Analysis of growing graphs as a way to identify their fundamental properties

Work in collaboration with Matthieu Latapy

Clémence Magnien
magnien@shs.polytechnique.fr

CREA - CNRS - École Polytechnique
Laboratoire J.-V. Poncelet - CNRS - HMY

Outline

Outline

Typical examples

- The Web graph [KRRT99, AJB99]
  Web pages linked by hyperlinks
  Data research, . . .
  Exploration: Crawl

Typical examples

- The Web graph [KRRT99, AJB99]
- Peer-to-Peer exchanges [Le Fessant 2004, GLS2004]
  Users linked by the exchange of data
  Design of efficient architectures, . . .
  Exploration: Capture from a server

Typical examples

- The Web graph [KRRT99, AJB99]
- Peer-to-Peer exchanges [Le Fessant 2004, GLS2004]
- Internet topology [FFF99, AB99]
  Routers linked by cables
  Protocol design, . . .
  Exploration: `traceroute`

Typical examples

- The Web graph [KRRT99, AJB99]
- Peer-to-Peer exchanges [Le Fessant 2004, GLS2004]
- Internet topology [FFF99, AB99]
- IP exchanges [Soule *et al*, 2005]
  Computers linked by the exchange of IP packets
  Usage analysis, . . .
  Exploration: Capture from a router

Typical examples

- The Web graph [KRRT99, AJB99]
- Peer-to-Peer exchanges [Le Fessant 2004, GLS2004]
- Internet topology [FFF99, AB99]
- IP exchanges [Soule *et al*, 2005]

Difficulty of capture:

- Access to the data
- Various technical constraints
- Size

The data set

Four data sets coming from different contexts

- Peer-to-Peer exchanges: e-Donkey trace of 16 hours (P2P)
  - 835,000 nodes
  - 2,800,000 links
- Web: crawl launched from French labs for 4 hours (Web)
  - 700,000 nodes
  - 1,900,000 links
- Internet: traceroutes from one source to 60,000 destinations (Inet)
  - 114,000 nodes
  - 118,000 links
- IP exchanges: 1 week traffic in a large French lab (IP)
  - 470,000 nodes
  - 1,700,000 links

Current situation

Get as much data as possible
(not possible to get all the data)

Consider it as representative of the whole

Current situation

Get as much data as possible
(not possible to get all the data)

Consider it as representative of the whole

Non trivial properties in common

Usual assumptions

- Average degree
  Degree of a node: its number of links

Usual assumptions

- Average degree constant

Usual assumptions

- Average degree constant
- Density
  Number of links / number of possible links

Usual assumptions

- Average degree constant
- Density tends to 0

Usual assumptions

- Average degree constant
- Density tends to 0
- Degree distribution
  p_k: number of nodes with degree k

Usual assumptions

- Average degree constant
- Density tends to 0
- Degree distribution heterogeneous, stable

Usual assumptions

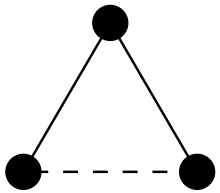
- Average degree constant
- Density tends to 0
- Degree distribution heterogeneous, stable
- Average distance and diameter
  - Distance between two nodes: Minimal number of links to go from one node to the other
  - Diameter: Largest distance in the graph

Usual assumptions

- Average degree constant
- Density tends to 0
- Degree distribution heterogeneous, stable
- Average distance and diameter small, increases

Usual assumptions

- Average degree constant
- Density tends to 0
- Degree distribution heterogeneous, stable
- Average distance and diameter small, increases
- Clustering



Usual assumptions

- Average degree constant
- Density tends to 0
- Degree distribution heterogeneous, stable
- Average distance and diameter small, increases
- Clustering large, constant

Methodology

Coarse-grained graphs:

$\longrightarrow$ divide the sample in slices (time, size, ...)

Observe the properties of the sample at the end of these slices

Look for stable properties

$\longrightarrow$ check the usual assumptions

Challenges:

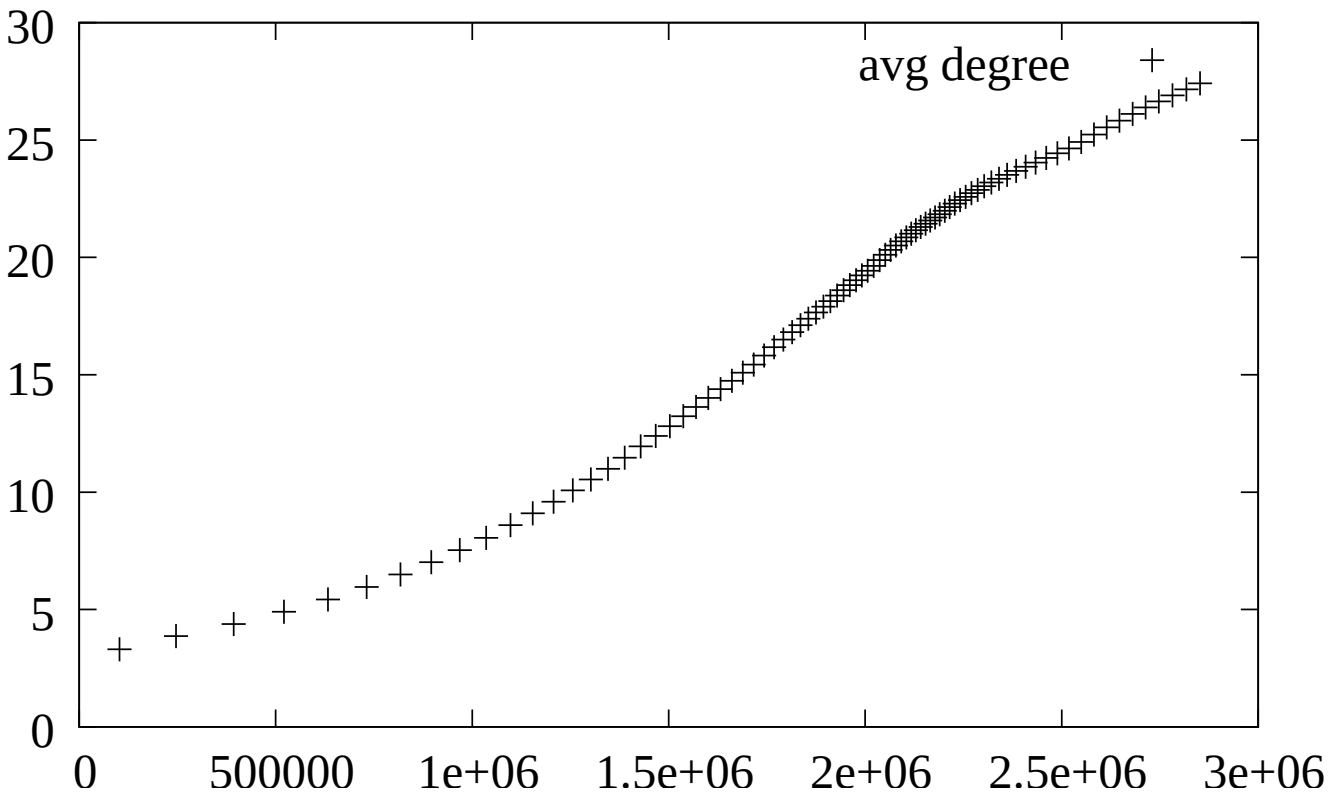
- Data acquisition, dynamics
- Size of data: use of heuristics

Outline

Average degree

Average degree: $2M/N$

Usual assumption: constant



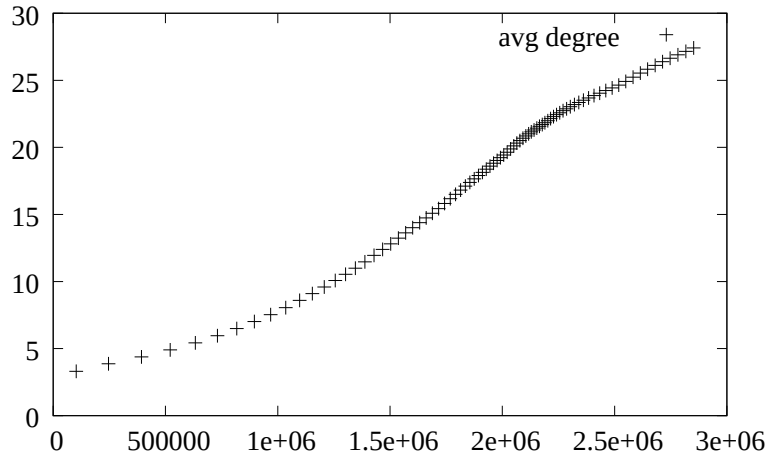
P2P

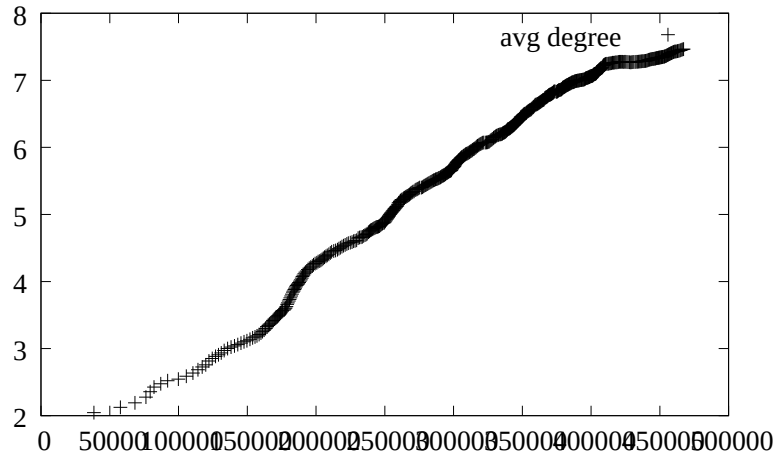
Average degree

Average degree: $2M/N$

Usual assumption: constant

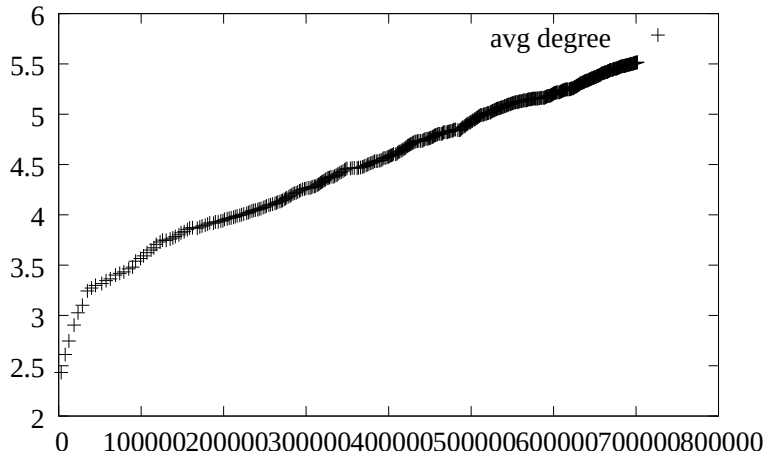
P2P

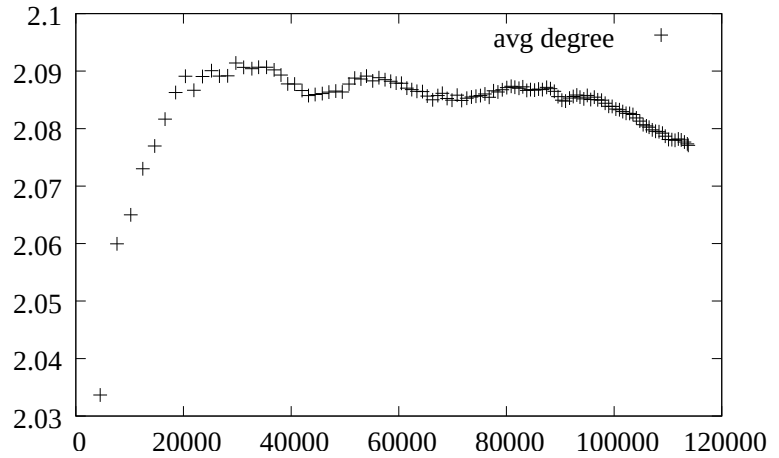




IP

Web

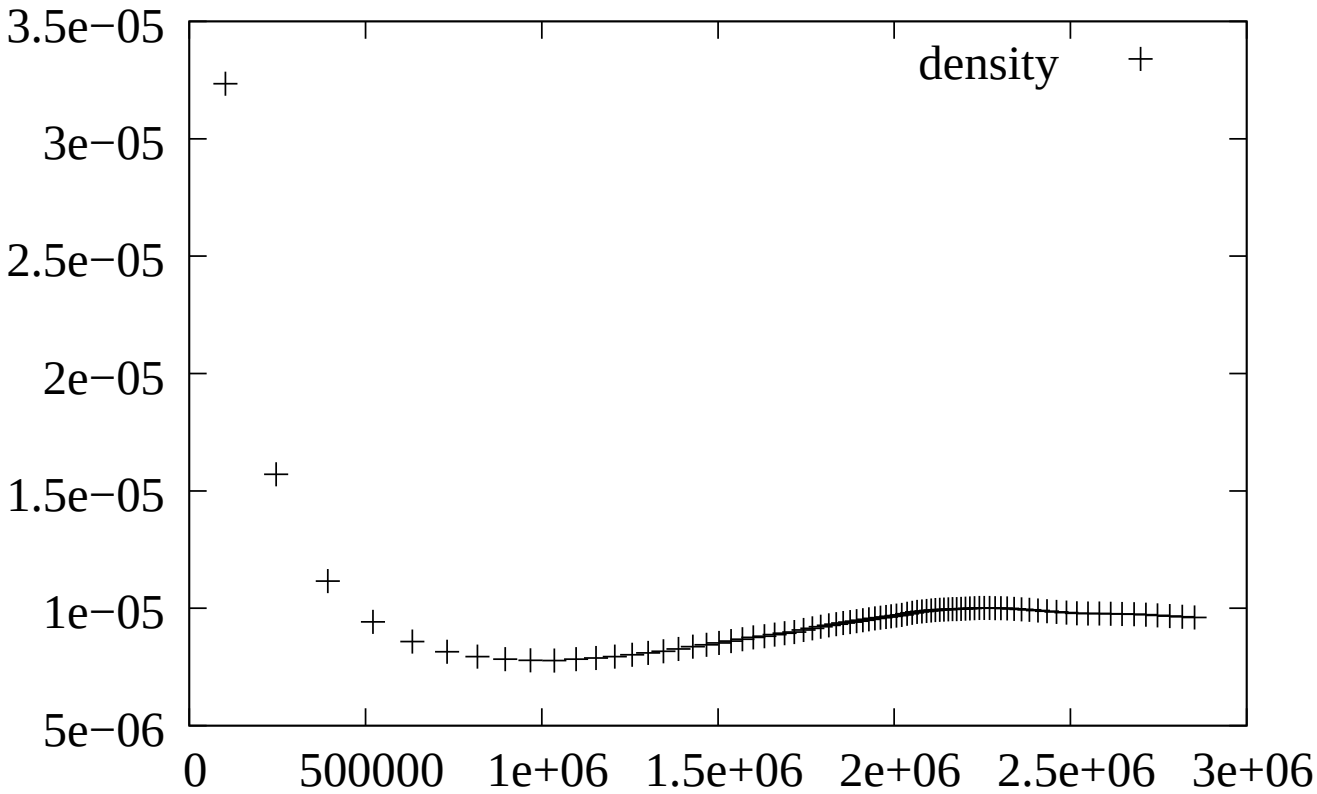




Inet

Density

Density: $2M/N(N-1)$
Usual: tends to 0

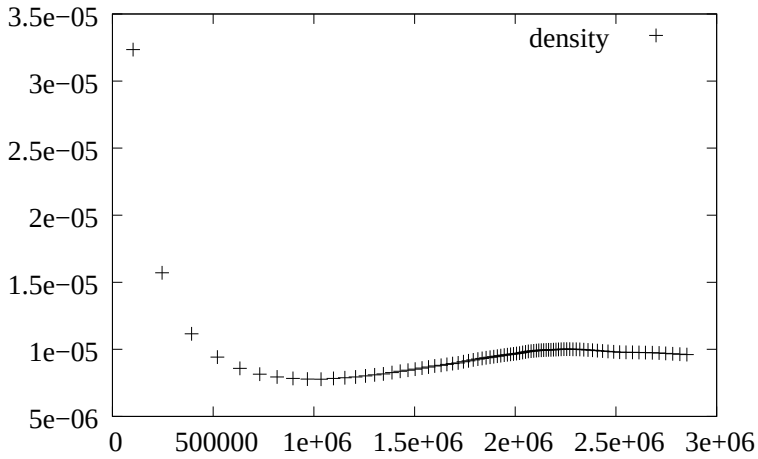


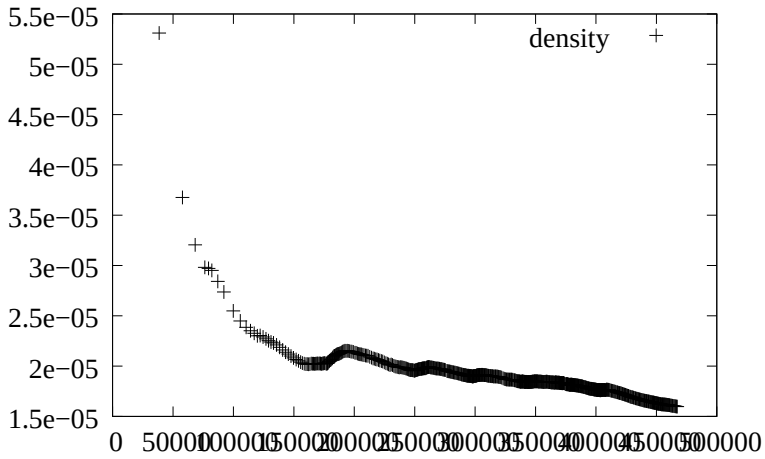
P2P

Density

Density: $2M/N(N-1)$
Usual: tends to 0

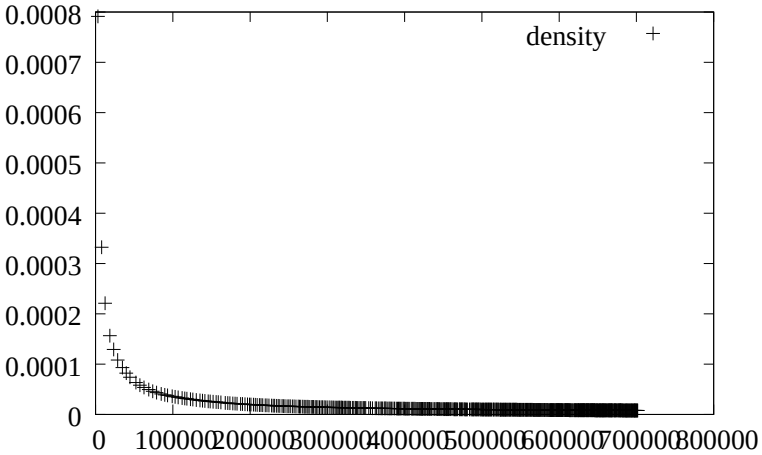
P2P

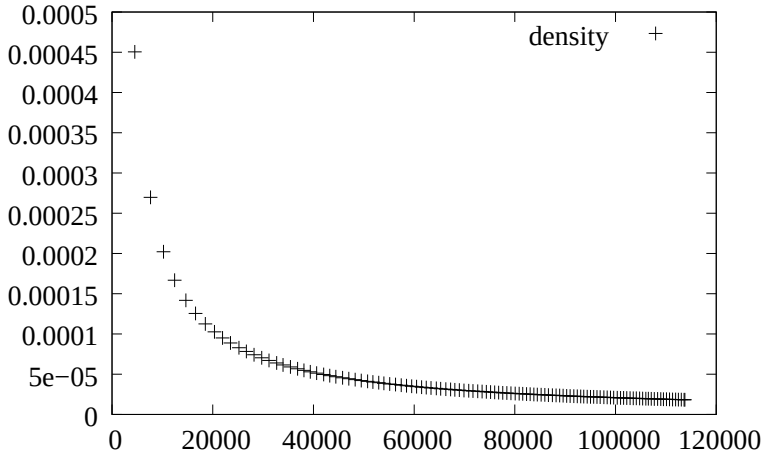




IP

Web





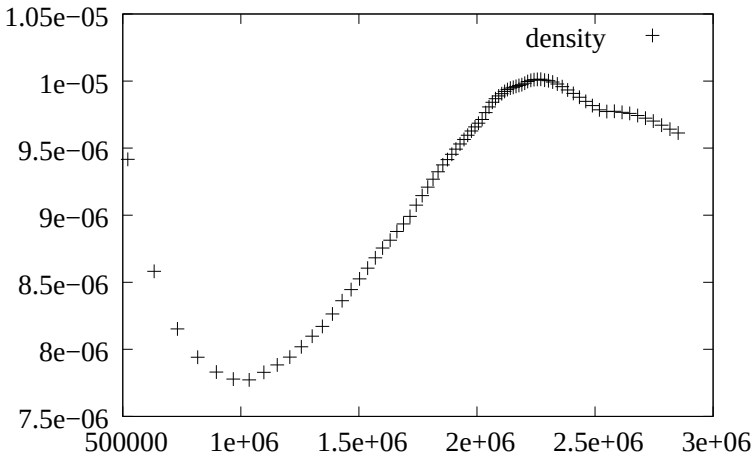
Inet

Density

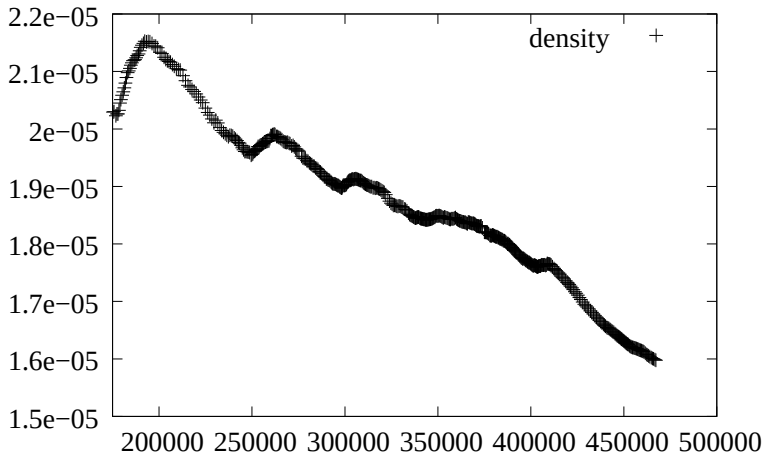
Density: $2M/N(N-1)$

Usual: tends to 0

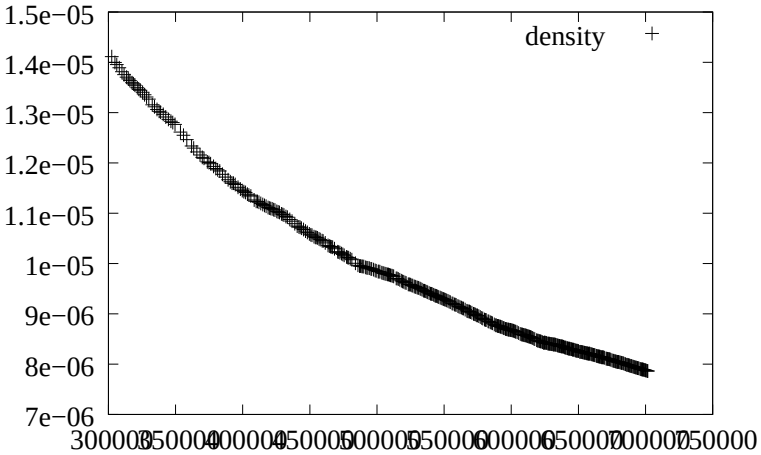
P2P



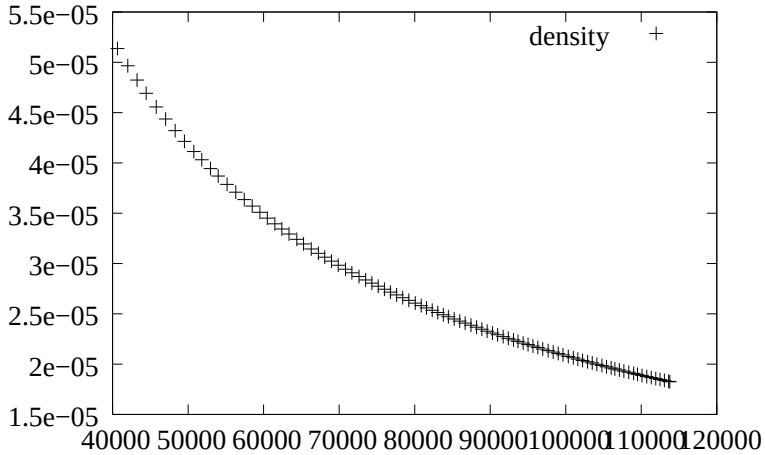
IP



Web

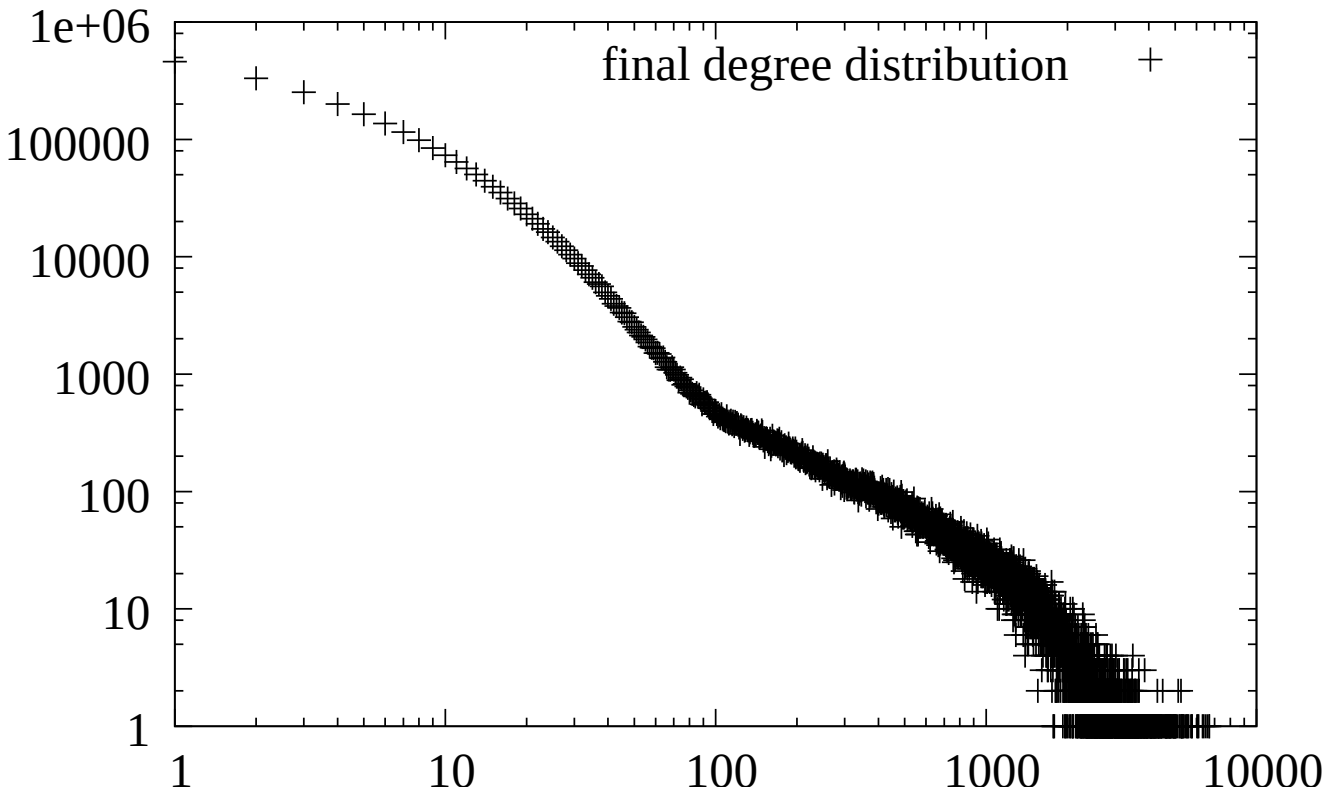


Inet



Degree distribution

p_k: number of nodes of nodes with degree k.
Usual assumption: heterogeneous, stable

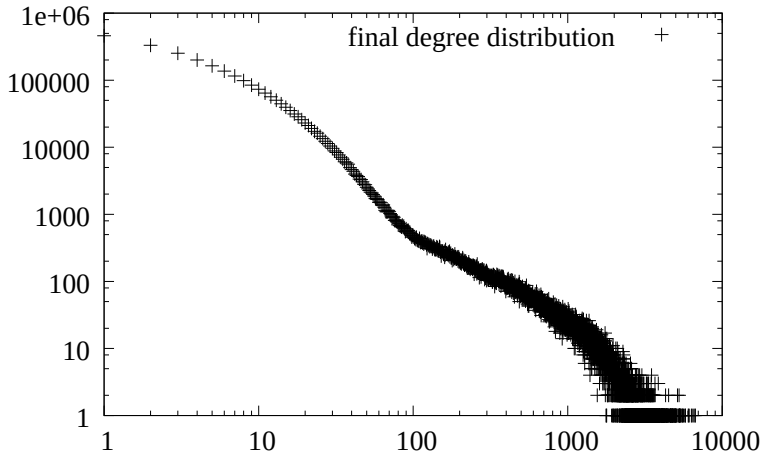


P2P

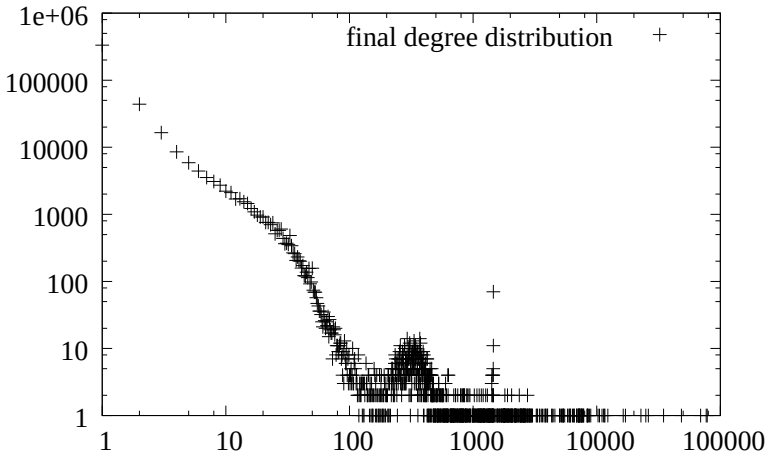
Degree distribution

p_k: number of nodes of nodes with degree k.
Usual assumption: heterogeneous, stable

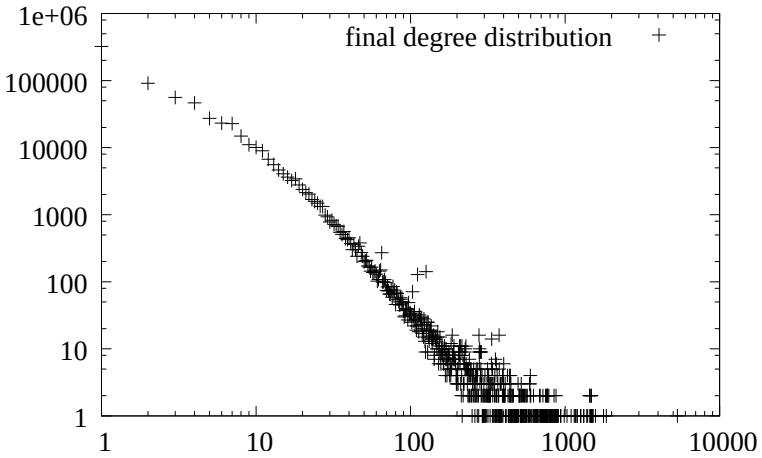
P2P



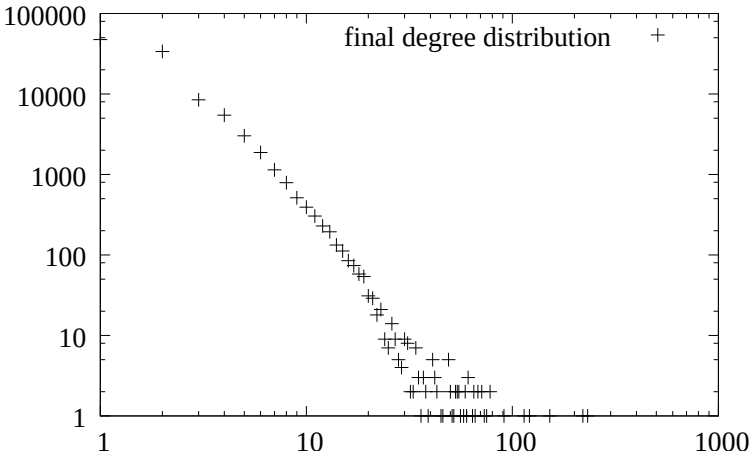
IP



Web



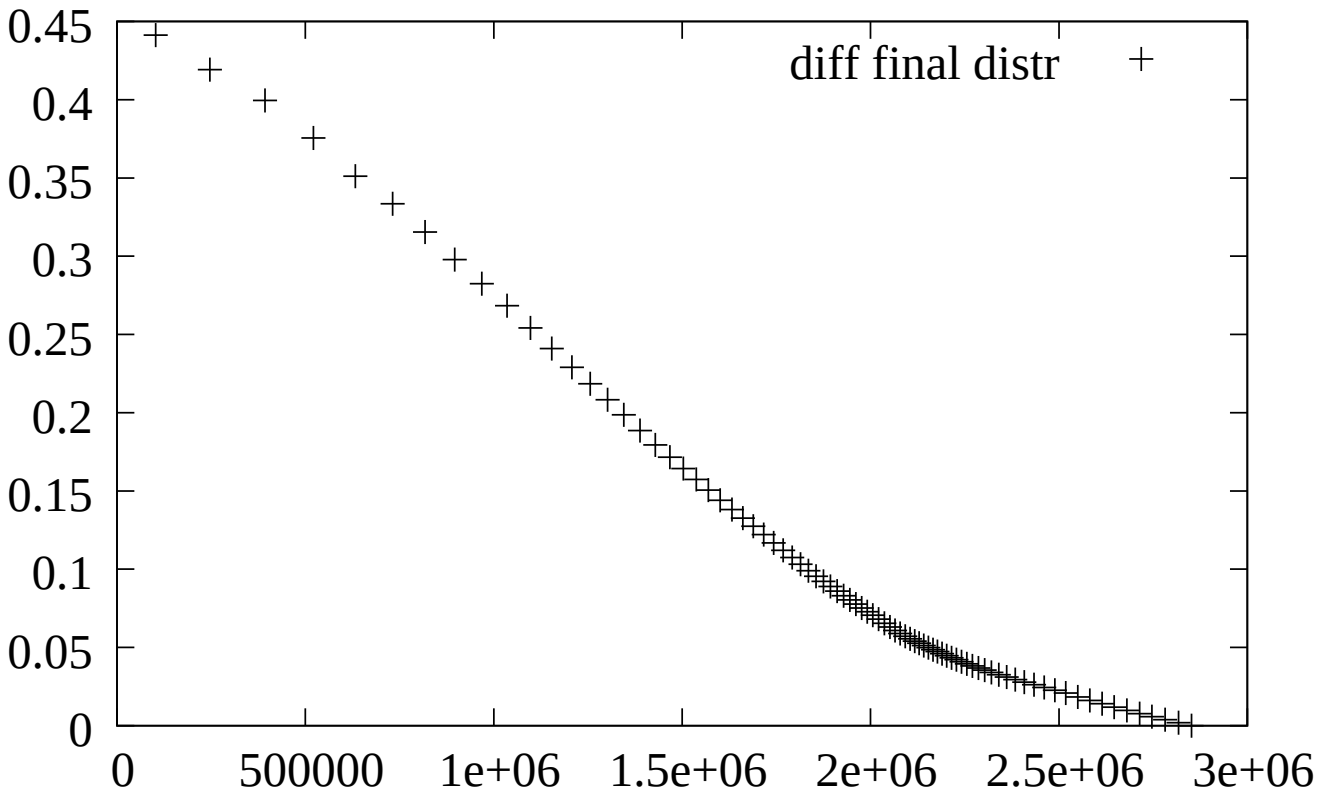
Inet



Always heterogeneous

Evolution of the degree distribution

K-S test: maximum difference between cumulative distributions
Difference with last degree distribution

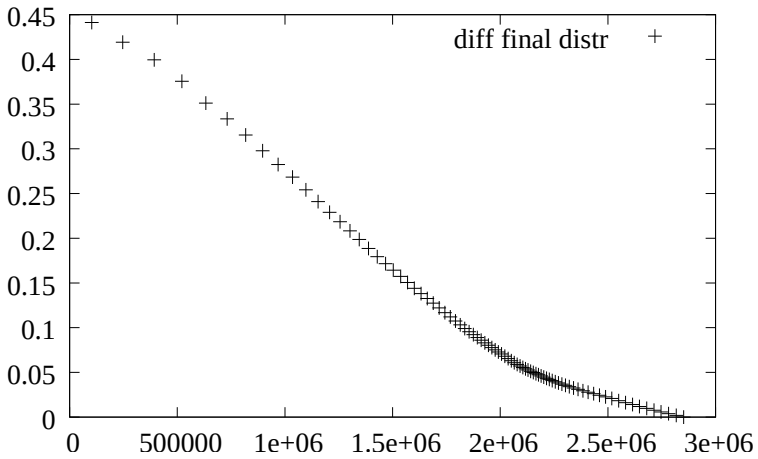


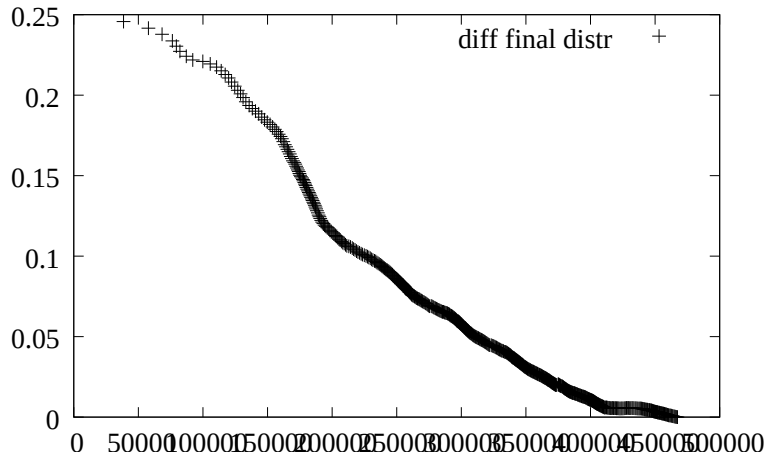
P2P

Evolution of the degree distribution

K-S test: maximum difference between cumulative distributions
Difference with last degree distribution

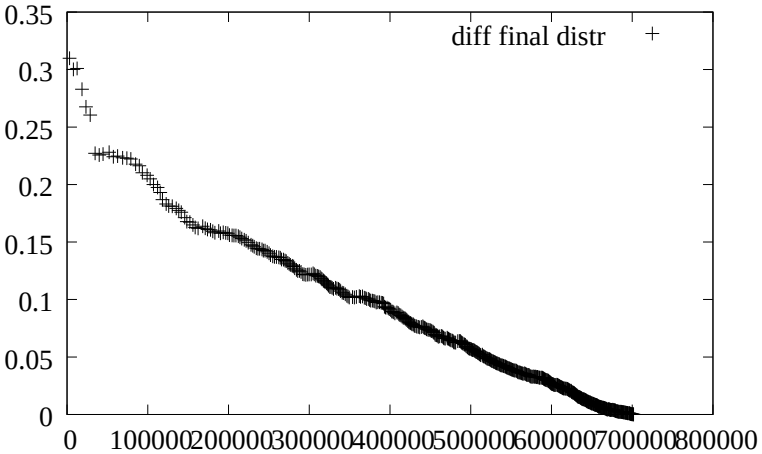
P2P

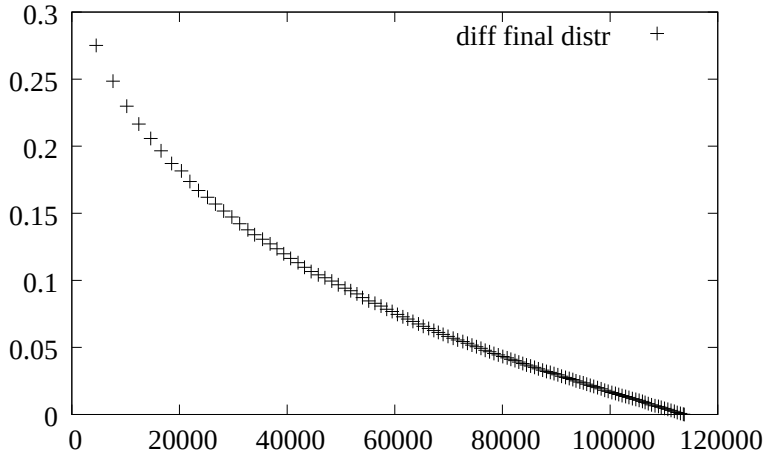




IP

Web

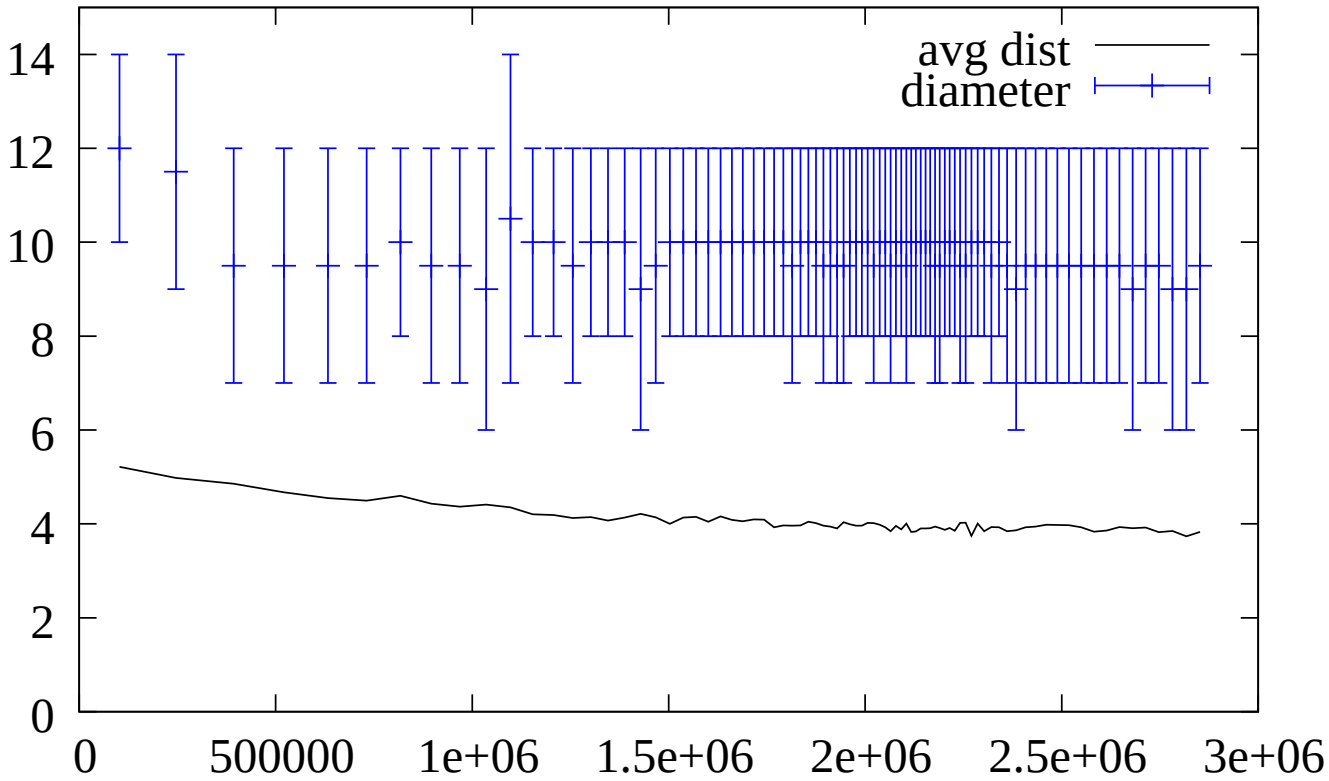




Inet

Average distance and diameter

Average distance: average on all pairs of nodes
Diameter: maximum distance
Use of heuristics
Usual assumptions: small, increasing



P2P

Average distance and diameter

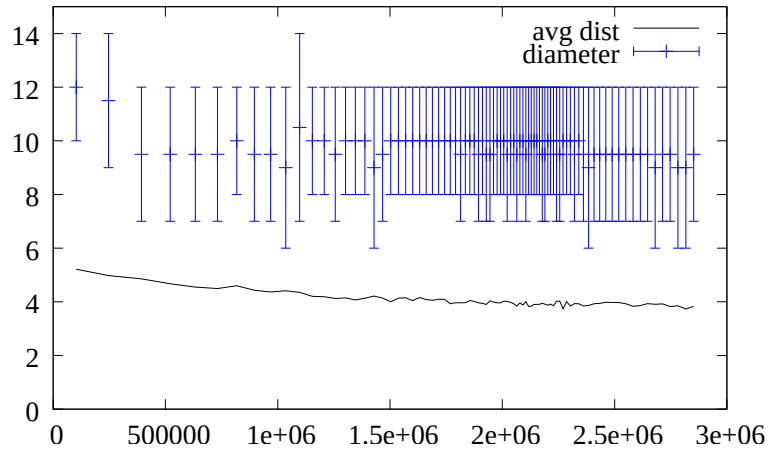
Average distance: average on all pairs of nodes

Diameter: maximum distance

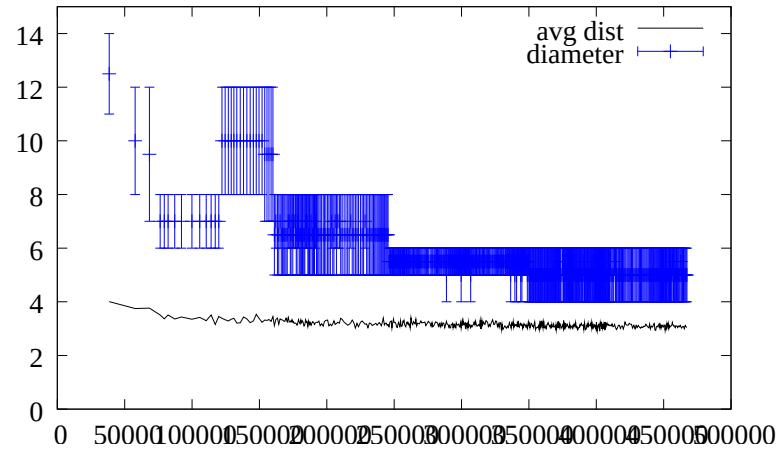
Use of heuristics

Usual assumptions: small, increasing

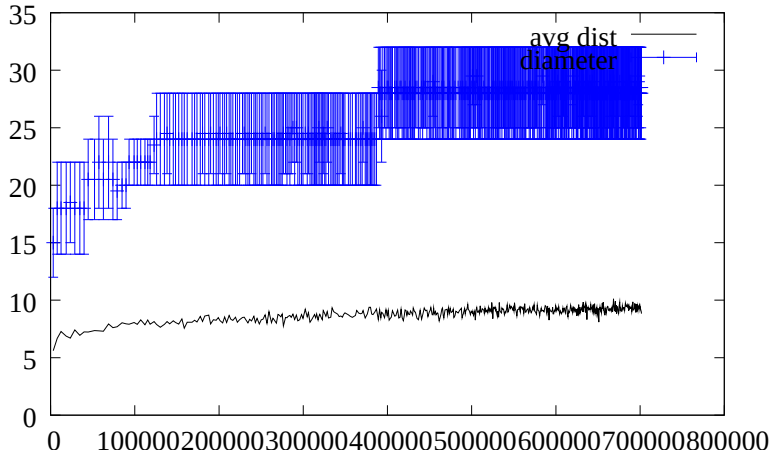
P2P



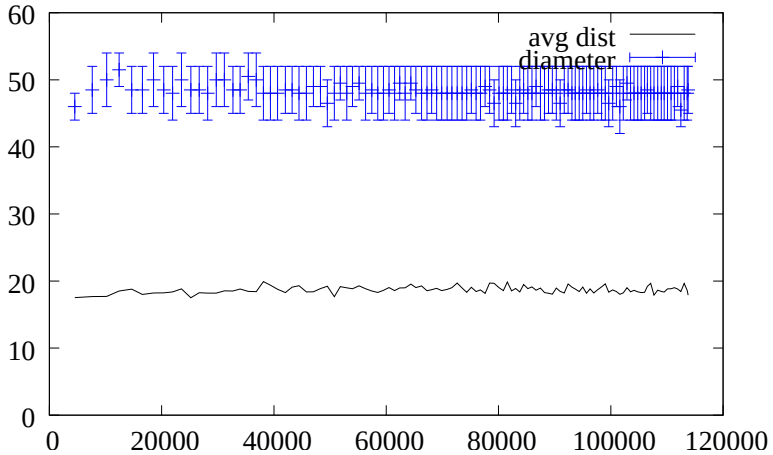
IP



Web

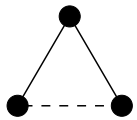


Inet

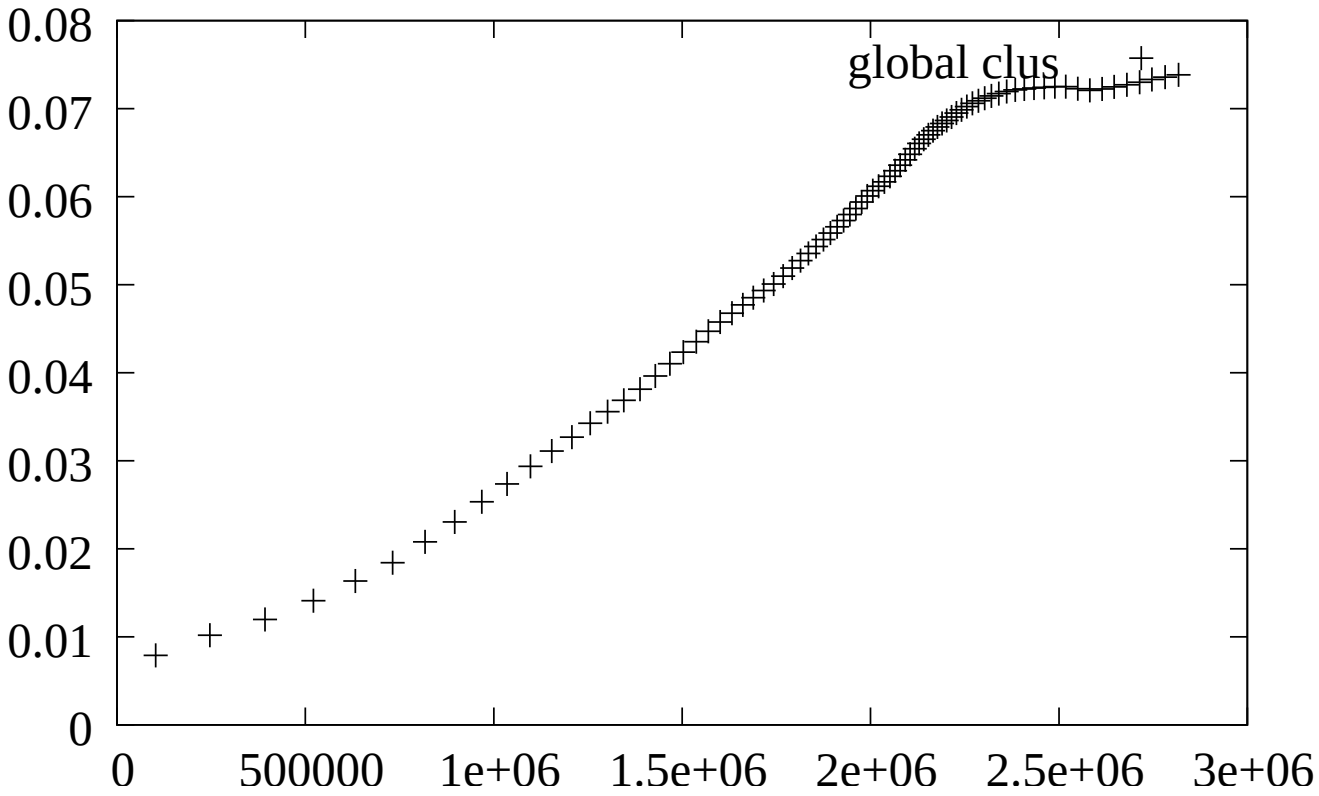


Clustering coefficient

Clustering coefficient:



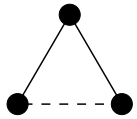
Usual assumption: significantly larger than density, constant



P2P

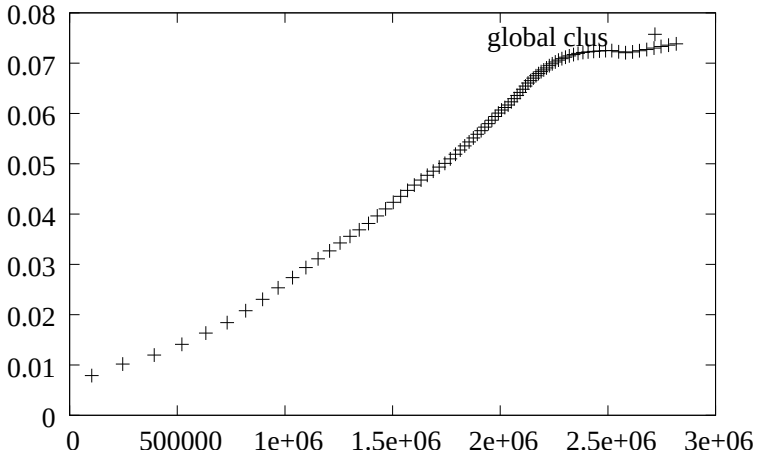
Clustering coefficient

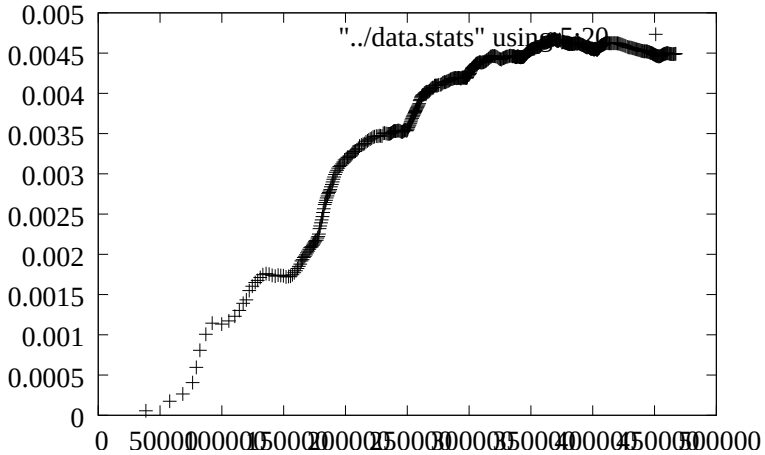
Clustering coefficient:



Usual assumption: significantly larger than density, constant

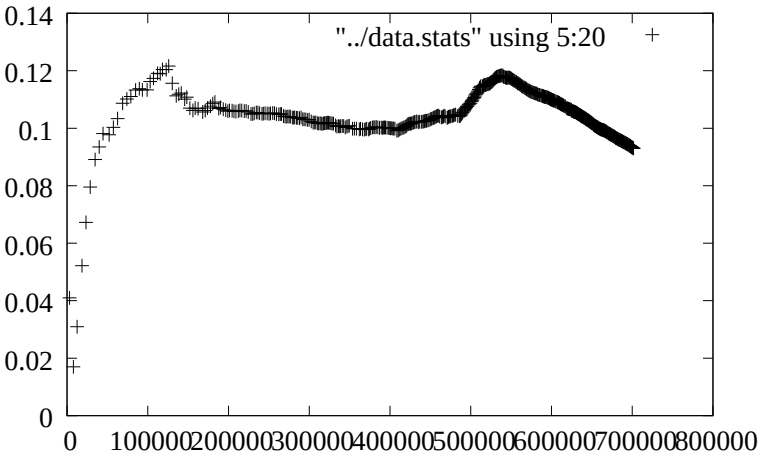
P2P

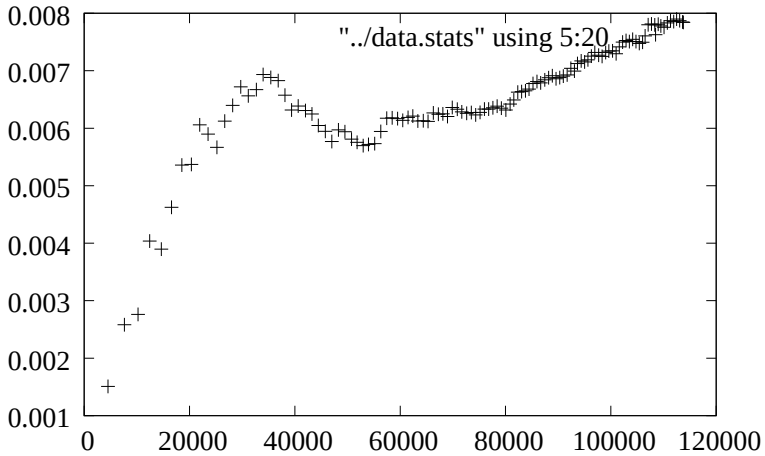




IP

Web





Inet

Conclusion and perspectives

Identification of universal properties

Some classical assumptions validated
Other properties infirmed

Perspectives:

- Study of different/larger data
- Links between these propertie
- New questions for modeling issues
- Investigations on slices of the samples