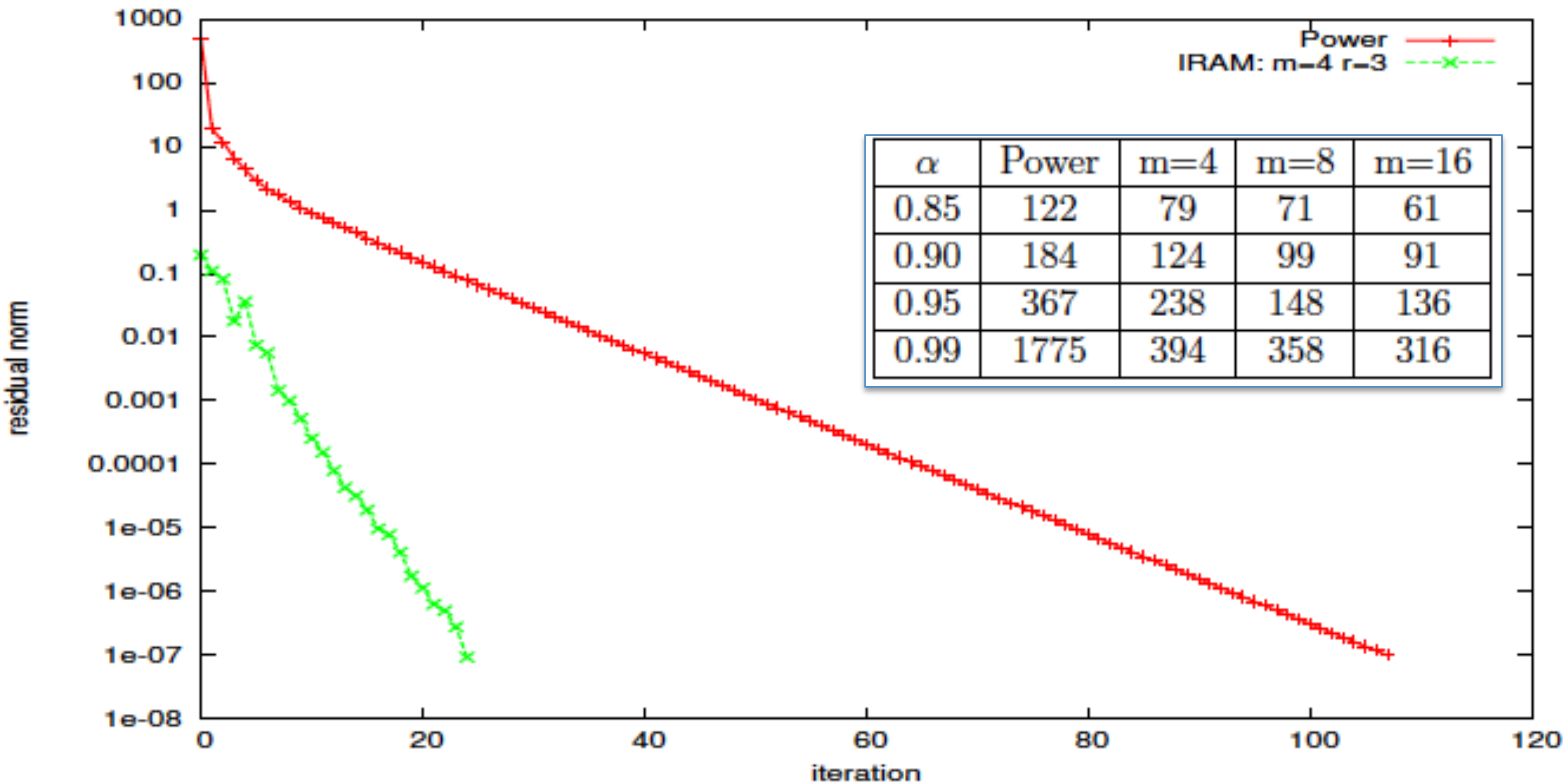
Experiments

Stochastic simulation using the infection vector



Time series of infection in an 7010-node power-law social graph ba, with $\nu = 0.2$, $\delta = 0.24$ and x=5

Experiments

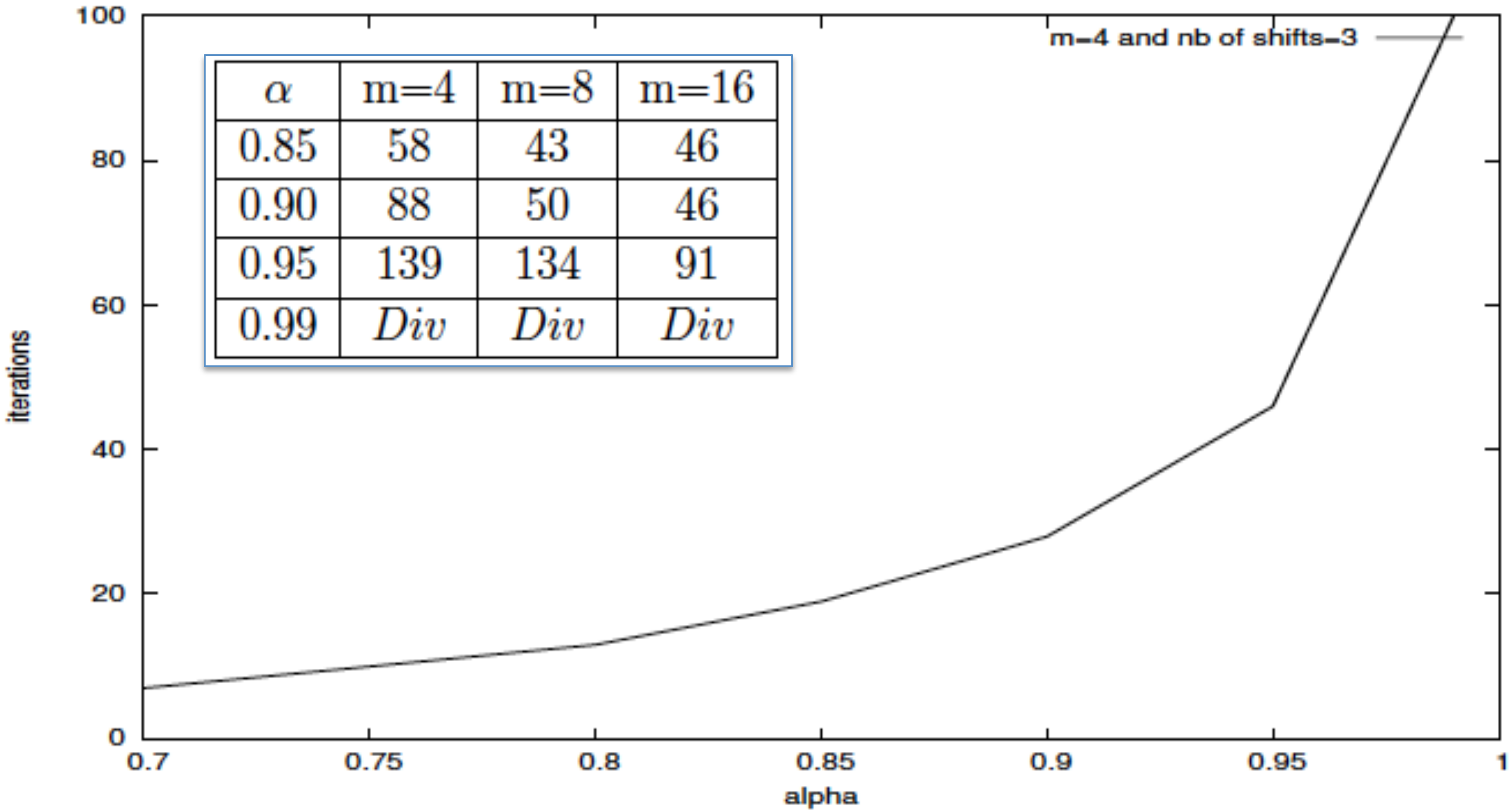


α	Power	m=4	m=8	m=16
0.85	122	79	71	61
0.90	184	124	99	91
0.95	367	238	148	136
0.99	1775	394	358	316

Convergence behavior for the 281903 X 281903 Stanford matrix, α= 0:85

Number of matrix vector products for the 281903 281903 Stanford graph

Experiments

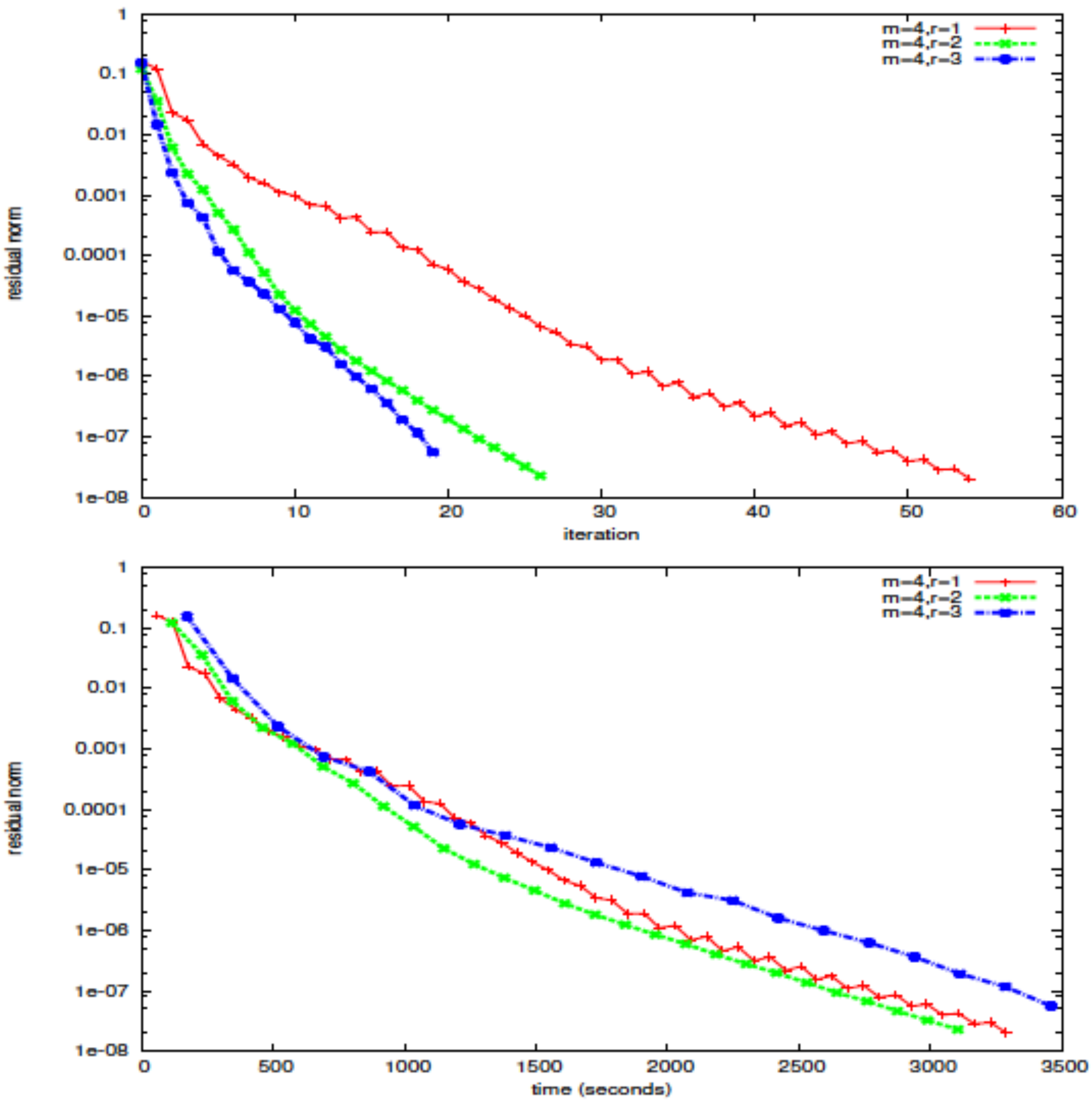


α	m=4	m=8	m=16
0.85	58	43	46
0.90	88	50	46
0.95	139	134	91
0.99	*Div*	*Div*	*Div*

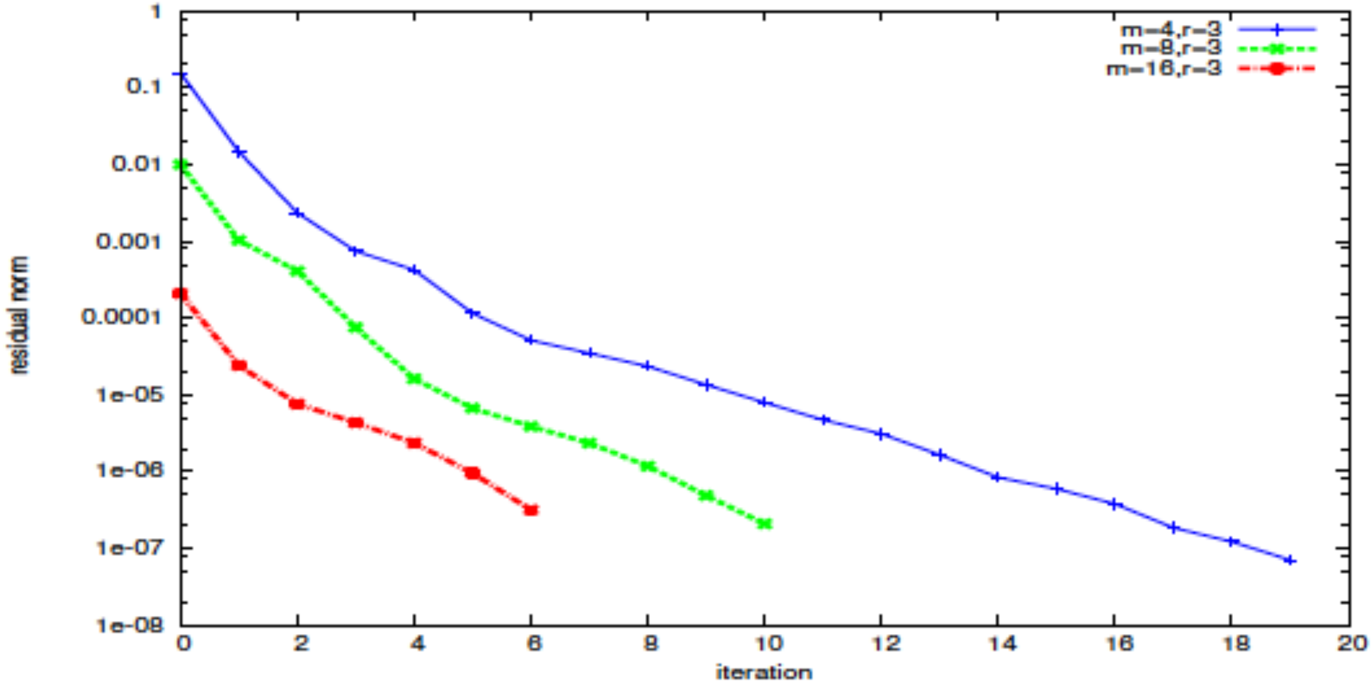
Number of iterations as grows for the 41652230 X 41652230 twitter graph.

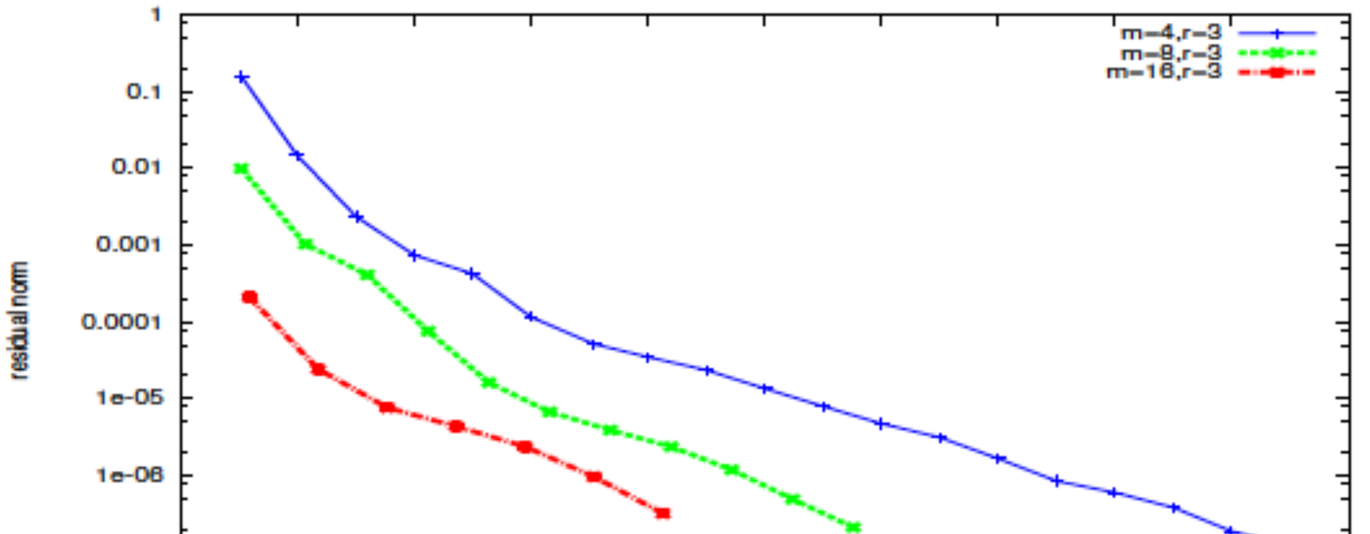
Experiments

Convergence with different number of shifts on twitter graph, where α= 0.85.

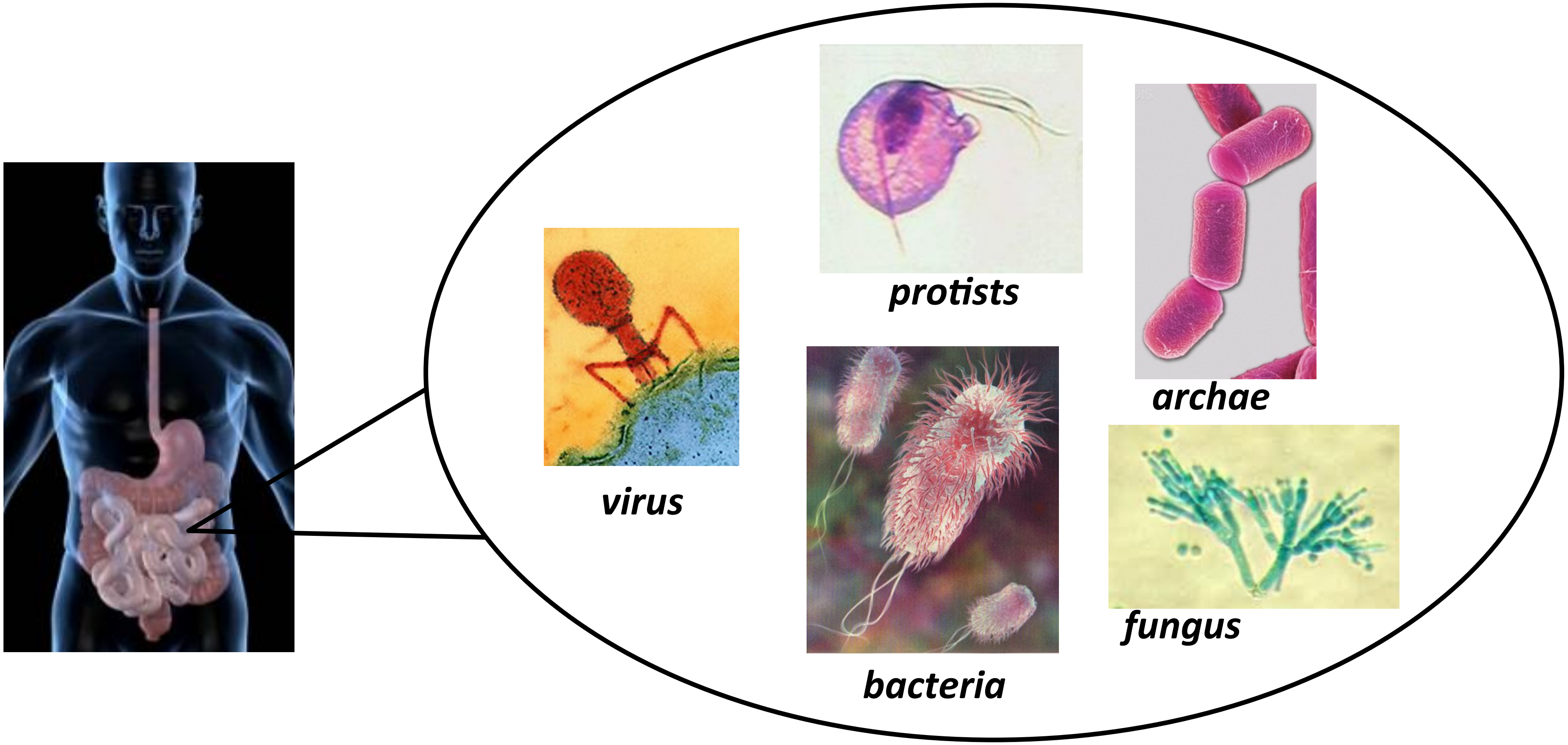


Convergence with different size of subspace on twitter graph where $\alpha = 0.85$.





Big Data: microbiota (J.-M. BATTO - INRA MGP)



- 2 kg – more bacteria than human cells (60.10^{16})
- An unknown organ: intestinal microbiota
- Amount of sequence generated has increased 10^9 times in 20 years.

Big Data: microbiota

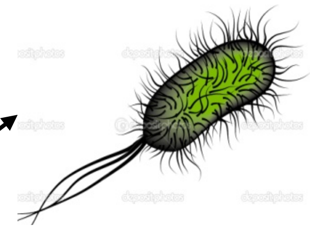
DNA preparation->Get Sequences->Compare to reference->Counting & analyzing

100 – 1000 individuals

Up to 10 millions of genes

id_fragment_	MH0001	MH0002	MH0003	MH0004	MH0005	MH0006	MH0007	MH0008
1	1.0246529439	8.3869647739	1.4874450848	2.3749503763	2.7242547919	4.7625759493	3.3844980635	1.5393655887
12	1.3467531952	1.19313569828	1.6585258148	9.8972700447	1.3267630323	1.3677934976	8.3035359532	8.5820901674
57	7.2892017460	4.3815309244	7.7707332906	2.5614971800	1.5525920094	7.7887477214	1.3471502008	1.6083954357
60	1.2526378532	1.89434148235	2.2545038192	5.3278079500	4.3180365068	5.41547566911	8.2863054205	3.5226732177
129	9.5875880325	4.56817276127	1.35029105751	2.7355209738	3.5472134610	6.7671067865	1.6093633649	4.6580802879
144	1.1328224535	5.8291787839	1.5753395670	3.6258692993	3.4426081345	3.5020169235	7.4943601770	2.4709660277
182	6.7023324214	6.7852874733	2.8881225396	1.11565209209	8.0145374559	1.1726692894	2.6077647116	2.6153179949
183	4.6178749698	0	2.2914030066	2.9504848716	0	3.18988198691	4.6551832869	1.48211645525
226	1.01068219623	4.5445646333	4.0299384274	1.2453790795	4.4732302079	7.4801496204	5.8219863330	5.5608090259
233	1.0614220079	1.0694423241	1.3794018099	4.4404063367	2.9665750583	3.4084920584	3.7364986913	7.1377548690
234	1.2627409521	9.9364209779	0	6.0509943978	0	4.9064710222	4.4552997108	0
241	1.20796851344	2.14481282084	3.6982056910	3.9183878356	4.1050069896	5.9720328692	1.06473125057	9.0761558293
305	3.8180469521	2.6821842247	2.66726611017	1.5225369727	3.3943923342	2.7671053485	2.3009688632	6.0657795514
336	4.6774055863	4.31859180624	9.5738868718	3.15587774118	7.7577180236	5.8109604316	5.3941829174	1.3078654156
378	9.0855751438	9.5325014259	0	0	0	1.17675524517	9.1589784995	4.3740510020
448	8.2490245187	4.2948632798	2.8563849293	2.0596654007	2.2898678445	8.1295252468	1.3067480313	1.9707262756
451	1.15537420813	5.1089223328	4.5303882975	4.0134307940	3.6206851565	9.2709854413	2.6261724625	1.87541010285
548	1.2977239878	8.08291179307	2.5297423705	3.8219906220	3.5568166228	1.2560645402	2.6496412253	4.3634085422
549	4.3167827556	0	3.8085132391	1.2554151014	8.4548966568	3.1811185748	5.7772018228	3.9414525513
550	6.7728832890	1.77651162938	0	0	0	1.31582631960	0	0
604	1.6843557917	2.3531803258	0	0	0	2.2734172014	6.2914030321	0
667	4.3613263603	5.9217054312	6.5639148629	2.9750722455	2.9143772567	2.7047541014	3.6982845002	4.6192629522
668	3.2878074218	6.4506344601	1.68240147942	2.0796621522	1.4939720306	1.3115679819	1.64061313900	2.6116906468
694	9.71111942183	9.5792293741	2.5483434173	2.8875701207	2.3572168988	8.51417030331	2.3930076177	1.75819697142
781	5.6827274172	0	1.18689967384	1.7116854015	1.4492040331	9.2528426127	1.15741885833	3.6849881729
793	5.81872302114	6.36535111506	8.4668087482	3.7794077387	5.01234915814	1.57156237954	2.1405755939	5.8415664848
854	6.9878954569	4.8250933143	3.0562143277	1.2592898394	1.1873388823	3.6589409122	1.0927775934	6.3257880453
888	1.10126367464	4.2537844423	2.4004184696	2.8730249126	2.5375844187	8.43367133167	2.5637285793	1.5023537973
903	6.9627772130	7.1226494299	0	0	1.79765326115	4.9599683698	2.35137513442	0
952	6.8983070536	7.2376399715	0	0	2.31531082061	1.8494669936	0	0
994	1.0974066213	4.15779317515	1.7016761281	4.0609005774	2.9592138299	6.31711490044	2.3047347697	5.8702484807
1009	7.0509488786	0	2.5408702695	3.1408387343	2.8203650871	1.2733803092	5.5061307695	0
1123	1.110839138175	4.9787587065	2.2074821959	4.6615781682	3.4304262231	9.2191652965	2.3320392708	3.9979383363
1124	9.12514215593	3.9005245355	0	1.0688882319	0	2.8890398234	1.8738429065	0
1126	8.32261082131	2.4841052444	2.2028053268	2.1934854692	1.8745867269	1.1243996092	2.3867677022	8.3588856014
1130	1.7287429374	0	0	4.49381541991	1.34509719541	0	5.2520016570	1.88114780859
1190	8.9810739477	3.8037231967	8.4324745684	3.5179686407	2.6676124671	7.0433525623	7.7661897982	1.1344862209
1191	7.7670679920	8.9640495060	1.5897922236	1.8014198918	2.6470031964	5.0902761293	1.7225601765	8.2264261965
1314	7.1794292847	9.9829516849	4.4262414401	1.4134442852	5.1587827302	1.5404522685	6.7142447749	0
1389	9.6459768114	6.2112753790	1.2710563744	2.9677982303	2.3044244616	6.7239046649	1.6985538366	8.7694910066
1390	1.6806162007	4.2671544844	2.2359658371	5.7582043463	3.3410478724	5.0282196828	1.21134876929	4.8950297814
1470	7.0550867594	2.171291991471	0	9.9169074853	1.0686049941	6.9690060630	7.3017411927	9.9631161714
1599	5.8386924905	3.3692461936	0	0	0	8.3184422503	6.4744503186	3.0920015704
1690	1.14049126649	2.24616412910	7.9672345922	4.3497608004	2.7857426743	8.7620925037	1.7696830871	1.6490675042
1798	7.2967527147	1.29557754187	1.7328735238	2.7219445517	4.8246629900	9.7826719194	1.9709514505	3.3092958244
1803	4.8897307755	2.3319365063	8.2714726673	1.15026659614	1.2394797783	3.2241408626	7.6179093248	4.2800976392
1806	2.3764502768	3.4283557760	0	0	0	3.3857519334	7.2468408829	0
1834	5.6440694075	2.6587248875	3.9294184213	6.88111947960	8.7233060745	5.2513703458	1.0643937598	0
1872	1.07411364936	7.4895225961	3.5878879127	7.8636270808	2.5420559331	8.5016678505	1.0752721454	9.4803219957
1877	8.6647871282	5.3503805265	7.9075063016	2.3505785929	2.0814809142	5.0637297088	1.92165399614	4.6762996371
1880	4.2330520556	3.9083255846	0	0	0	0	1.4081929443	0
1883	5.3574565079	2.9007103948	1.28611706845	1.70136944044	4.0573595870	2.2804229575	3.6671691258	6.8652097368
1884	3.5621072053	0	0	3.0865302836	0	1.1123219608	1.0821847794	0
1900	6.0560452767	6.2761432746	0	0	0	3.52167628117	1.7405401842	0
1942	4.7673207617	1.4229340720	0	0	9.3373251914	1.17104284107	1.36717761583	0
1958	4.3503406516	1.5259205683	0	0	0	5.3351177979	9.9318394057	4.0655537021

bacteria = ~3000 genes

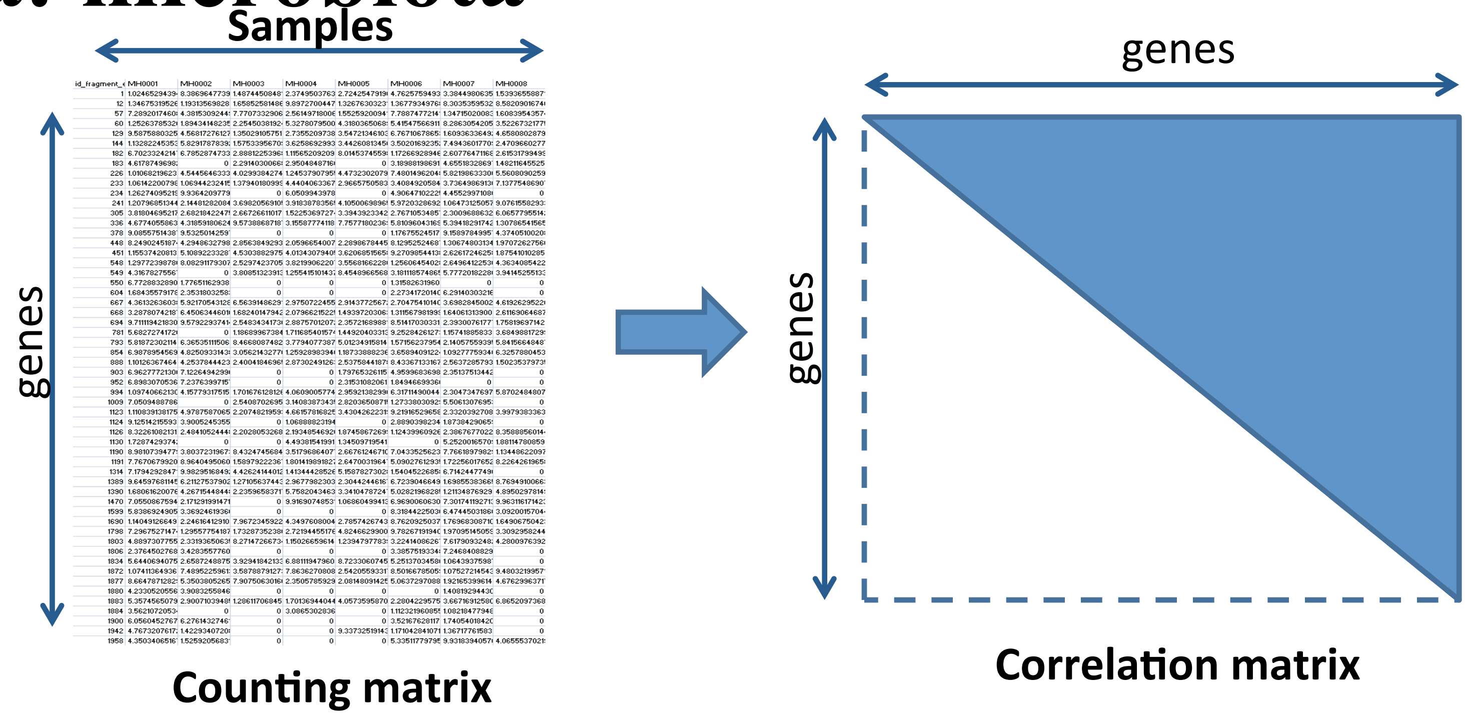


parasites = ~6000 genes

Virus = ~50 genes

Big Data: microbiota

Matrix 10^6 genes by 800 samples



MetaProf→

energy efficiency multiplied by 4.7 with the GPU implementation

Principal Coordinates Analysis applied on the matrices of distances between samples, concentrating the major variations in the samples in a small space implies using many linear algebra techniques.

Numerical methods/algorithms & HPC techniques have to be defined/adapted to increase data-scalability.

Concluding remarks and future work

- Conventional means of investigation are essential;
- Our predictions provide complementary solutions ;
- The virus/individual characteristics have to be integrated
- The impact of social graph structure on propagation of virus have to be extended

For efficient computation solver, many problems arise:

- Methods / algorithms
- Data Processing
- Programming models for Exascale computing (graph computation, PGAS ...)
- …