A stochastic model for the link analysis of the Web

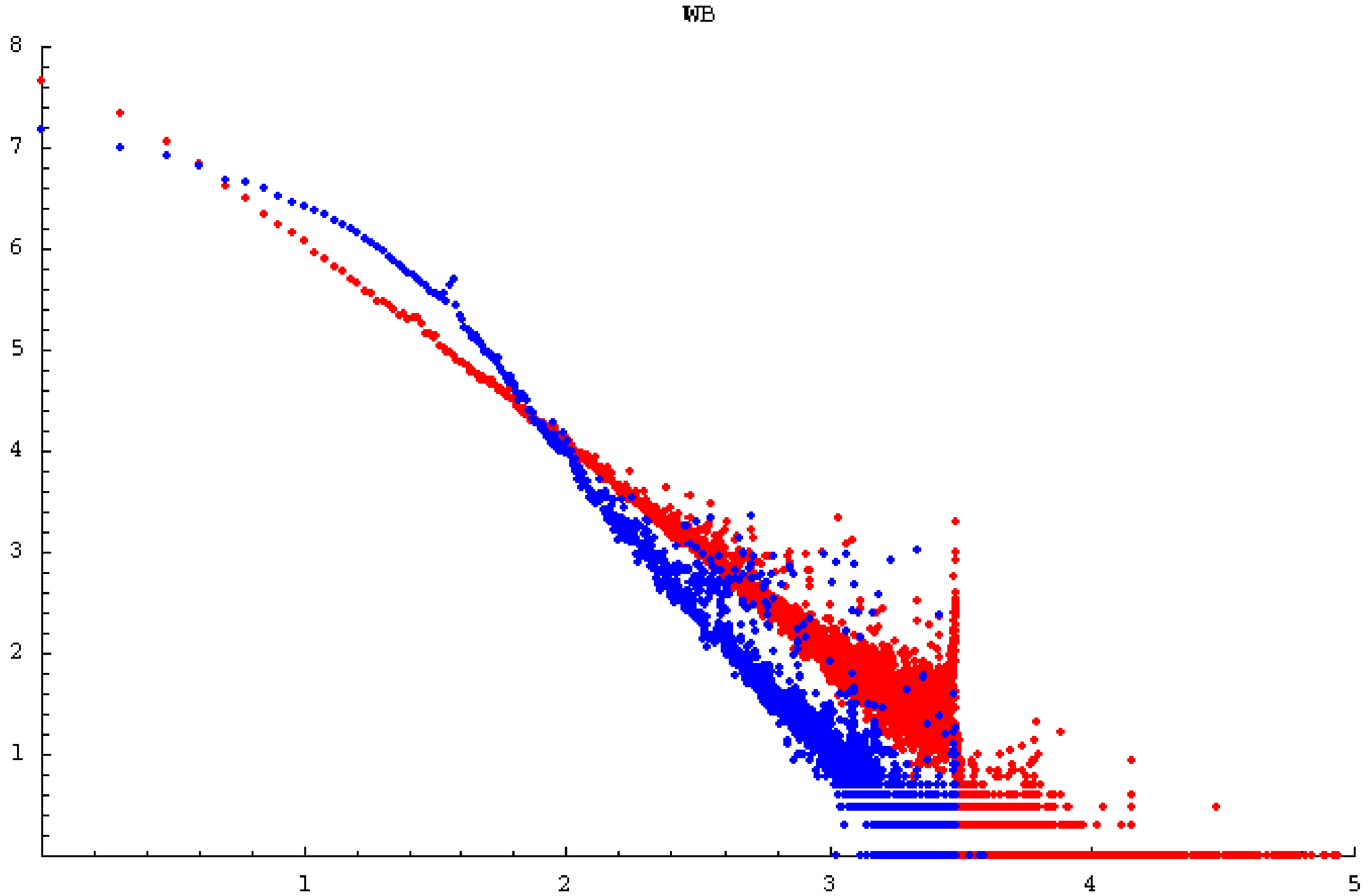
Paola Favati, IIT-CNR, Pisa.

Grazia Lotti, University of Parma.

Ornella Menchi and Francesco Romani, University of Pisa

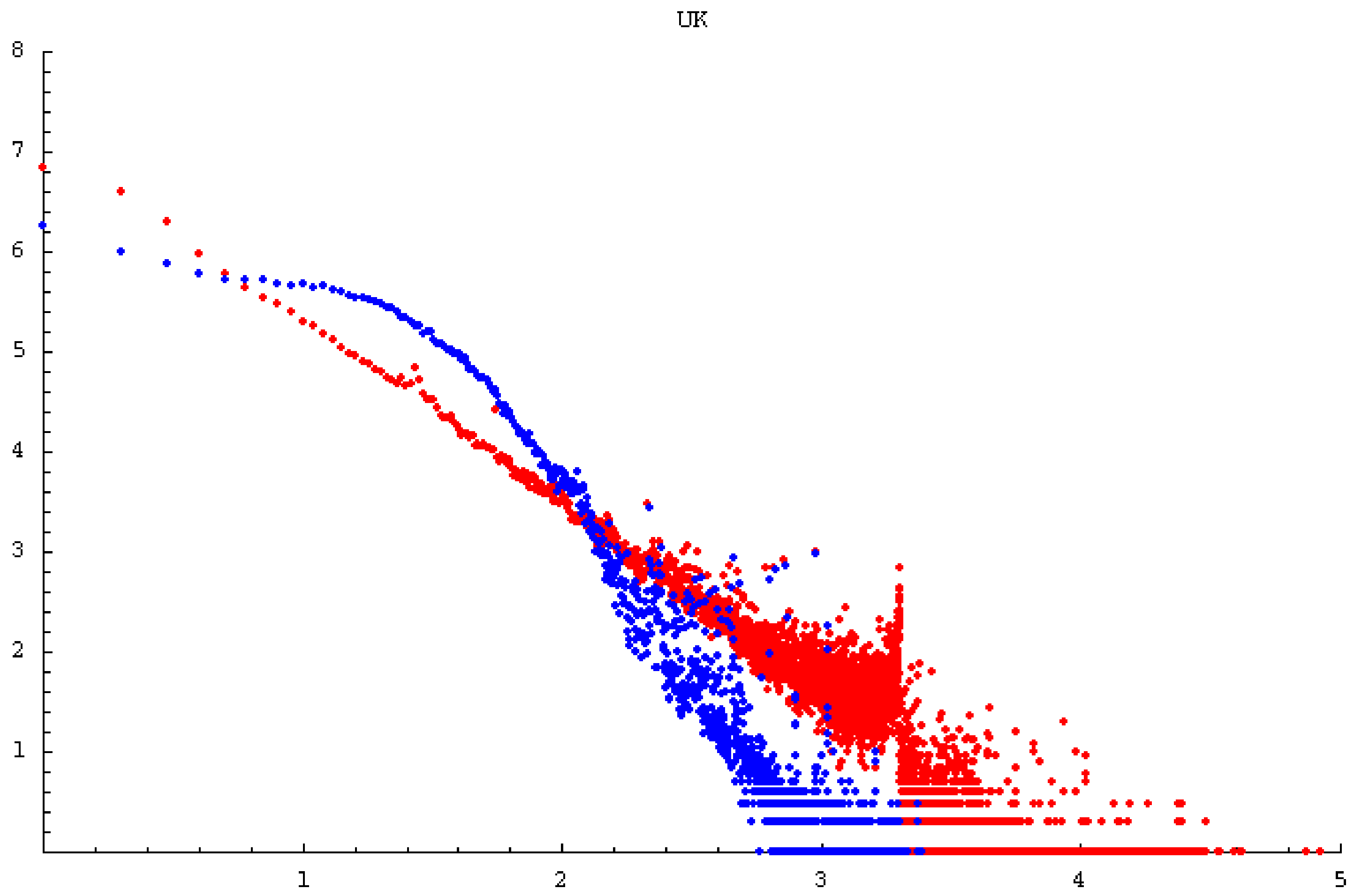
Three examples of link distribution in the web

Number of pages versus degree, Log-Log scale

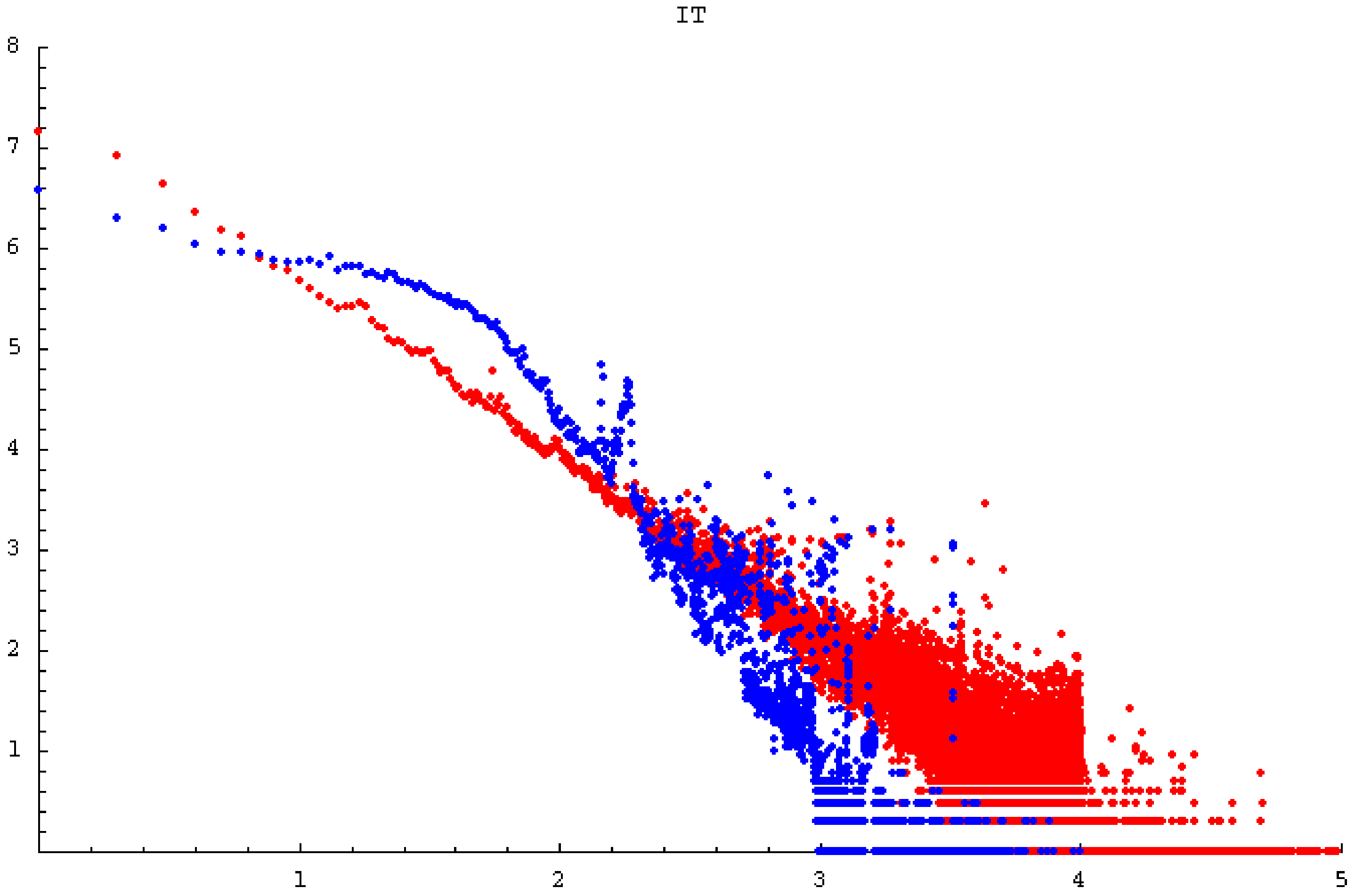


Data set WB, 118M pages and 1G links obtained by WebBase crawler
red inlink distribution, **blue outlink distribution**.
(see: `http://webgraph-data.dsi.unimi.it/`)

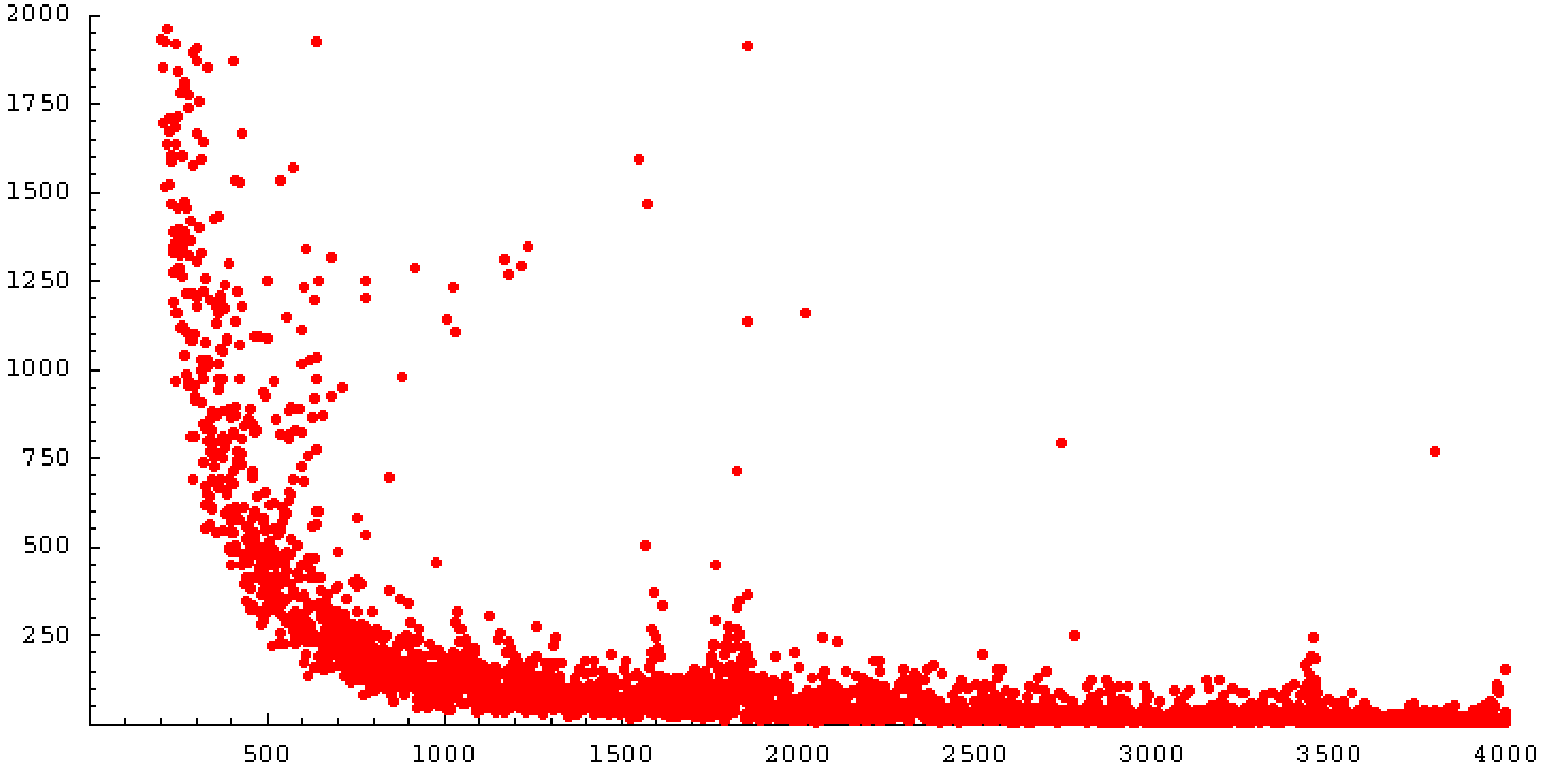
Log-Log plots are usually employed for these graphic representations since the data spans many orders of magnitude



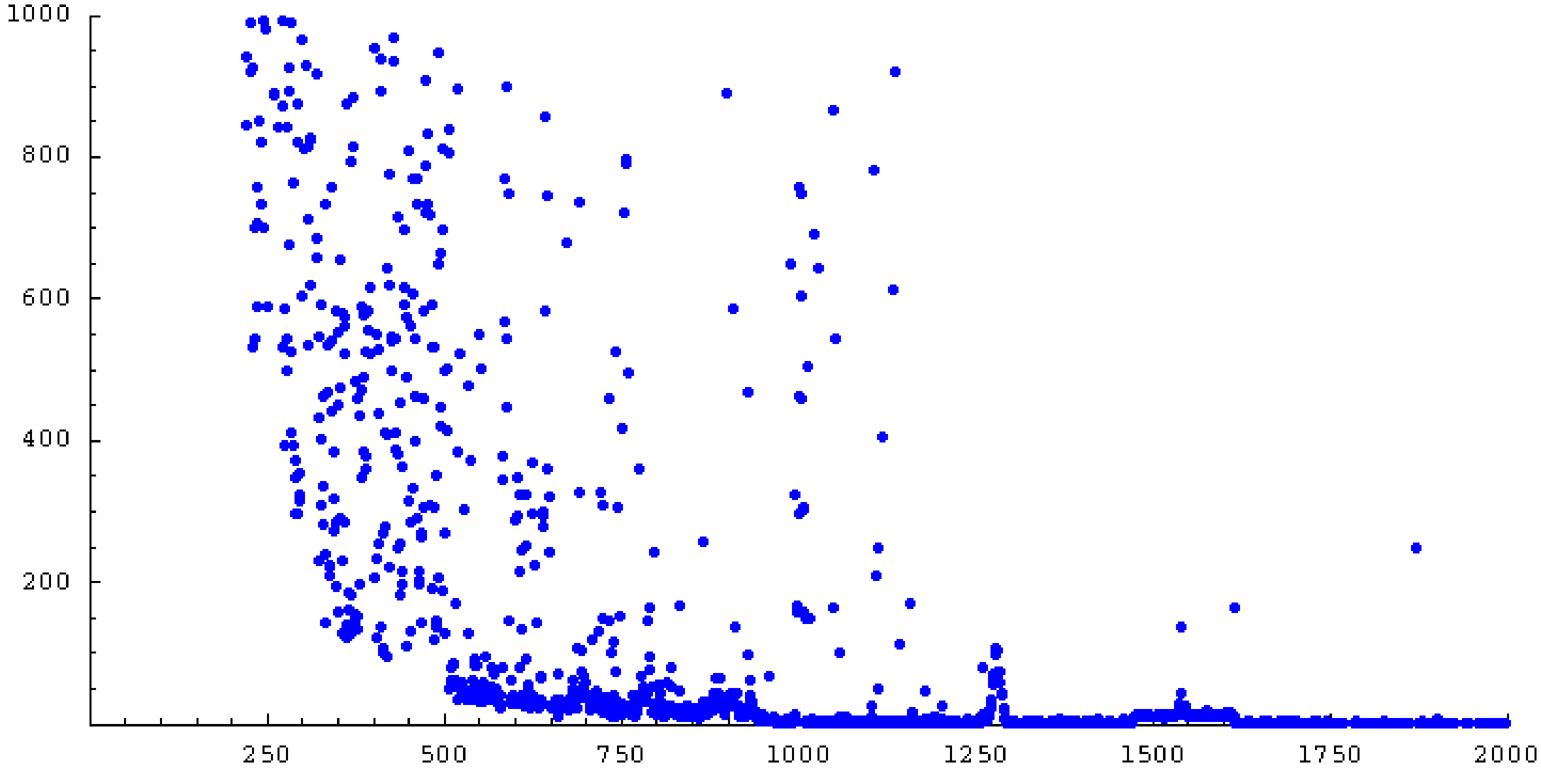
UK, 18.5M pages and 300M links of the .uk domain obtained by UbiCrawler
(`http://webgraph-data.dsi.unimi.it/`)
Log-Log scale, **red inlink distribution**, **blue outlink distribution**.



IT, 41.3M pages and 1.15G links of the .it domain obtained by UbiCrawler
Log-Log scale, **red inlink distribution**, **blue outlink distribution**.



IT Data Set, Natural scale, **inlink distribution**.

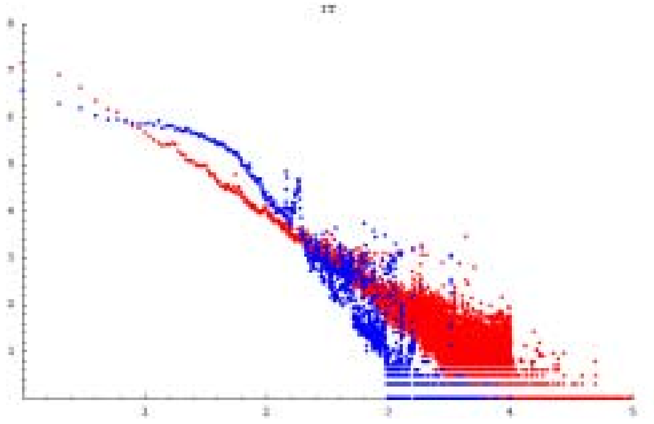


IT Data Set, Natural scale, **outlink distribution**.

Remarks

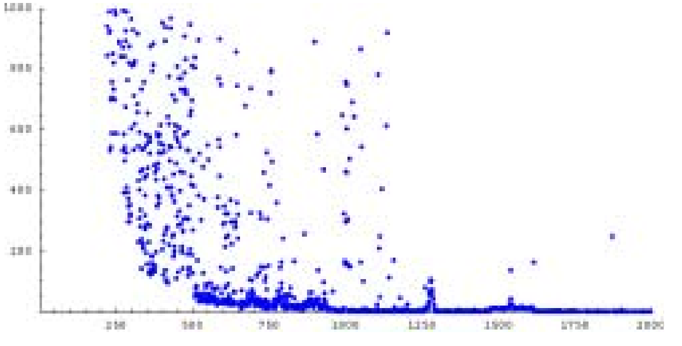
- The points have (obviously) integer coordinates.
- The link distributions do not follow a power law;

  the discrepancy is larger for outlinks than for inlinks.



- In the Log-Log scale the data appears to be spread in the tail;

  actually the dispersion occurs everywhere, as seen from the natural scale plots.



- The graphs represent small subsets of the whole Web:

  the crawler limitations influence the data.

Models of (in)link distribution

A discrete-time stochastic process is considered.

- A simple model can be based on uniform attachment:

  when a new link is created, it points to a page chosen at random.

- Models based on preferential attachment (Simon, 1955, and many others, for the web see Barabasi, Albert, 1999):

  when a new link is created, it points to a page chosen proportionally to its indegree.

- Mixed models based on uniform attachment and preferential attachment (Dorogovtsev et al. 2000, Cooper, Frieze 2001, Pennock et al. 2002, Mitzenmacher, 2003):

  models of this kind depend on some parameters, e.g.

  number of new links generated at each time step,

  probability of pointing a new page instead of an existing one,

  probability of choosing uniform instead of preferential policy in the mixed model.

Similar models can be devised for outlink distribution.

Our choice of model for both inlinks and outlinks

We adopt a mixed model.

At any time step ONE new link is created

with probability α connected to a new page;

with probability $1-\alpha$ connected with an already existing page,

with probability β chosen at random,

with probability $1-\beta$ chosen proportionally to the degree of that page.

Equation of the model

Let $X_j^{(t)}$ be the number of pages having degree j at time t.

The expected value of the variation of $X_j^{(t+1)}$ with respect to $X_j^{(t)}$ is

$$\mathcal{E}\left[X_j^{(t+1)} - X_j^{(t)}\right] = p(j-1,t) - p(j,t), \quad j = 2,\ldots,t,$$

where

$$p(j,t) = (1-\alpha)\left[\frac{\beta}{n(t)}X_j^{(t)} + \frac{1-\beta}{t}\,j\,X_j^{(t)}\right]$$

is the probability that the new link is connected with a page having degree j and n(t) is the number of existing pages at time t.

The equation holds also for j = 1 and for j = t+1 provided that

$$p(0,t) = \alpha \quad \text{and} \quad X_{t+1}^{(t)} = 0$$

Solution of the model

The model can be solved as a difference equation by replacing the expected values by the actual ones.

The steady state solution is given in terms of the Beta function: B(a,b) = Γ(a) Γ(b)/Γ(a + b).

$$S_j^{(t)} = c\,B\big(\sigma + j, \rho + 1\big)$$

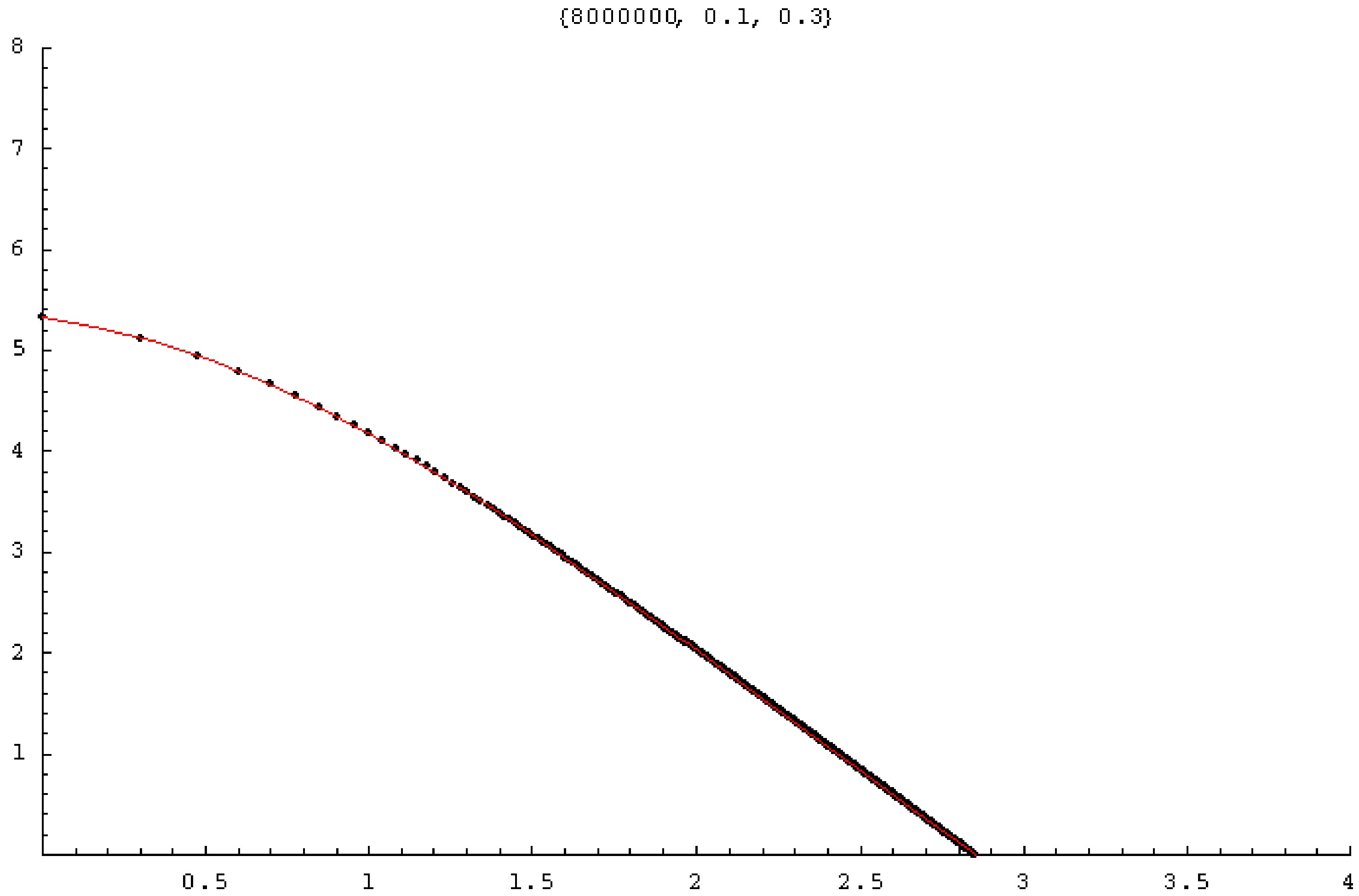
where

$$c = \frac{\alpha\rho\,t}{(\sigma+\rho+1)\,B\big(\sigma+1, \rho+1\big)},$$

$$\sigma = \frac{\beta}{\alpha(1-\beta)}, \quad \rho = \frac{1}{(1-\alpha)(1-\beta)}.$$

Example of simulation

The difference equation can be used recursively to generate a deterministic solution.



Example of simulation, t=8000000, $\alpha = 0.1$, $\beta = 0.3$

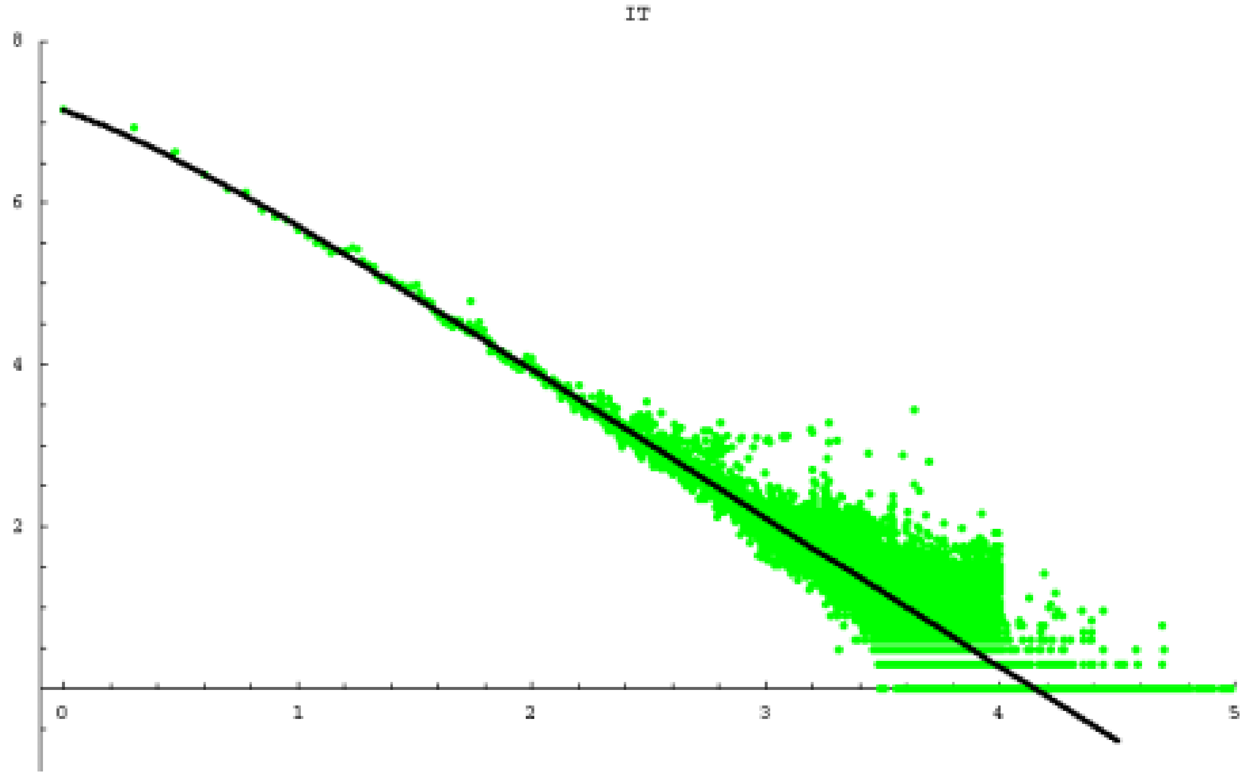
black: the discrete points X obtained by the simulation,

red: the continuous approximation S.

Deriving the parameters of the model from the experimental data

A continuous monotonic function S can hardly be used to represent integer spread data.

S_j could be viewed as the expected value of an integer random variable P_j and the parameters of the model could be determined by a Least Squares fit of the experimental data with a Beta function in Log-Log space.

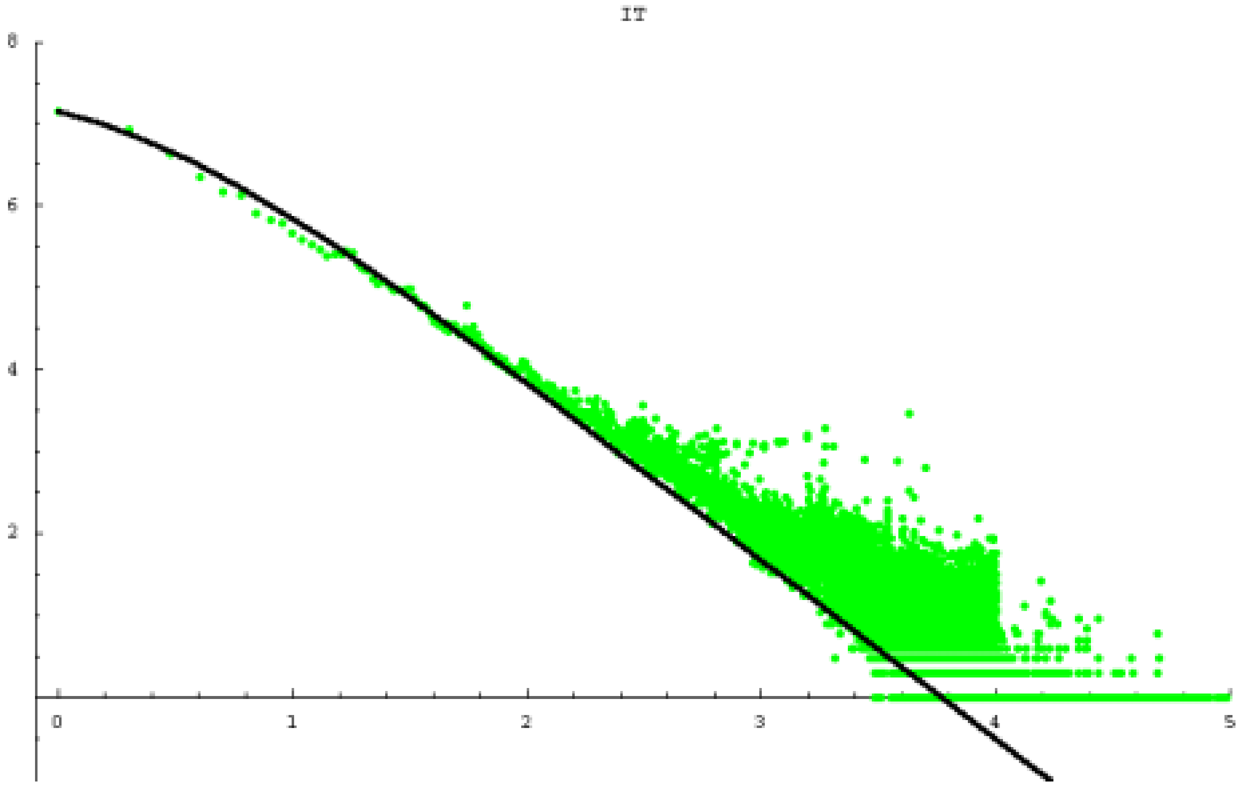


Least Squares fit of **inlink distribution** for data set IT

Unfortunately this approach may produce negative values of α and β not compatible with the model, e.g. in this example $\alpha = -0.12$, $\beta = -0.06$.

Numerical experiments suggest a different approach

We approximate the lower envelope of the data.



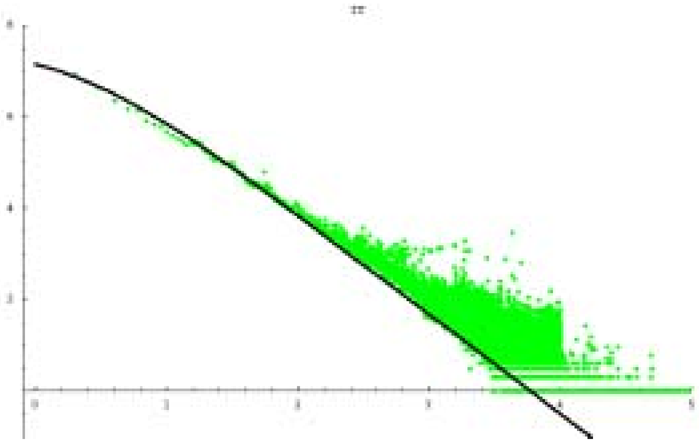
Least Squares fit of lower envelope of the **inlink distribution** for data set IT,

In this example $\alpha = 0.07$, $\beta = 0.09$.

Integer Approximation

To give a motivation to the use of the lower envelope, we conjecture that the integer values of the data are generated as the rounded sum of two terms:

- the values of the continuous approximation S_j ,
- the realizations of an integer nonnegative random variable ξ_j with probability function $p_j(n)$ decreasing with respect to n.



In our experiments we choose a geometric random variable with mean $S_j^{0.75}$, i.e.

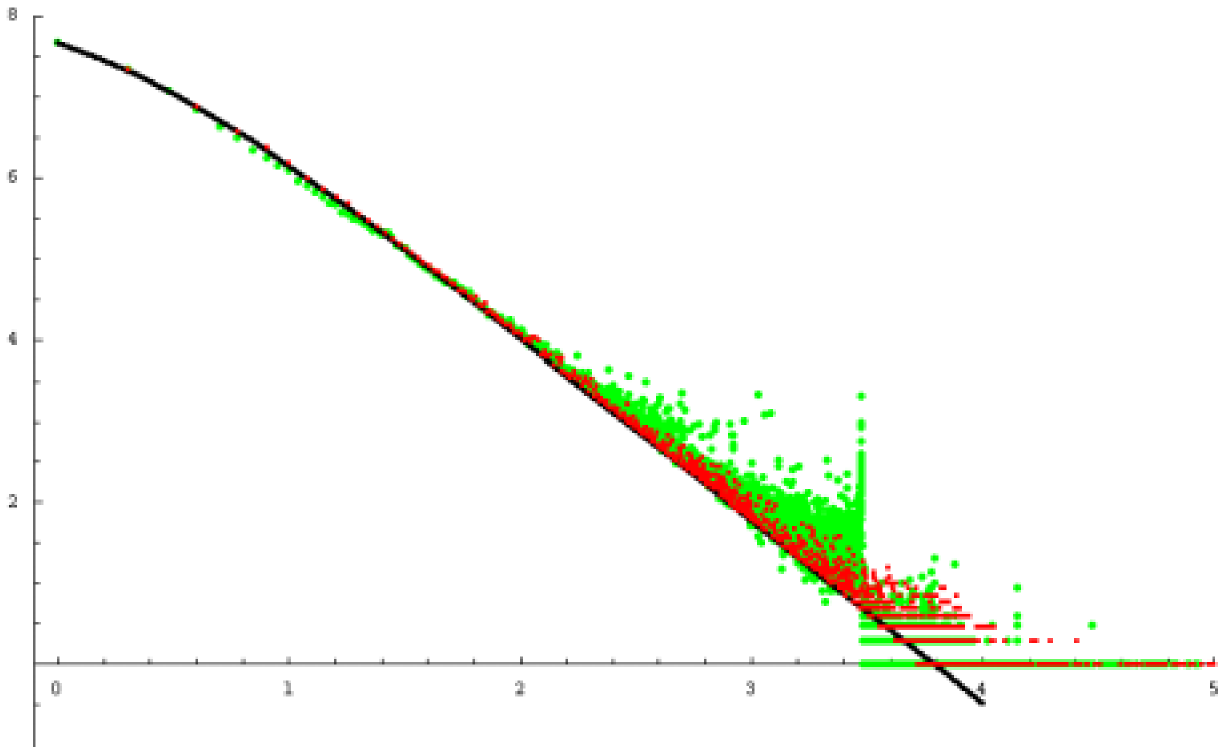
$$p_j(n) = \tau_j (1-\tau_j)^{n-1}, \quad \text{where} \quad \tau_j = 1/(1+ S_j^{0.75})$$

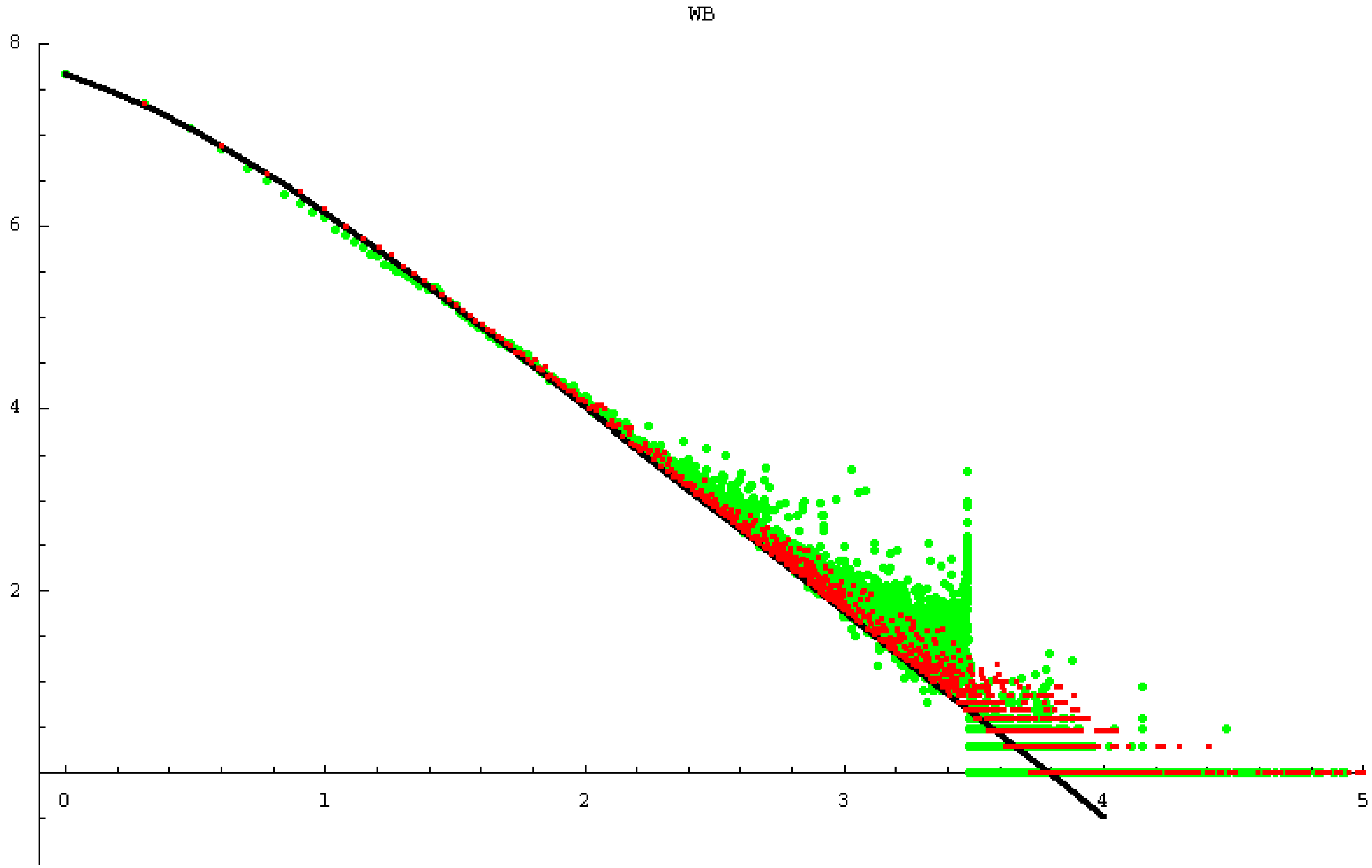
Fitting the experimental data

In the following six plots we see together:

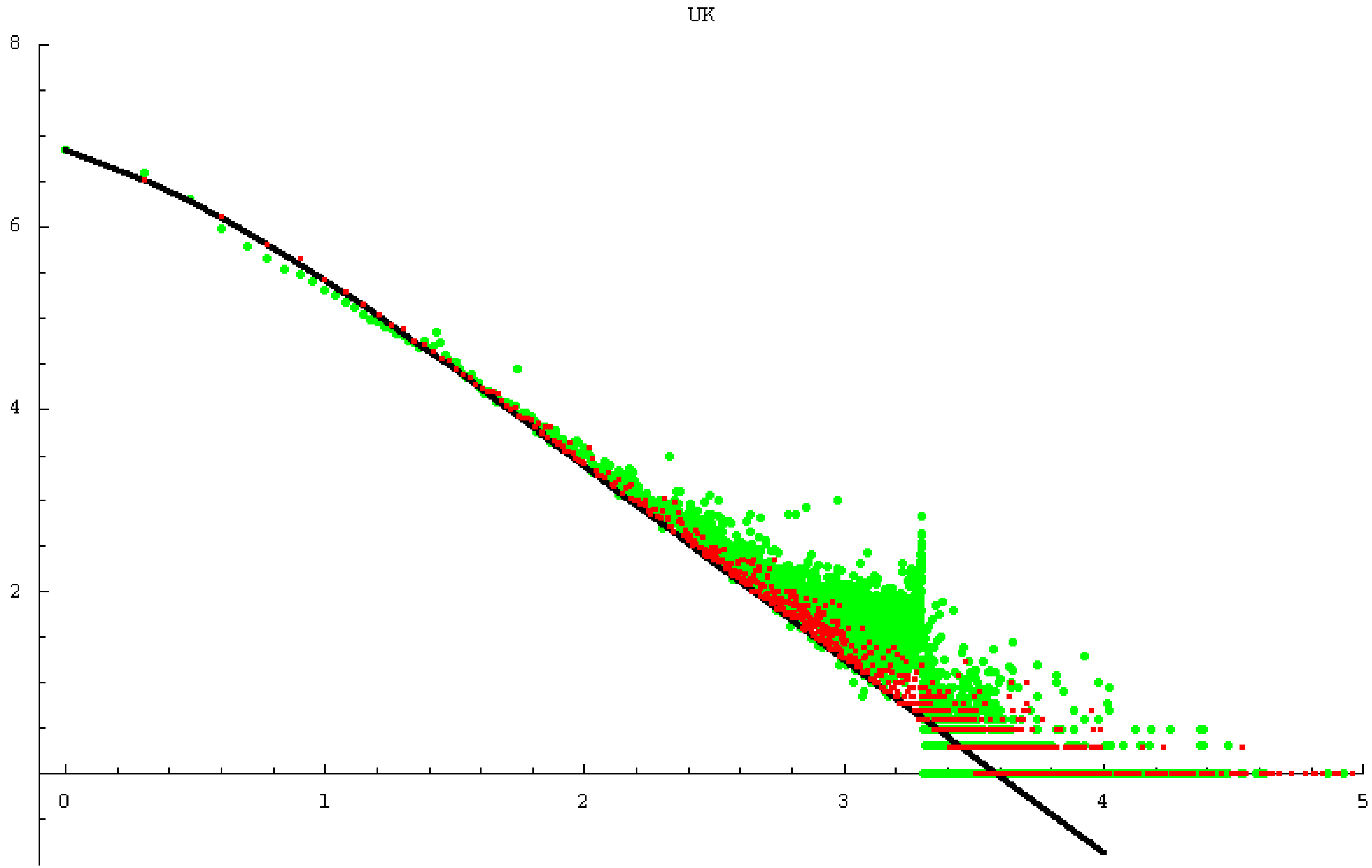
- the crawler data (**green points**)
- the continuous approximation S_j (**black line**) fitted on lower envelope
- the integer approximation (**red points**)

e.g.

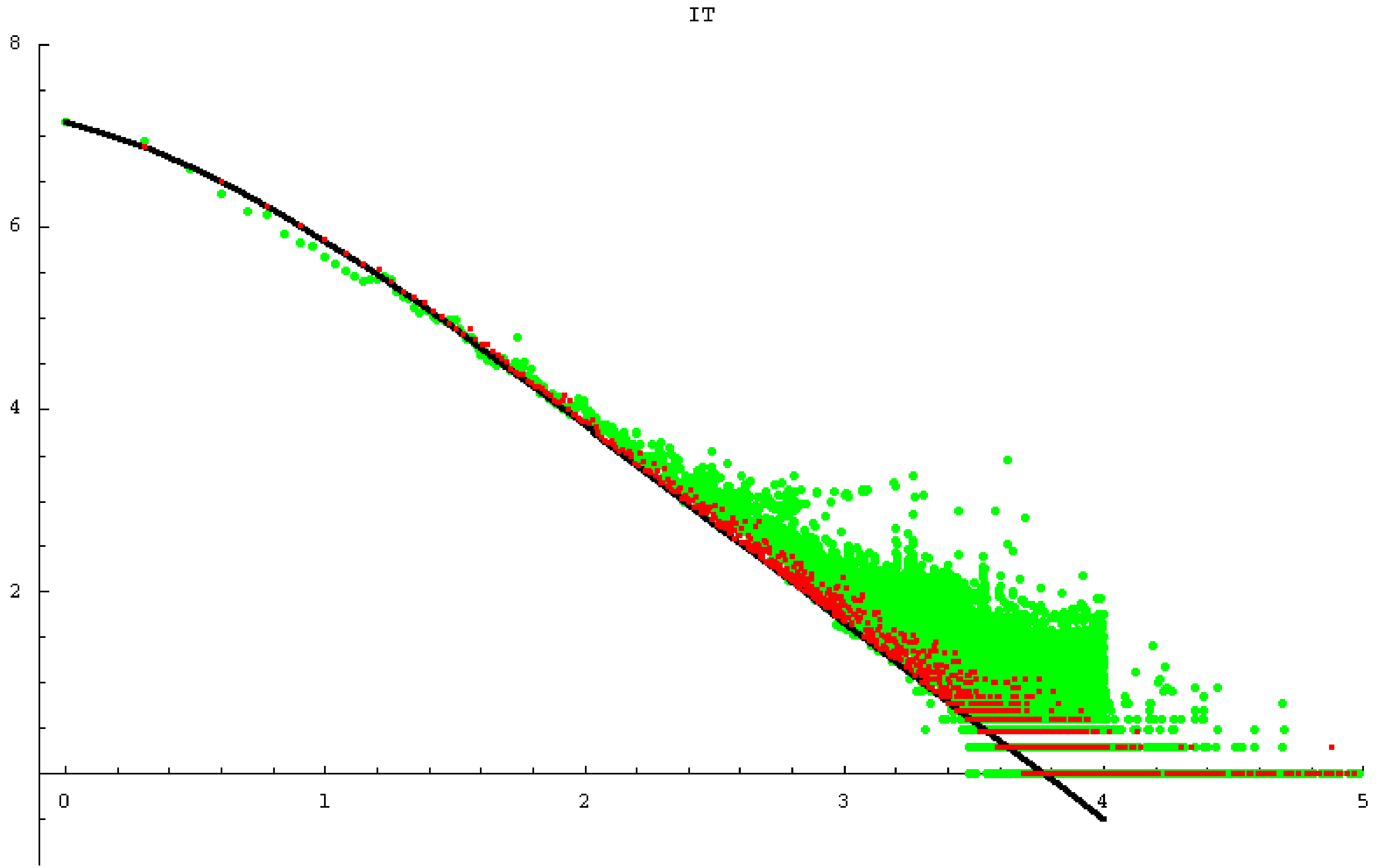




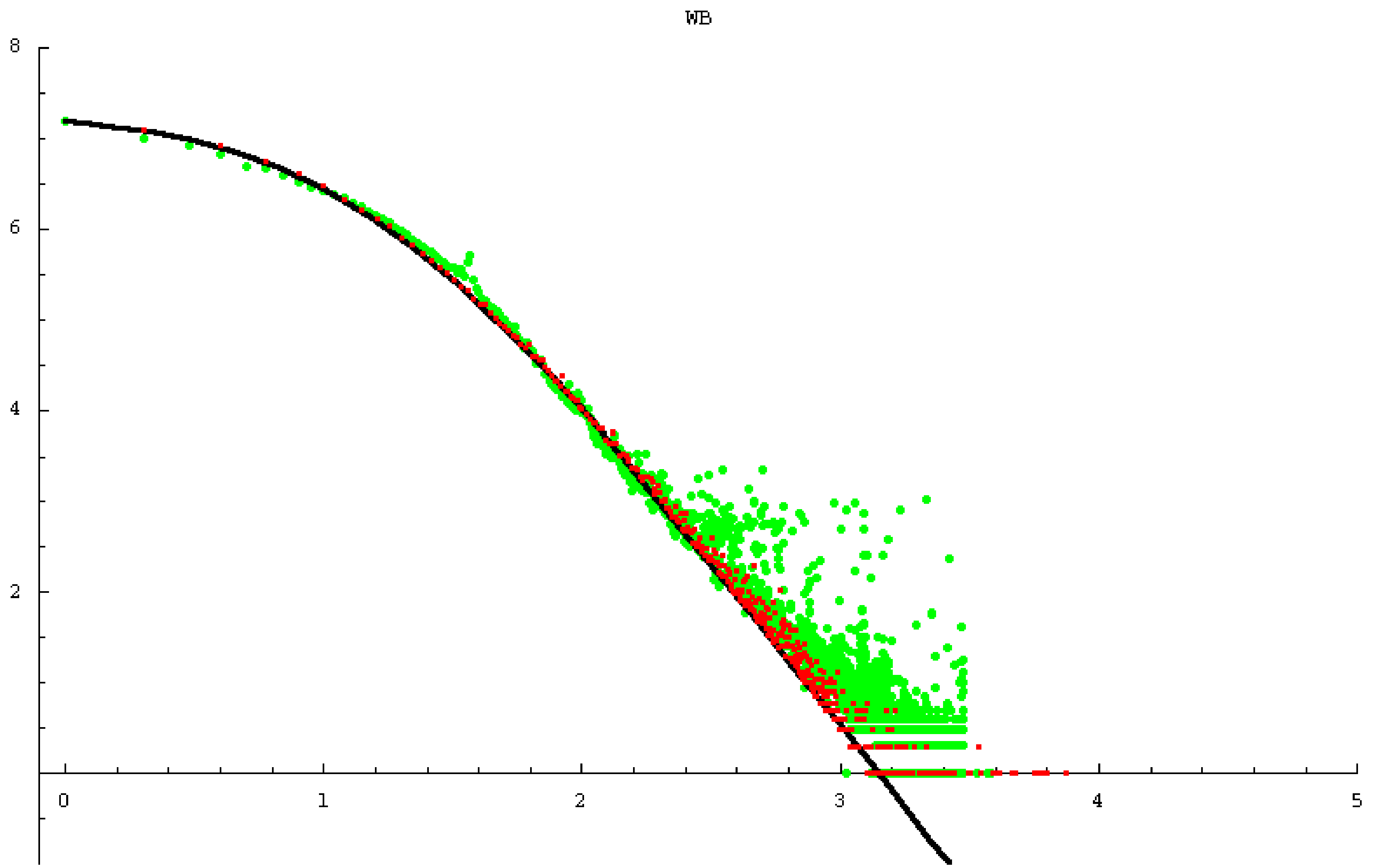
Approximation of **inlink distribution** for data set WB,
black continuous approximation, **red integer approximation**
$\alpha = 0.12$, $\beta=0.10$



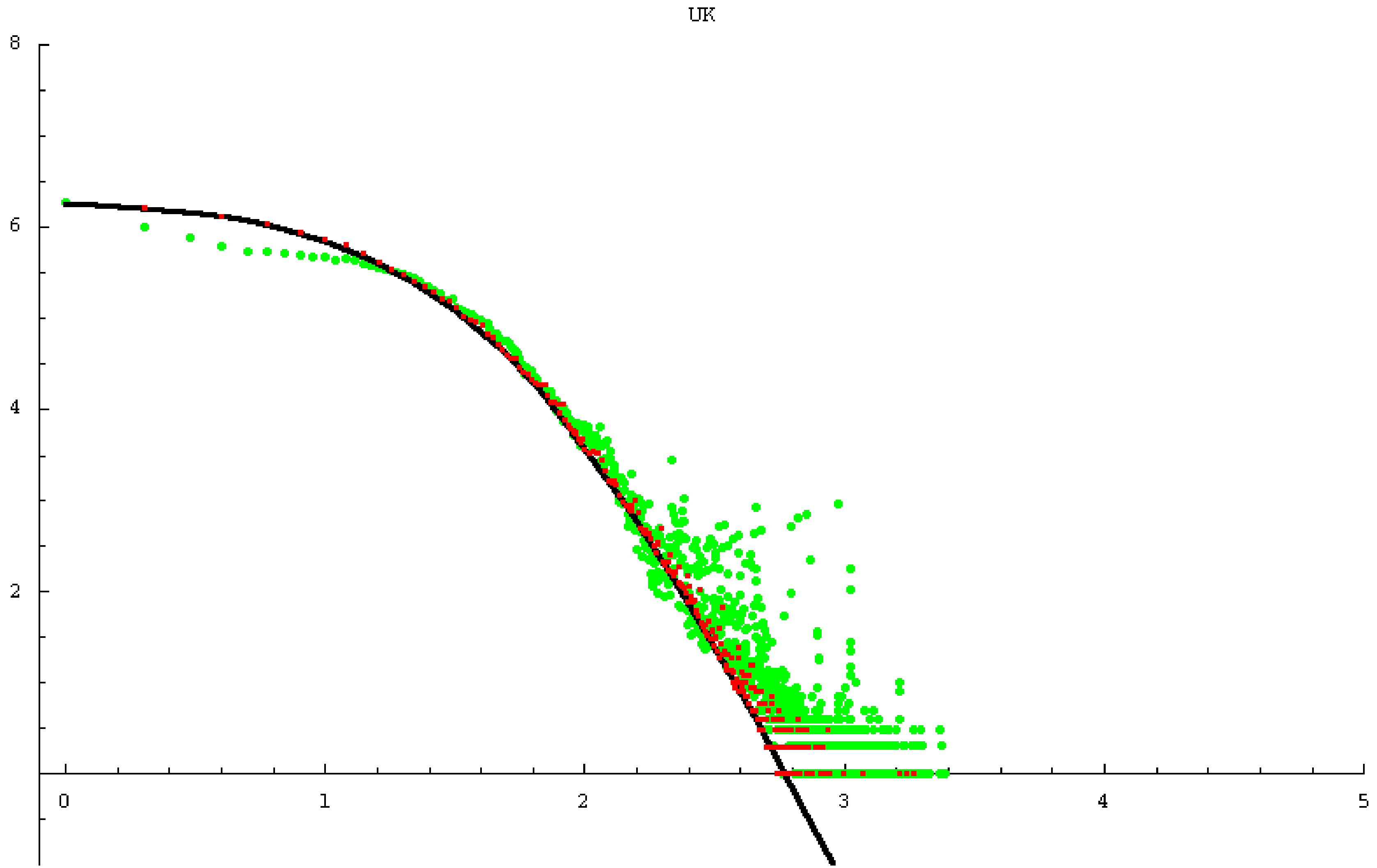
Approximation of **inlink distribution** for data set UK,
black continuous approximation, **red integer approximation**
$\alpha = 0.07$, $\beta = 0.06$



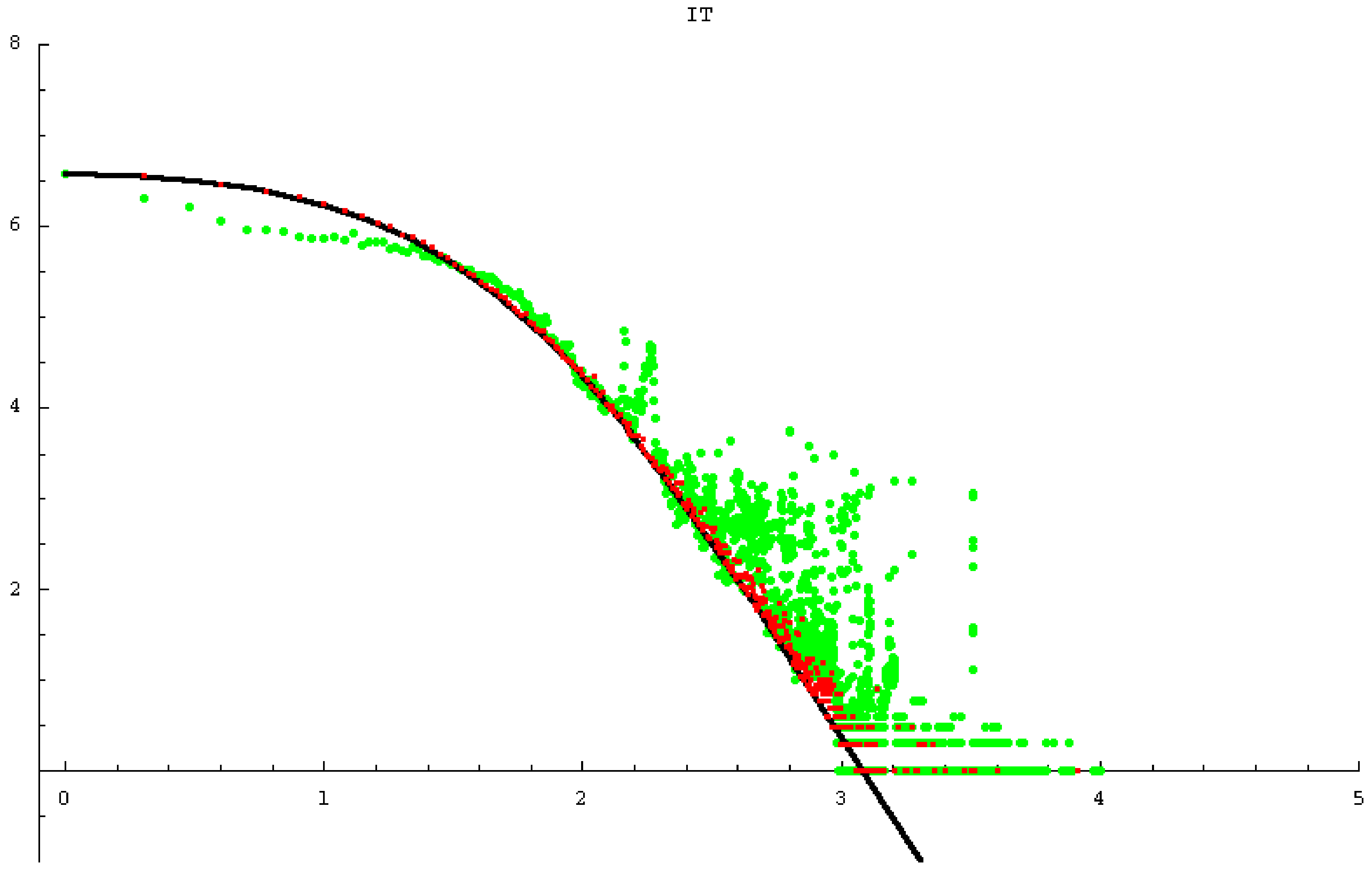
Approximation of **inlink distribution** for data set IT,
black continuous approximation, **red integer approximation**
α = 0.07, β=0.09



Approximation of **outlink distribution** for data set WB,
black continuous approximation, **red integer approximation**
$\alpha = 0.11$, $\beta=0.58$



Approximation of **outlink distribution** for data set UK,
black continuous approximation, **red integer approximation**
$\alpha = 0.07$, $\beta=0.77$



Approximation of **outlink distribution** for data set IT,
black continuous approximation, **red integer approximation**
$\alpha = 0.05$, $\beta=0.71$

Summary of the estimated values for α and β

		α	β
Inlink	WB	0.12	0.10
	UK	0.07	0.06
	IT	0.07	0.09
Outlink	WB	0.11	0.58
	UK	0.07	0.77
	IT	0.05	0.71

The values of β are much smaller for the inlinks than for the outlinks. This means that the preferential attachment is the dominant policy in the inlink distribution, while the outlink distribution appears to be significantly ruled by the uniform attachment.

Another possible approach: fitting of the reversed data

- The inverse of a Beta function is well approximated by a Yule function

$$j(y) = c\, b^{y}\, y^{-r}, \quad \text{where } c, b, r \text{ are suitable parameters}$$

- The approximation with a Yule function, in the Log-Log space, can easily be computed with a Linear Least Squares Fit.

- From the values of c, b, r one can derive approximations to the parameters of the mixed model.

The numerical experiments give results to those ecposed above.

Research problem:

Find a model in the reverse coordinates space which as solution produces a Yule function

References

A.L.Barabasi, R. Albert, Emergence of scaling in random networks, Science, 286, pages 509-512, 1999.

A.L.Barabasi, R. Albert, H. Jeong, Mean-field theory for scale-free random networks, Physica A, 272, pages 173-187, 1999.

A. Broder, R. Kumar, F. Maghoul, P. Prabhakar, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, Graph structure in the web. Proceedings of the Nineth International World Wide Web Conference, 2000.

C. Cooper, A. M. Frieze, A general model of undirected web graphs, Proceedings of the Nineth Annual European Symposium on Algorithms, LNCS n.2161 , pages 500-511.

S. Dorogovtsev, J. Mendes, A. Samukhin, Structure of Growing Networks: Exact Solution of the Barabasi-Albert's model, Phys. Rev. Lett. 85, pages 4633-4636, 2000.

F. Menczer, Growing and navigating the small world Web by local content. Proceedings of the National Academy of Science, 99, pages 14014-14019, 2002.

M. Mitzenmacher, A Brief History of Generative Models for Power Law and Lognormal Distributions. Internet Mathematics, 1, pages 226-251, 2003.

D.M. Pennock, G.W. Flake, S. Lawrence, E.J. Glover, C.L. Giles, Winners don't take all: Characterizing the competition for links on the web. Proceedings of the National Academy of Science, 99, pages 5207-5211, 2002.

H.A. Simon, On a Class of Skew Distribution Functions. Biometrika, 42, pages 425-440, 1955.