Towards two dimensional search engines

A.D. Chepelianskii

Experiments :

Cavendish Laboratory (Cambridge, Uk)

Since almost no experiments here

Main work done in Toulouse :

Leo Ermann, O. V. Zhirov, A. O. Zhirov
Klaus Frahm, B. Georgeot
Dima L. Shepelyansky

Quantware (LPT, Toulouse)

Outline

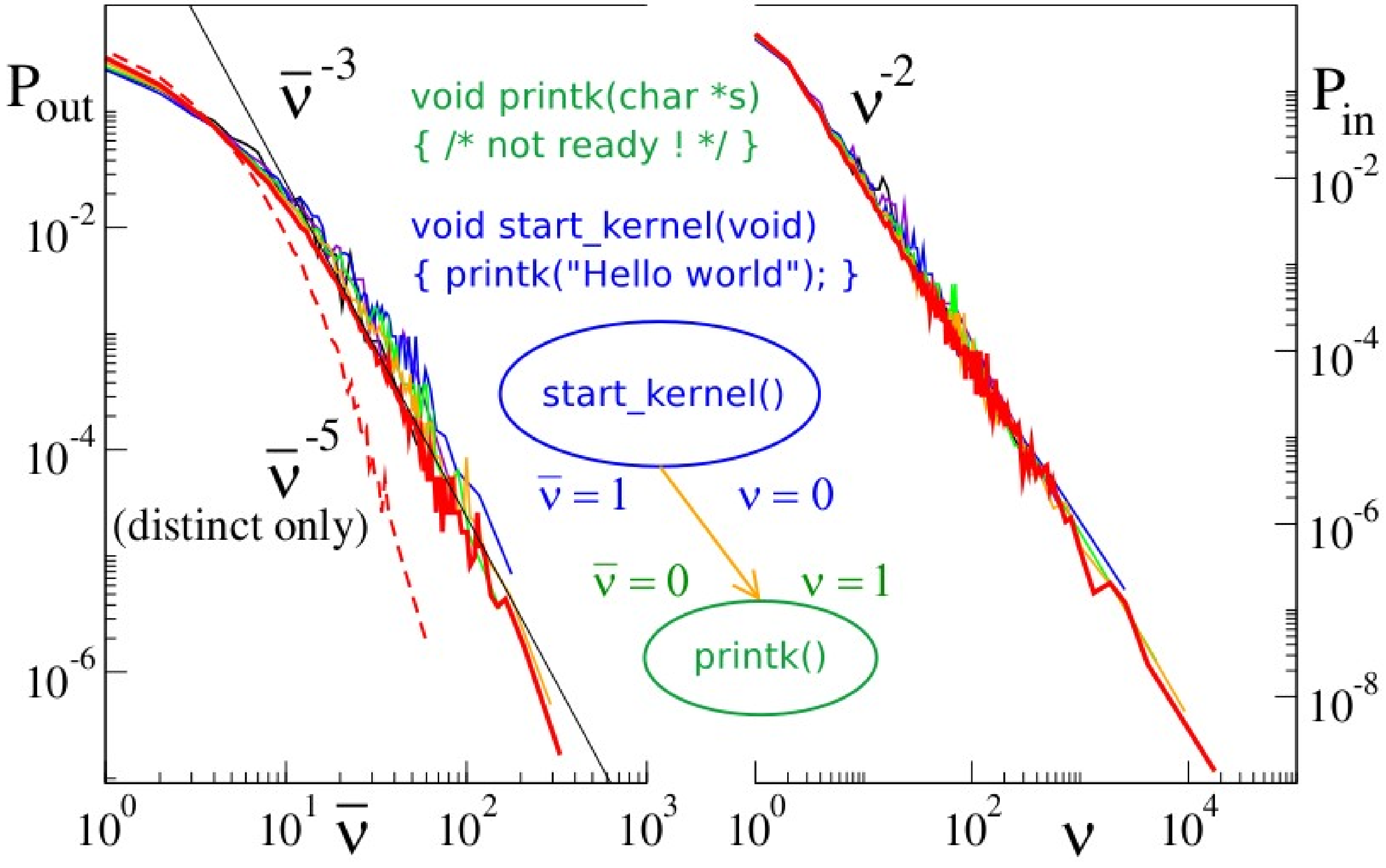
1. Introduction on the procedure call network in computer programs

2. A rating based on PageRank only is not sufficient, need for another rank based on time reversed dynamics

3. On the statistical correlation between the two ranks

4. Stability against “spam” links, manipulation ?

Scale free properties of the procedure call network in the Linux kernel



```
void printk(char *s)
{ /* not ready ! */ }

void start_kernel(void)
{ printk("Hello world"); }
```

Number of Procedures

N = 2751

N = 4358

N = 14079

N = 38766

N = 85756

N = 285509

- number of incoming procedure calls ν

- number of outgoing procedure calls $\bar{\nu}$

A.C. arXiv:1003.5455

Probability distributions $P_{in}(\nu)$ and $P_{out}(\bar{\nu})$ follow power laws

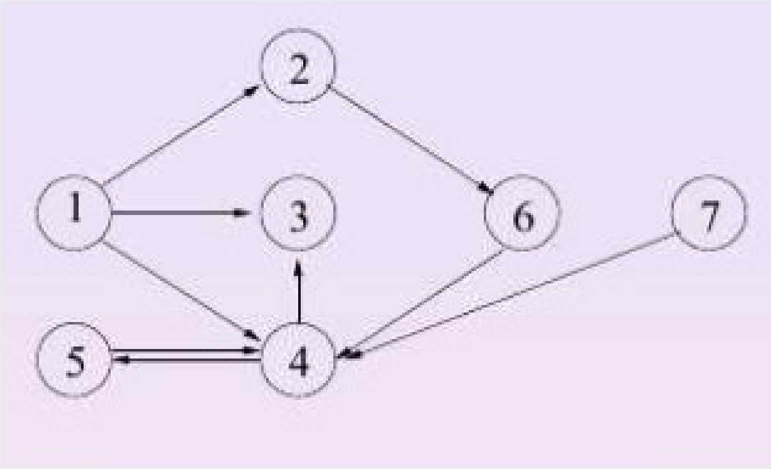
Reminder of the Pagerank method

$$\mathbf{G} = \alpha \mathbf{S} + (1-\alpha)\mathbf{E}/N$$

- $\mathbf{S}$ is constructed from the adjacency matrix $\mathbf{A}$ of directed network links between N nodes.
  1. $S_{ij} = A_{ij}/\sum_k A_{kj}$
  2. columns with only zero elements are replaced by $1/N$
- The second term describes a finite probability $1-\alpha$ for WWW surfer to jump at random to any node so that the matrix elements $E_{ij} = 1$.

PageRank: p

- $\mathbf{G}$ follows PFT (with $\lambda_1 = 1$)
- $\alpha = 0.85$ (random after 6 clicks)
- $\mathbf{G}p = p$



$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

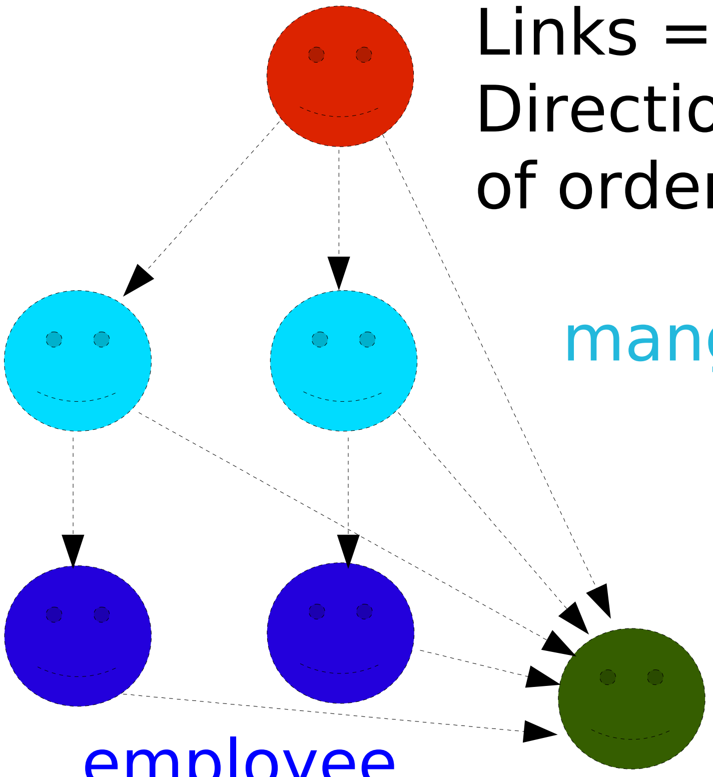
$$\begin{pmatrix} 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & 0 & 1 & 1 & 1 \\ 0 & 0 & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Pagerank method an organization networks

Example : organization of a small company

studied in M. Abel, D.L. Shepelyansky Eur. Phys. J. B v.84, p.493 (2011)

CEO

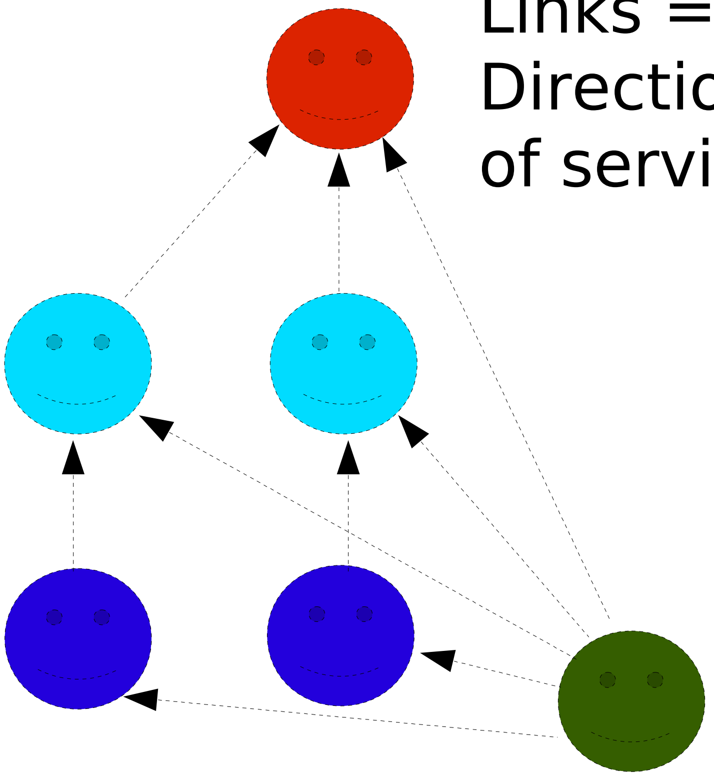


Links =
Direction
of orders

manger

employee

System
administrator



Links =
Direction
of services

Direct PageRank : System administrator will lead

PageRank on the inverse “service” network : CEO leads

Experimental slide: time reversal symmetry

Motions under opposite magnetic fields are related

by time-reversal symmetry

In equilibrium microscopic dynamics is time-reversal symmetric

↓

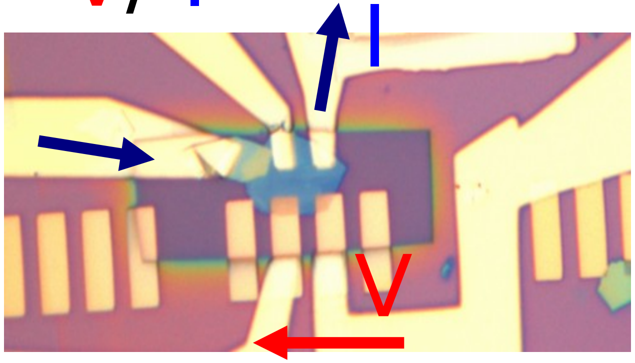
Onsager-Casimir reciprocity relations

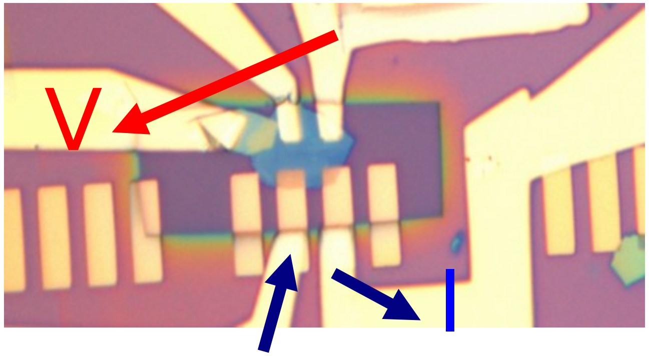
$R(H) = R(-H)$ and $R(H) = R^*(-H)$

$R = V/ I$

$R^* = V/ I$

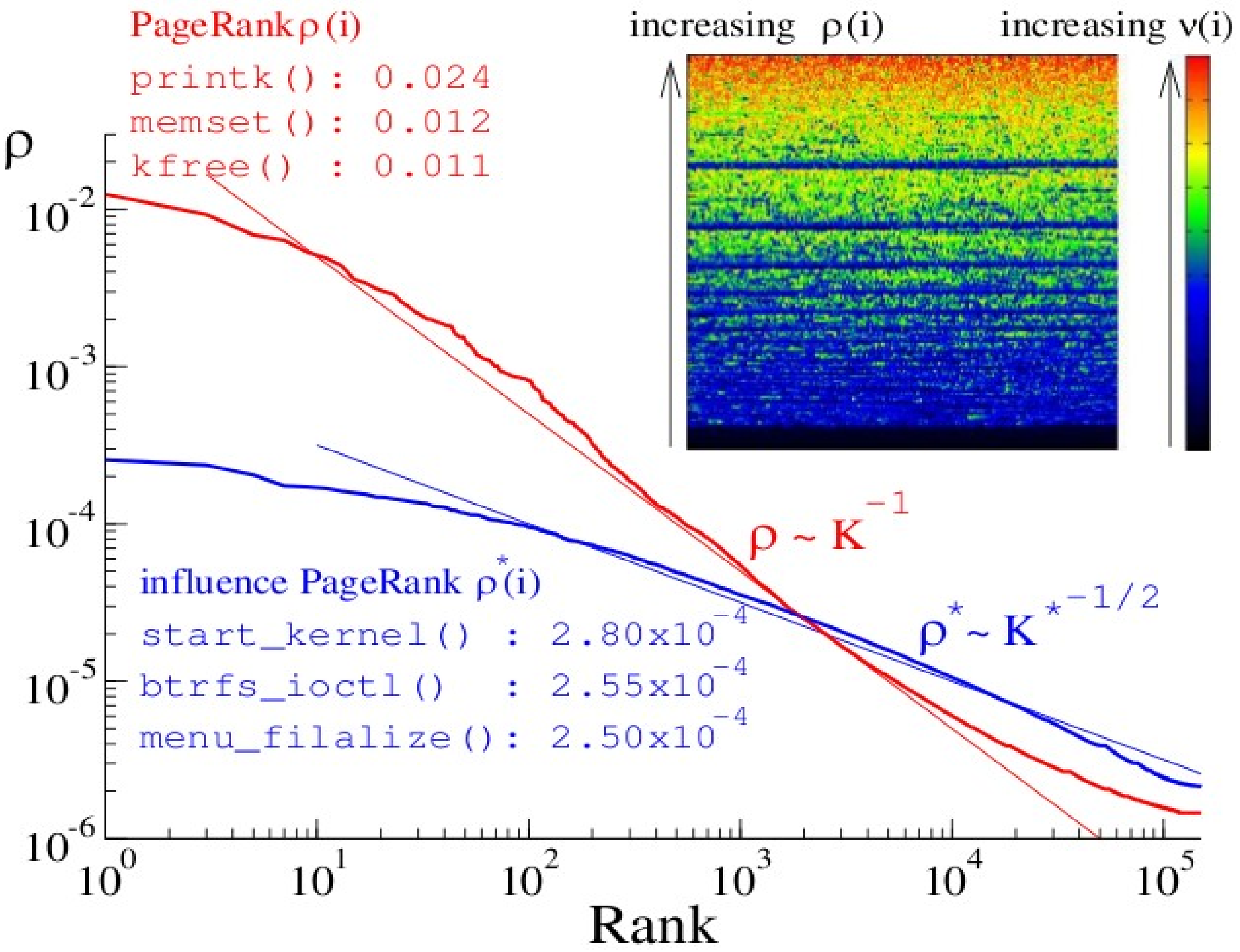
M. Büttiker (1986)





Studying time reversed dynamics can lead to intersting results !

Application to Linux kernel



Color: PageRank introduces mixing as compared to ordering by incoming links

~ Talk by

S. Vigna

on PageRank of actors

PageRank : general purpose procedures are leading

Service PageRank (CheiRank) : coordination/task distribution procedures

On the Similarity with HITS

Both approaches are similar in the sense that two ranks are obtained (Hubness/Authorities for HITS)

However $\rho(i)$ and $\rho^*(i)$ are the steady state distributions of two distinct (“time reversed”) Markov processes

While in HITS Hubness and Authorities are computed together thus strongly inter-dependent

We can study correlations between $\rho(i)$ and $\rho^*(i)$

Correlations between the two Ranks

We introduce the correlator : $\kappa = \mathrm{N}\,\Sigma_{\mathrm{i}}\rho(\mathrm{i})\rho^{*}(\mathrm{i}) - 1$

If ρ(i) and ρ*(i) are statistically independent

$$\kappa = N^2 \frac{1}{N} \sum_i \rho(i)\rho^*(i) - 1$$

$$= N^2 \int \rho\rho^* P(\rho,\rho^*) d\rho d\rho^* - 1$$

$$= N^2 \left(\int \rho P(\rho) d\rho\right) \left(\int \rho^* P(\rho^*) d\rho^*\right) - 1$$

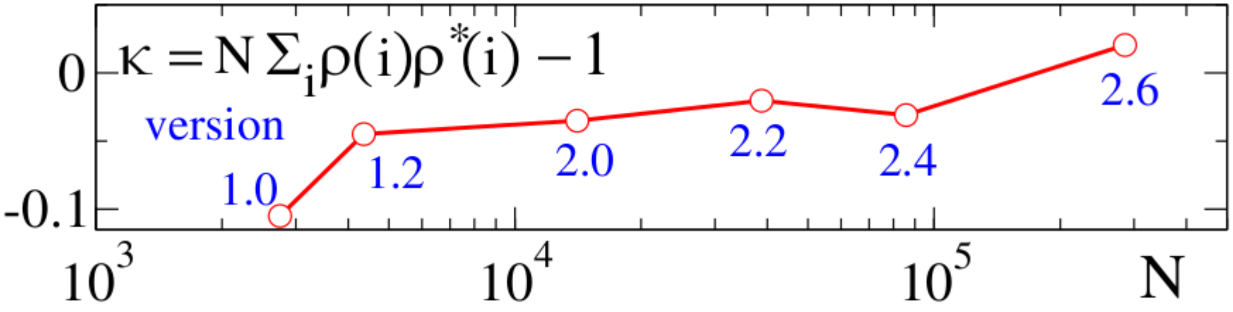
~ Connection with correlator discussed by Nelly Litvak yesterday

However we have $$\int \rho P(\rho) d\rho = \frac{1}{N} \sum_i \rho(i) = \frac{1}{N}$$

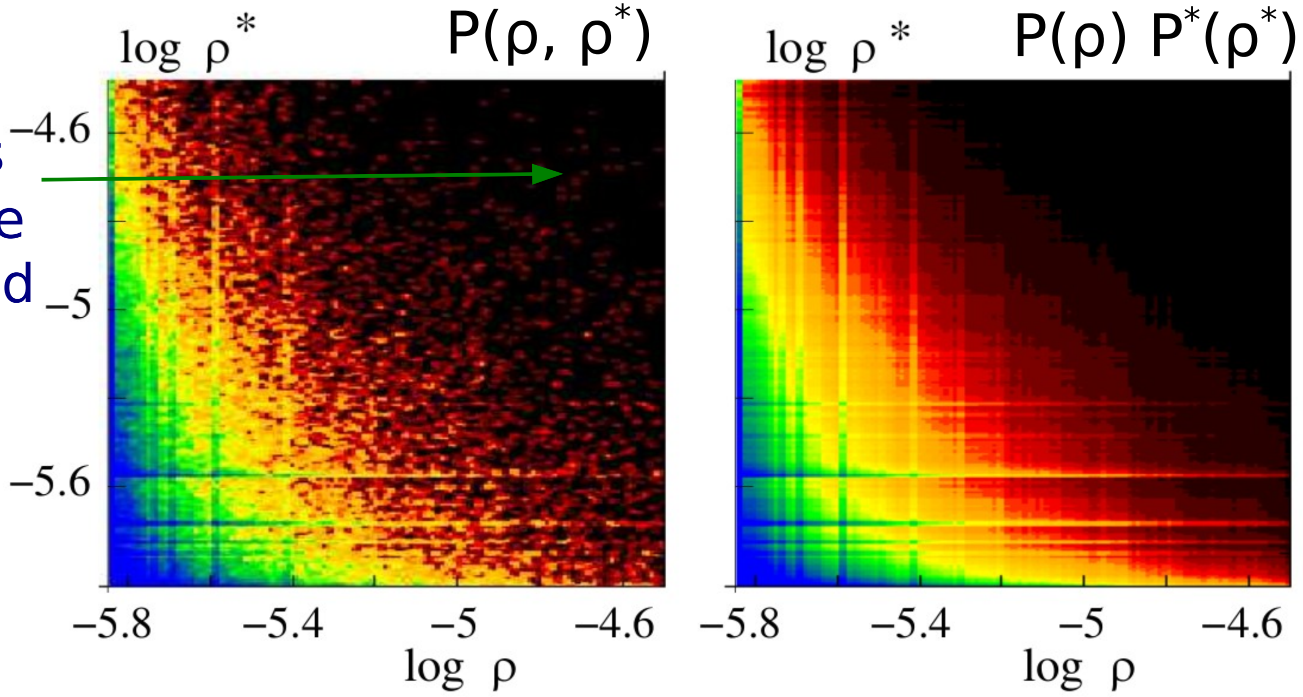
Thus κ = 0 when ρ(i) and ρ*(i) are independent

Correlation between ranks for Linux kernel

For Linux kernel κ ≃ 0 or slightly negative



Direct comparison between P(ρ, ρ*) and P(ρ) P(ρ*)



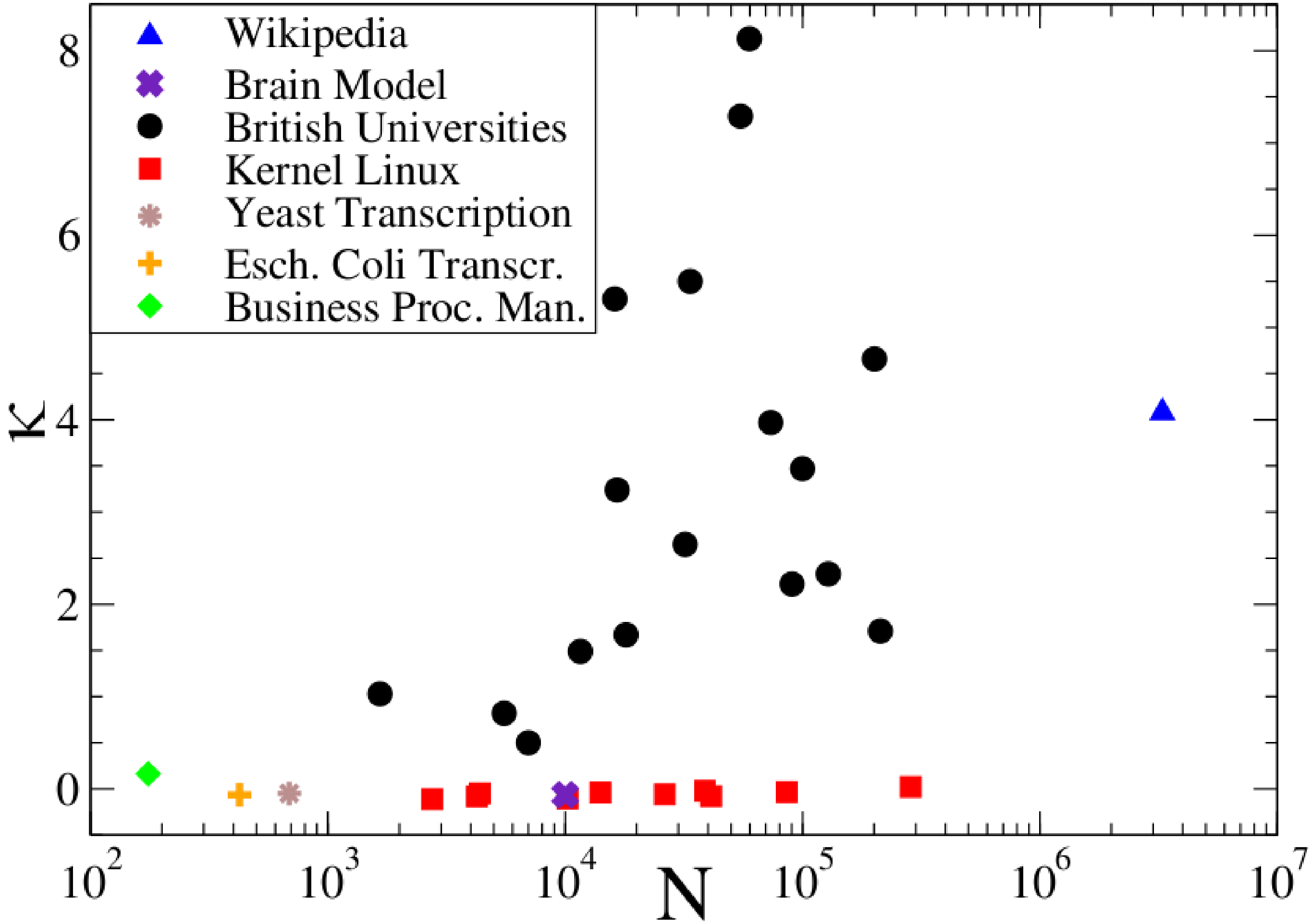
Procedures that provide services and distribute tasks: do_fork()

Colour: Probability amplitude

Data for procedure call network in Linux kernel 2.6

For Linux kernel, the two Ranks are statistically independent

Correlator values in other networks



Information storage networks have κ ≳ 1 (web, wikipedia, ...)

Functional networks have κ≃ 0 (Linux, Gene regulation, ...)

L. Ermann, A.C. D.L. Shepelyansky J. Phys. A: Math. Theor. 45 (2012) 275101

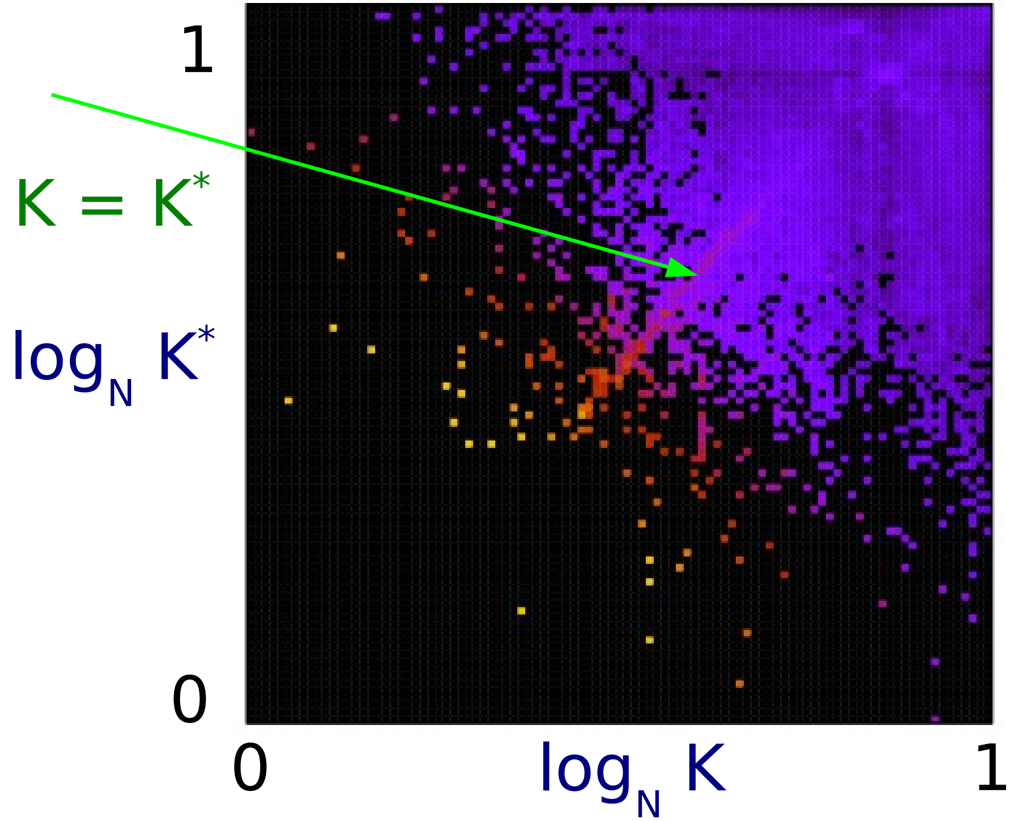
Density representation in (K, K*) plane

We introduce dN_i the number of sites with ranks in the interval [K + dK, K* + dK*]

The density W is then: $$W(K,K^*) = dN_i/dKdK^*$$

Peaked

Diagonal K = K*



κ= 1.72

Colour:

W(K, K*)

Data for Cambridge university website (2006)

N = 212710

Peaked diagonal, strong correlation between the two Ranks

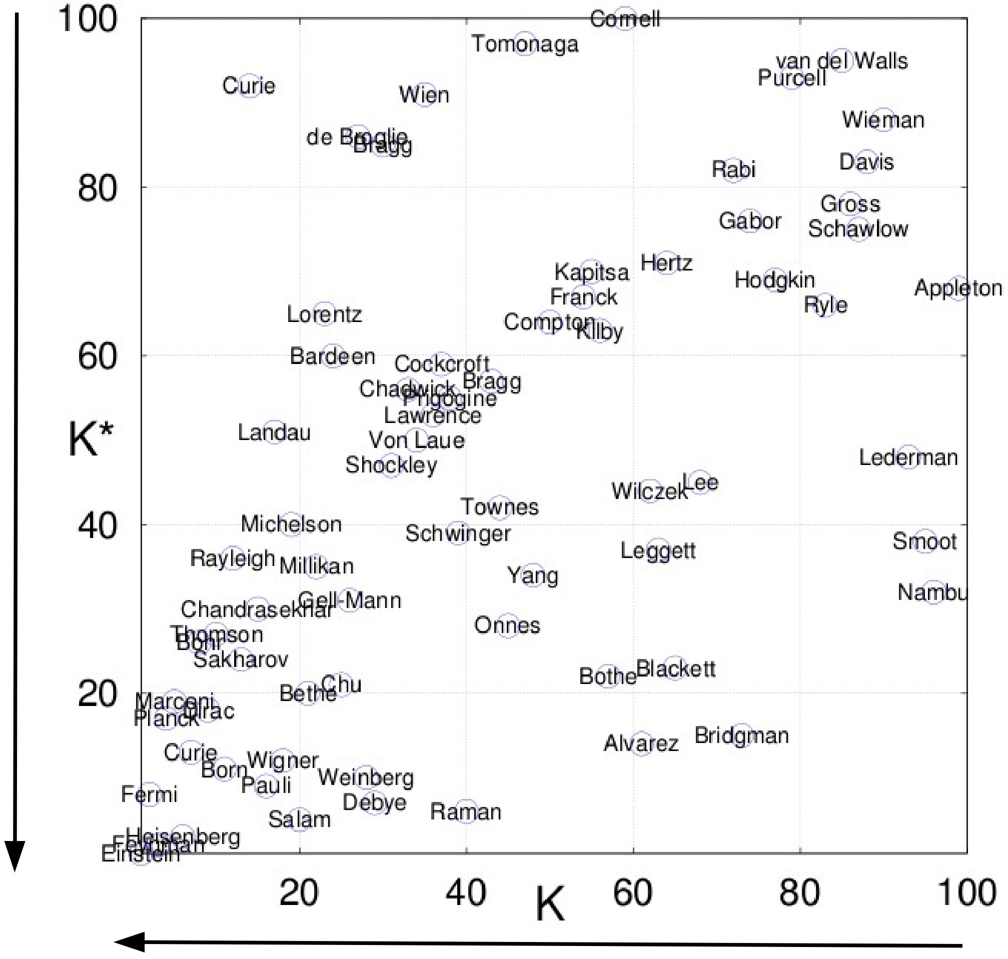
Towards two dimensional ranking

It is possible to organize search results, in the two dimensional (K, K*) plane

Example :

Nobel prize winners in physics, classified on the basis of the English wikipedia (2006)

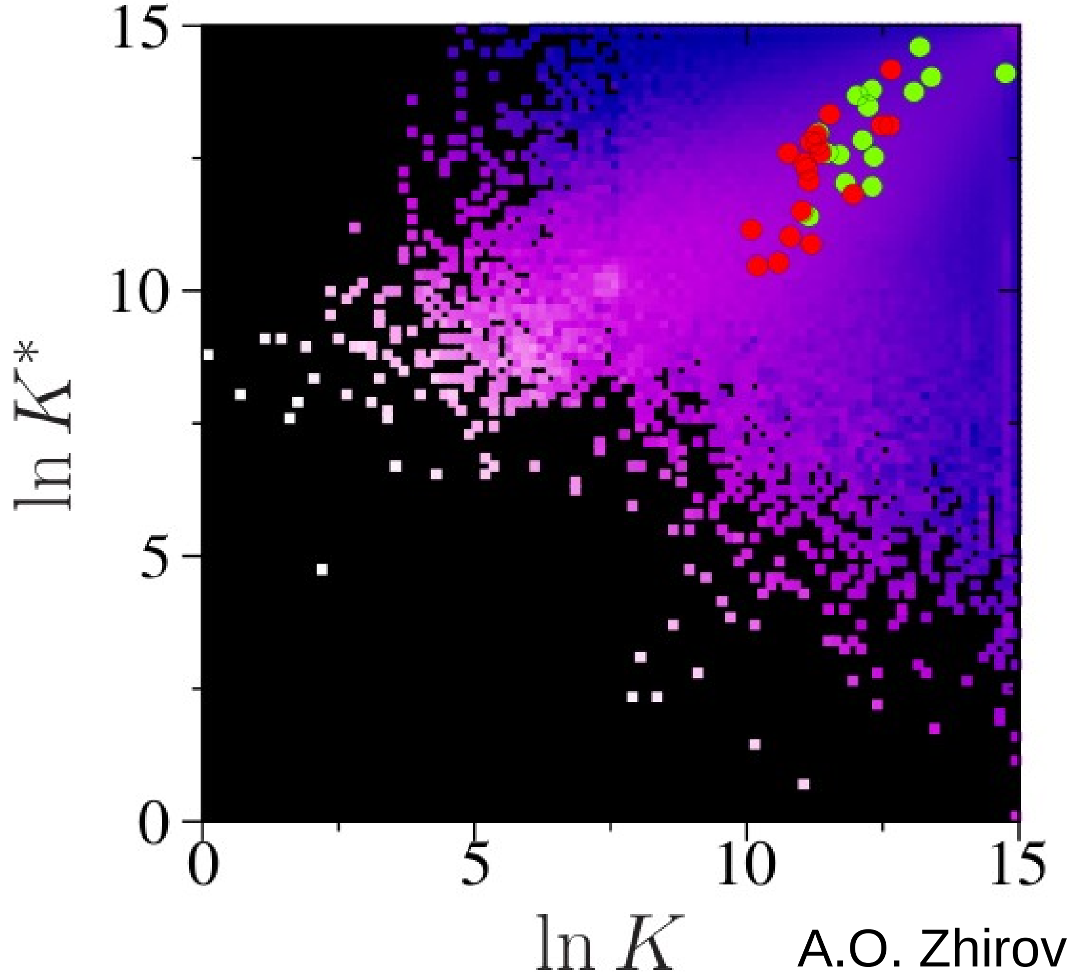
Dispersion on the (K, K*) plane



Good K* score, may highlight influence in other fields

2D classification of chess players

Chess players (Red points World Champions)



Ordering by PageRank K

1. Garry Kasparov
2. Bobby Fischer
3. Alexander Alekhine

Ordering by K^*

1. Bobby Fischer
2. Alexander Alekhine
3. Wilhelm Steinitz

A.O. Zhirov, O.V. Zhirov, D.L.Shepelyansky (2010)

More concentration around $K = K^*$ but still strong spreading

Protecting K^* against bias and manipulation

Since K^* is based on out-going links it can be easy to manipulate (for web, ...)

Many links are automatically generated (links to root, ...)

They should not influence the results

Interest in a filtering procedure

We invert only the links j → i for which

$$K(j) < \eta_K K(i)$$

where K(i) and K(j) PageRanks of sites i and j

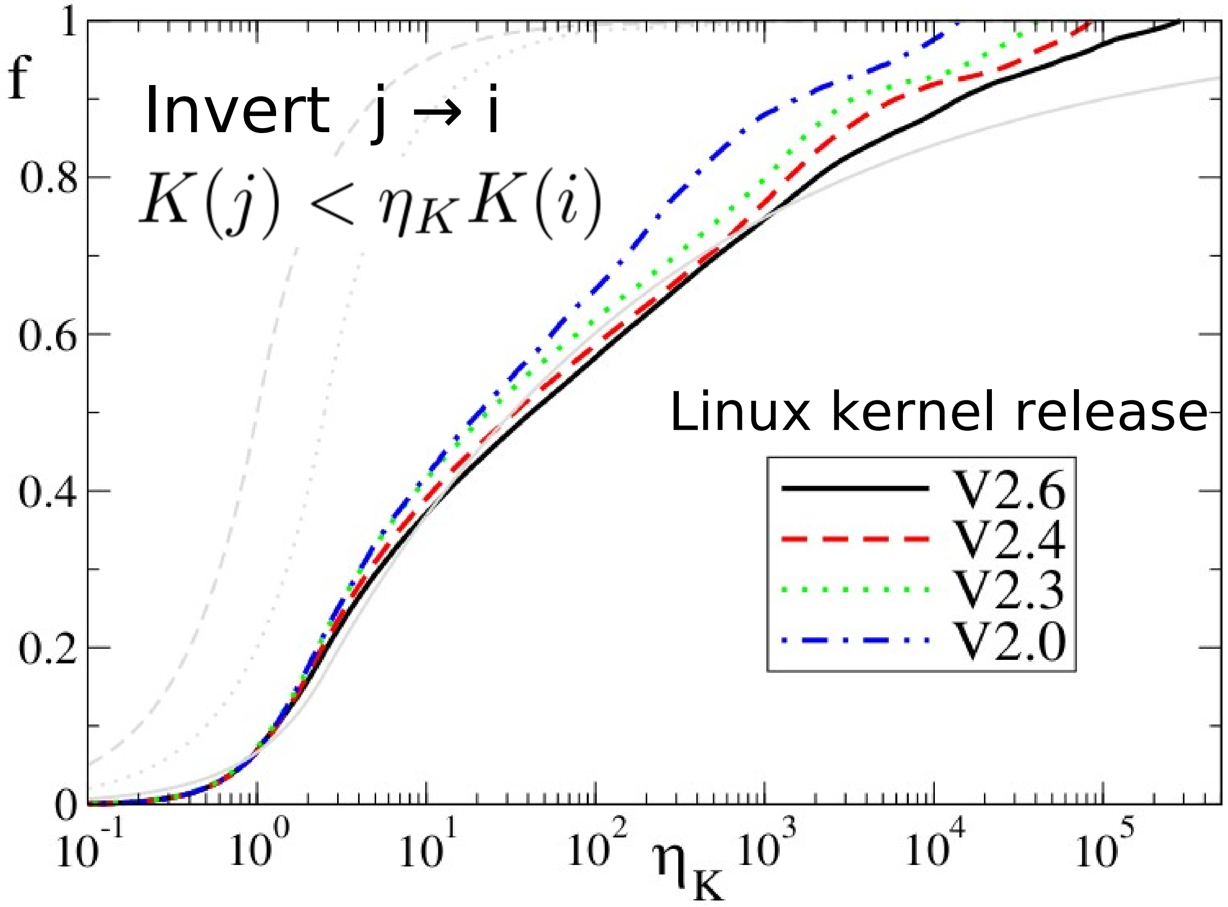
Invert links only between sites of comparable Rank

Here $\eta_K > 1$ filtering parameter (all links inversed for $\eta_K \to \infty$)

Fraction of inverted links as function of η_K

f : Fraction of inverted links

Data for procedure call network in Linux kernel



Invert j → i

$K(j) < \eta_K K(i)$

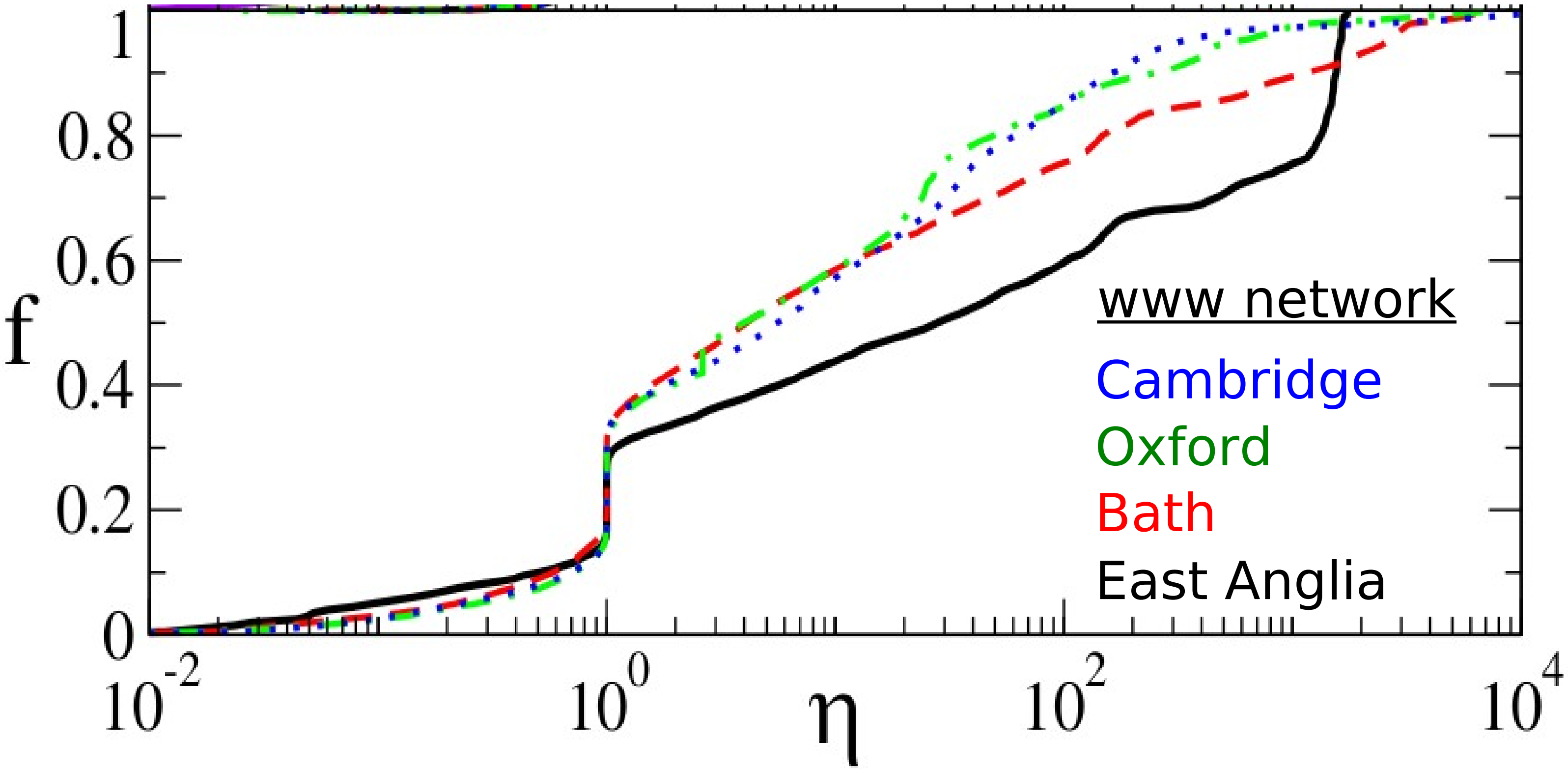
Analytical approximation : links only to sites with K(i) < a N

Use density of incoming links ∝ $1/K^{\nu}$

$$f(\eta_K) = \begin{cases} \frac{1-\nu}{2-\nu}(a\eta_K) & \eta_K \le 1/a \\ 1 + \left(\frac{1-\nu}{2-\nu} - 1\right)(a\eta_K)^{\nu-1} & \eta_K > 1/a \end{cases}$$

Good fit for
a = 0.4
ν = 0.8

Fraction of inverted links for universities

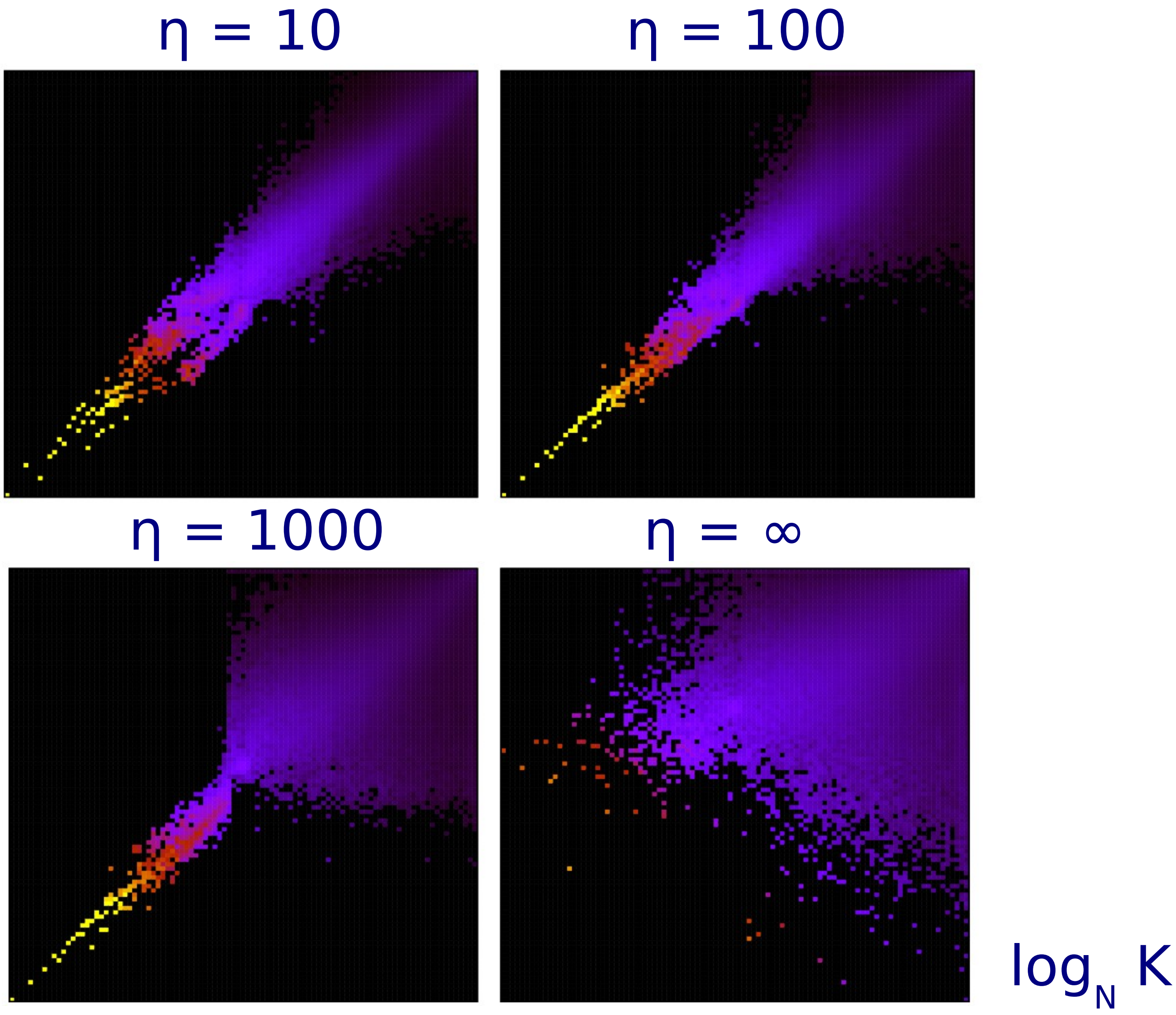


For British university networks, the fraction of inversed links has a strong jump at η = 1

(many sites with similar K are linked)

Except for the jump at η = 1, dependence f(η) (relatively) well understood ...

Formation of the 2D rank for Wikipedia



Spreading around diagonal $K = K^*$ increases with η

Even a finite spread already leads complementary information to PageRank, but which η to choose ?

Summary on 2D Ranking

1. 2D Ranking based on PageRank and its time-reversed conjugate
   (PageRank on the network where all links are reversed)

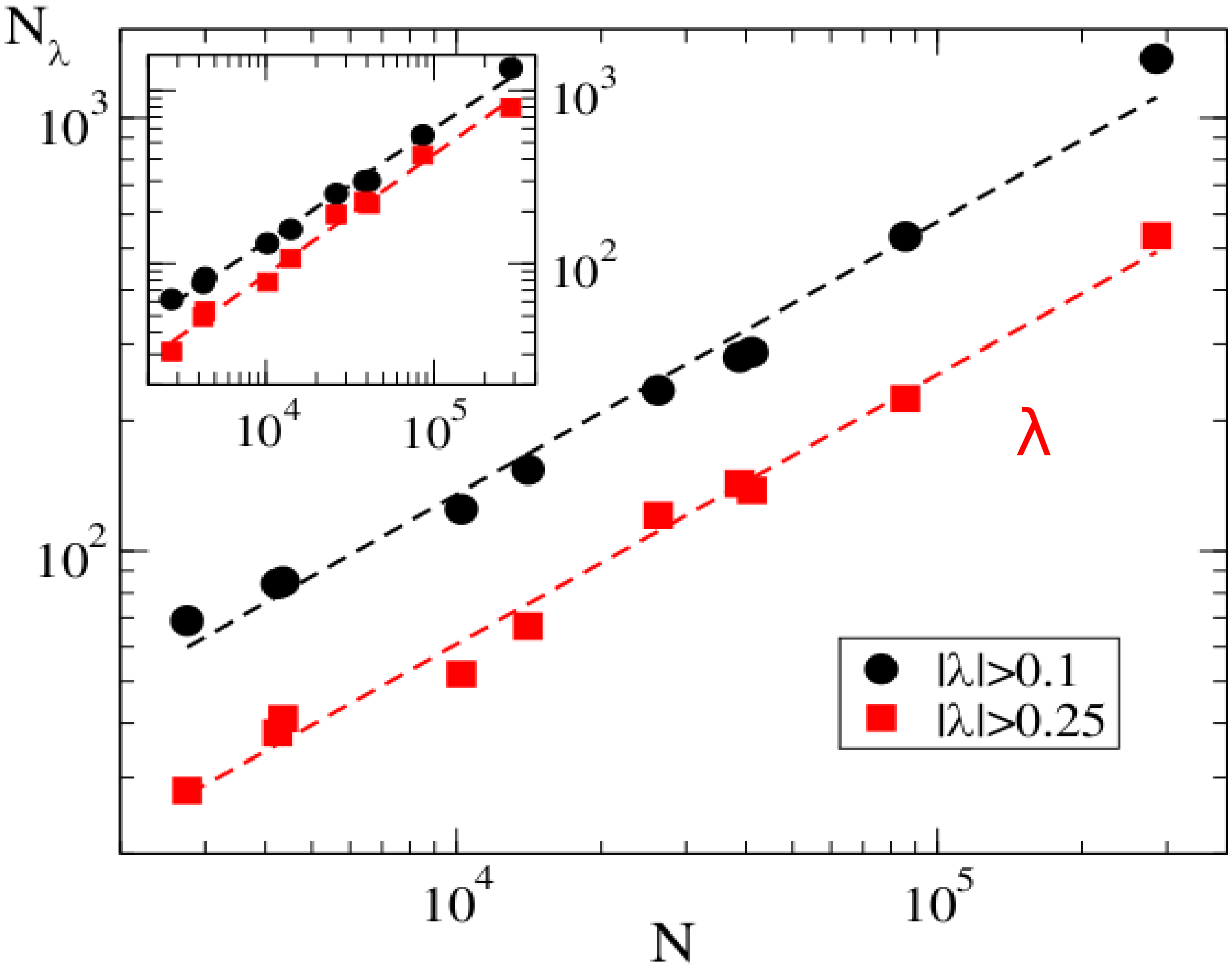
2. computer programs avoid correlations between the two Ranks (correlator $\kappa \lesssim 0$)

3. For web correlations between K and K^* are higher
   However they still provide distinct information

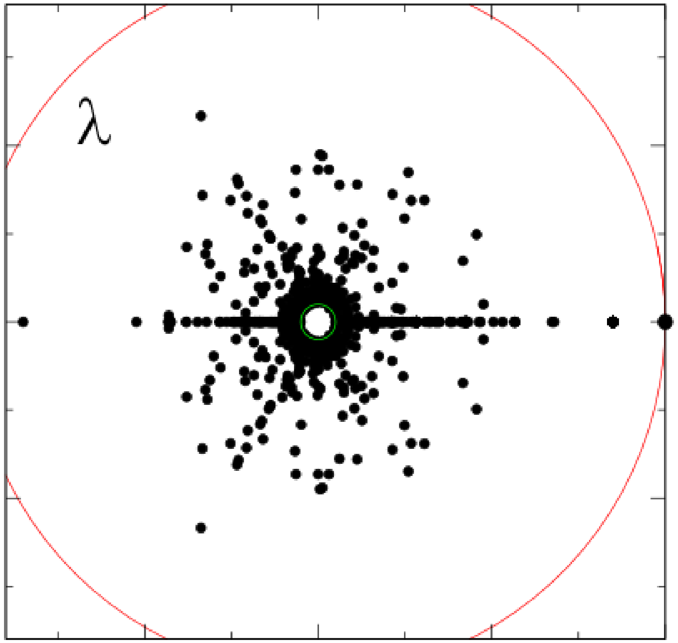
4. Filtering method to make K^* stable against manipulation

Fractal dimension of the Linux kernel

Number of eigenvalues with |λ| > 0.25, |λ| > 0.1



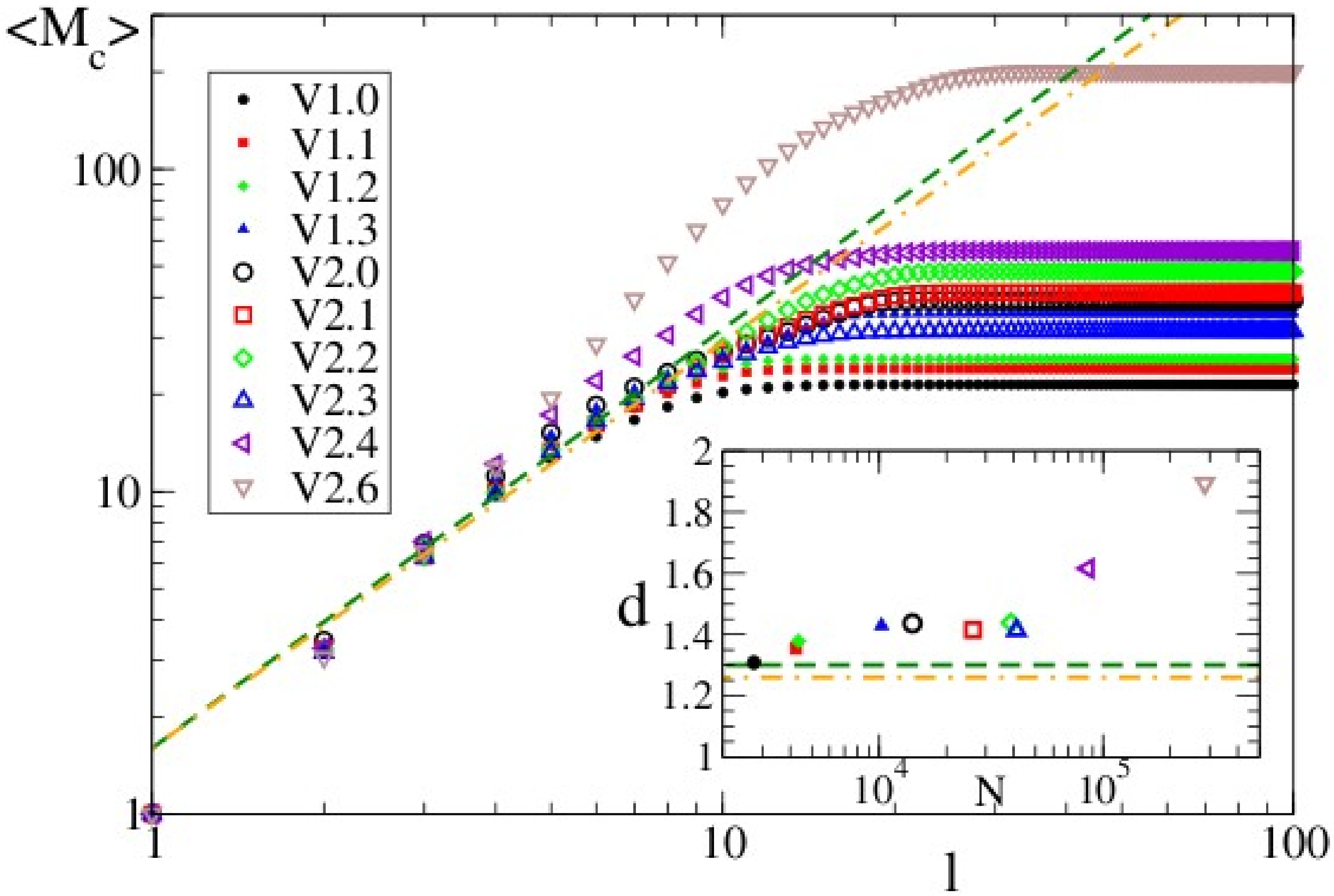
Eigvenvalues circle



Only largest |λ| computed using Arnoldi method

Procedure number in Kernel (Google matrix size)

Power law distribution of the Eigenvalues $N^\nu \simeq N^{0.65}$



Geometrical fractal dimension from cluster grwoth method

Fractal Weyl law : connection between the

the two exponents $\nu = d / 2$: fractal growth

L. Ermann, A.C., D.L. Shepelyansky (2011)

Thank you, for your attention !

Supported by EC FET Open project NADINE