Networks Structure and Dynamics
11. Internet topology metrology

Maximilien Danisch, Marwan Ghanem, Lionel Tabourier

LIP6 – CNRS and Sorbonne Université

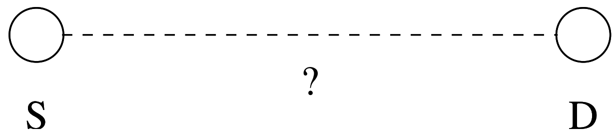
first_name.last_name@lip6.fr

December 18th 2018

Outline

Topology of the internet

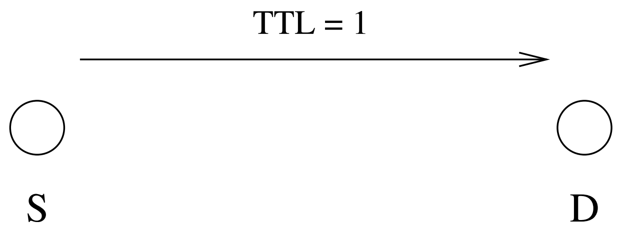
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

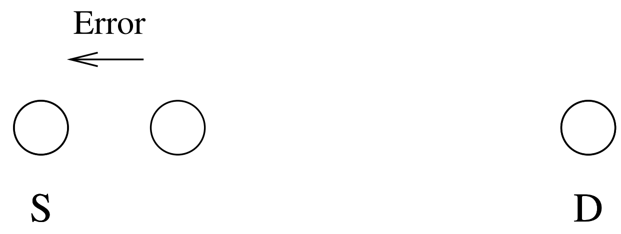
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

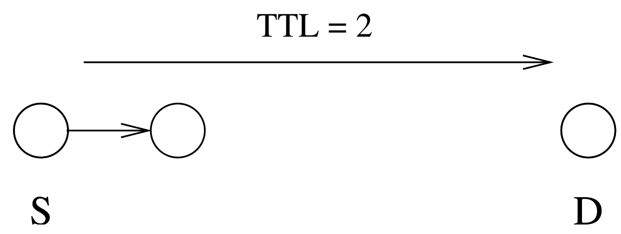
Topology of the internet

Measurement: exploration using `traceroute`

Principle: packets with same destination and increasing TTL

Topology of the internet

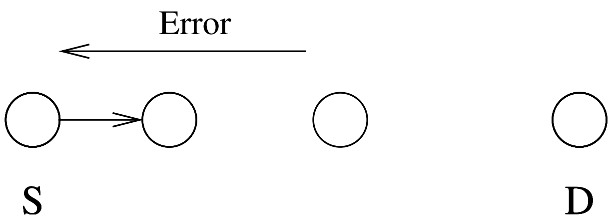
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

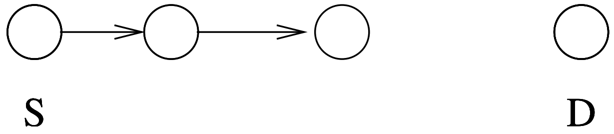
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

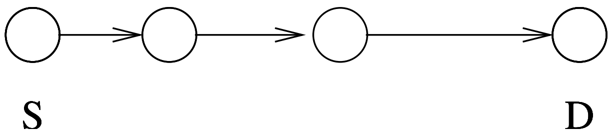
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

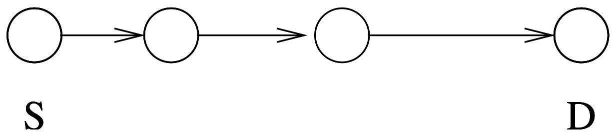
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

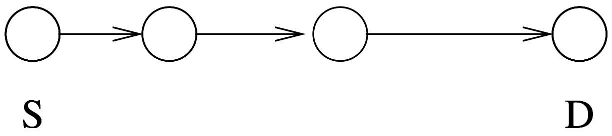
Measurement: exploration using `traceroute`



If no answer: *
ICMP filtered for various reasons:
Rate limiting
Time exceeded
...

Topology of the internet

Measurement: exploration using `traceroute`

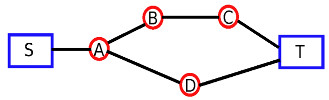


Remark:
one router = several IP addresses
answers with the IP address that sends the packet
$\Rightarrow$ simplified description of the process

Measurement bias

A very general but largely ignored fact about Internet-related measurements is that what we can measure in an Internet-like environment is typically not the same as what we really want to measure (or what we think we actually measure)

Mathematics and the internet: A source of enormous confusion and great potential, W. Willinger et al., Notices of the AMS, 2009.



Problematic

Information collection

A few sources, a lot of destinations:

- We know that we don't see everything
- How to get a meaningful view? ($\rightarrow$ evaluate bias)

Measured property

The degree distribution, we discussed this property a lot...
Degree distribution of the Internet: heterogeneous, even a power-law

Pansiot, Grad - *1998*

Faloutsos, Faloutsos, Faloutsos - *1999*

Surprising degree distribution observed $\rightarrow$ bias?

How to procede?

- Measure from a large number of sources
- Call to theoretical and experimental studies

Lecture goal: understand and comment research papers

Surprising degree distribution observed → bias?

How to procede?

- Measure from a large number of sources
- Call to theoretical and experimental studies

Lecture goal: understand and comment research papers

Outline

1. Introduction: traceroute measurement
2. Metrology
   - Influence of sources and destinations
   - Bias on degree

Volume of information

Barford, Bestavros, Byers, Crovella - *On the Marginal Utility of Network Topology Measurements, 2001*

General idea of the article

- Use data from measurements (rather than simulations)
- Evaluate number of nodes/links seen vs number of sources/destinations → unit of the information volume

Interest of using more sources and destinations

→ Does it increase the volume of information?
→ Does it decrease the bias?

Data

Two datasets

- 8 sources
- 1277 destinations
- 1 traceroute every 30 minutes
- approximately 7 months

- 12 sources
- $>$ 300 000 destinations
- same measurement method
- duration unknown

Data

Remark about the benefit of repeating measurements

Because of load-balancing, . . .
→ repeating give more information (and more noise too...)

Methodology

Assess the number of nodes seen as a function of

- the number of sources
- the number of destinations

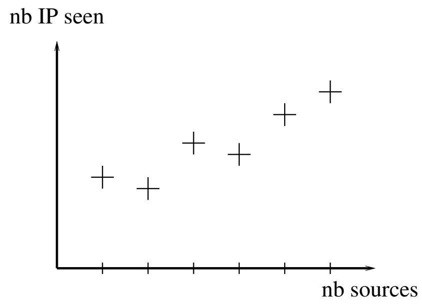
s sources, d destinations $\rightarrow s \times d$ possible parameter values

A lot of possibilities. . .

Interpretation?

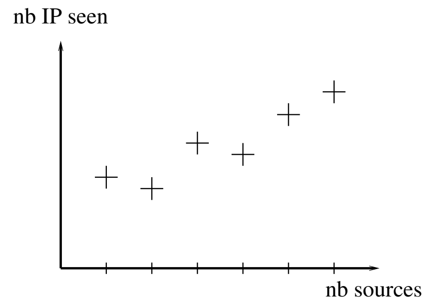
Methodology

What do we want?



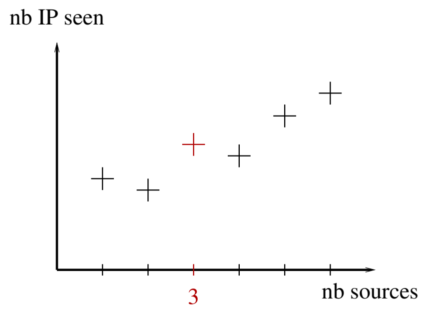
Methodology

What do we want?



same thing with destinations

Problem



Number of IPs seen with 3 sources: which 3 sources?

Example

One source $\rightarrow$ set of IPs seen

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

$s_1 + s_3 + s_6 \rightarrow$ 5 IP
$s_1 + s_4 + s_5 \rightarrow$ 10 IP

Depends on how complementary the sources are
no obvious choice

Example

One source $\rightarrow$ set of IPs seen

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

$s_1 + s_3 + s_6 \rightarrow$ 5 IP
$s_1 + s_4 + s_5 \rightarrow$ 10 IP

Depends on how complementary the sources are
no obvious choice

Greedy strategy

At each step: add the source which adds most information

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

Greedy strategy

At each step: add the source which adds most information

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

1 source: s_1

Introduction: traceroute measurement
Metrology
Influence of sources and destinations
Bias on degree

Greedy strategy

At each step: add the source which adds most information

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

2 sources: $s_1 s_5$

Introduction: traceroute measurement
Metrology
Influence of sources and destinations
Bias on degree

Greedy strategy

At each step: add the source which adds most information

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

3 sources: $s_1 s_5 s_4$

Introduction: traceroute measurement
Metrology
Influence of sources and destinations
Bias on degree

Greedy strategy

At each step: add the source which adds most information

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

4 sources: $s_1 s_5 s_4 s_2$

Greedy strategy

At each step: add the source which adds most information

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

5 sources: $s_1 s_5 s_4 s_2 s_3$

Greedy strategy

At each step: add the source which adds most information

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

6 sources: $s_1 s_5 s_4 s_2 s_3 s_6$

Greedy strategy

At each step: add the source which adds most information

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{a, d\}$

sources: $s_1 s_5 s_4 s_2 s_3 s_6$

Motivation: close to “best” case, without testing all combinations

Complexity

Complexity of the union of two sets

Complexity of step 2

compute $n - 1$ unions

Complexity of step i

compute $n - (i - 1)$ unions

Complexity

Complexity of the union of two sets

proportional to size of the smallest
(minimum, depends on the implementation)

Complexity of step 2

compute $n-1$ unions

Complexity of step i

compute $n-(i-1)$ unions

Complexity

Complexity of the union of two sets

proportional to size of the smallest
(minimum, depends on the implementation)

Complexity of step 2

compute $n-1$ unions
$\rightarrow (n-1)\times k$ if all sets are of size k

Complexity of step i

compute $n-(i-1)$ unions

Complexity

Complexity of the union of two sets

proportional to size of the smallest
(minimum, depends on the implementation)

Complexity of step 2

compute $n-1$ unions
$\rightarrow (n-1)\times k$ if all sets are of size k

Complexity of step i

compute $n-(i-1)$ unions
$\rightarrow (n-i+1)\times k$

Complexity

At step i

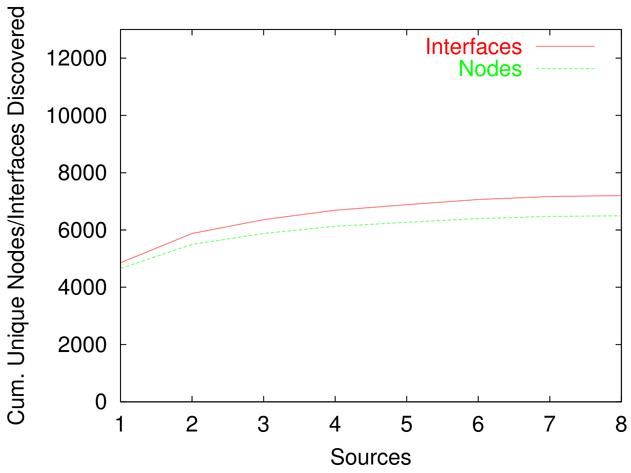
$n-(i-1)$ unions
$\rightarrow (n-i+1)\times k$

$$k((n-1)+(n-2)+\ldots+2+1)=\frac{kn(n-1)}{2}$$

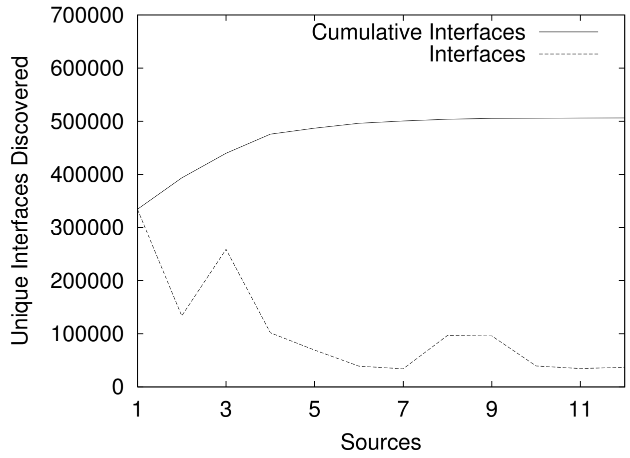
$$\mathcal{O}(kn^2)$$

long if large number of sources (n)

Results



Results



Observations

Convergence of the curve:
the last ones bring nearly no new information
→ authors conclude marginal utility of source addition

to be discussed later...

Destinations utility

In the ideal case, inverse approach:
Every destination → set of IPs seen

Greedy strategy is expensive → random strategy

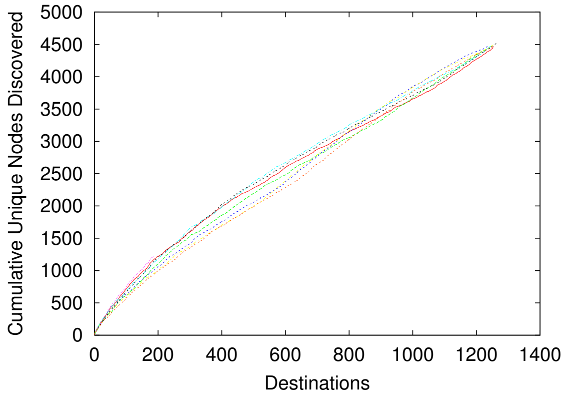
For one source

At each step:

- add randomly a destination

Compare curves for all sources

Results



Observation: roughly linear increase
similar benefit for all destinations

Comparison sources and destinations

Difference between curves
→ Why such difference between sources and destinations?

Intuition:
s sources, d destinations $\Longleftrightarrow$ d sources, s destinations

→ Importance of the strategy used
greedy vs random

Comparison sources and destinations

Difference between curves
→ Why such difference between sources and destinations?

Intuition:
s sources, d destinations $\Longleftrightarrow$ d sources, s destinations

→ Importance of the strategy used
greedy vs random

Comparison sources and destinations

Difference between curves
→ Why such difference between sources and destinations?

Intuition:
s sources, d destinations $\Longleftrightarrow$ d sources, s destinations

→ Importance of the strategy used
greedy vs random

Critical look

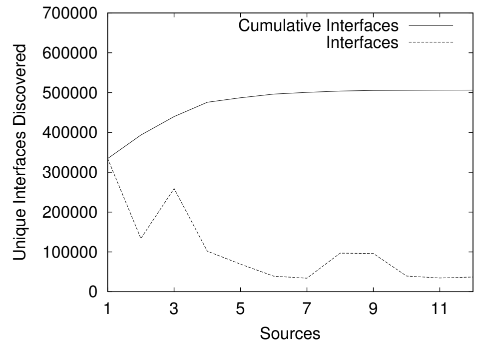
Interesting study, but...

Lack of details on

→ **disparity** between sources
(one source only sees 184 nodes , > 4000 for the largest one)
→ influence of the **strategy**

Q: is the choice of sources more important than their number?

Critical look



Last sources: bring few information
but the greedy strategy induce the shape of the curve
no obvious best strategy...

Datasets

To get a better understanding: compare different strategies

Ouédraogo, Magnien - *Computer Communications, 2011*

Data

- 11 sources
- 3 000 destinations
- 100 traceroutes per day
- ∼ 2 months

Difference between sources

Number of IPs seen per sources

Vary between:

- ∼ 16,500
- ∼ 26,500

→ Every sources are not equivalent

Influence of sources and destinations

Three different strategies

- greedy-max:
  add the source which brings the most information
- random:
  add a random source
- greedy-min:
  add the source which brings the least information

Influence of sources and destinations

Greedy strategy $\neq$ maximum possible with k sources

Example

$s_1 : \{a, b, c, d, e\}$ $s_3 : \{a, c, d, g\}$
$s_2 : \{a, b, e, f\}$

Influence of sources and destinations

Greedy strategy $\neq$ maximum possible with k sources

Example

$s_1 : \{a, b, c, d, e\}$ $s_3 : \{a, c, d, g\}$
$s_2 : \{a, b, e, f\}$

1 sources : s_1

Influence of sources and destinations

Greedy strategy $\neq$ maximum possible with k sources

Example

$s_1 : \{a, b, c, d, e\}$ $s_3 : \{a, c, d, g\}$
$s_2 : \{a, b, e, f\}$

2 sources : $s_1 s_2$

Influence of sources and destinations

Greedy strategy $\neq$ maximum possible with k sources

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, e, f\}$
s_3 : $\{a, c, d, g\}$

3 sources : $s_1 s_2 s_3$

Influence of sources and destinations

Greedy strategy $\neq$ maximum possible with k sources

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, e, f\}$
s_3 : $\{a, c, d, g\}$

3 sources : $s_1 s_2 s_3$

$s_2 + s_3$: 7 IP

Representativeness of maximum? (close to "standard" case?)
Cost to compute the maximum?

Influence of sources and destinations

Other strategies

- Max $\rightarrow$ max over 1000 random orders
- Min $\rightarrow$ min over 1000 random orders
- Random $\rightarrow$ average over 1000 random orders

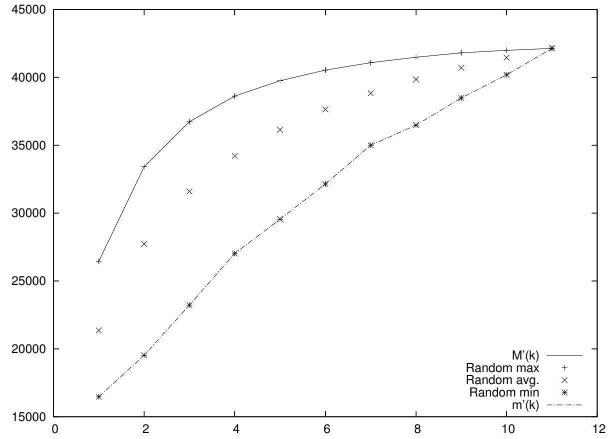
Influence of sources and destinations

Example

s_1 : $\{a, b, c, d, e\}$
s_2 : $\{a, b, c, d, f\}$
s_3 : $\{a, b\}$
s_4 : $\{g, h\}$
s_5 : $\{i, j, k\}$
s_6 : $\{i, j\}$

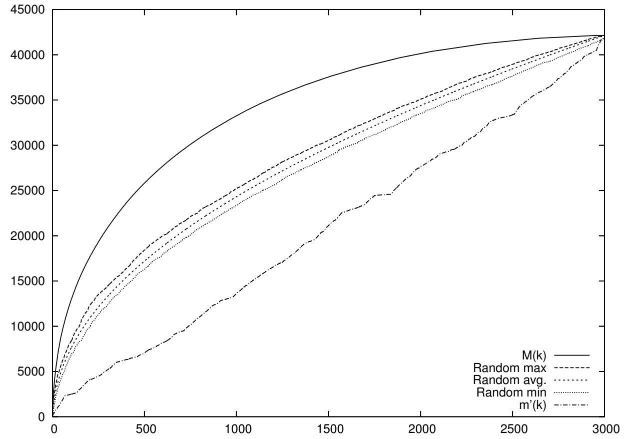
	s_3	s_4	s_6	s_5	s_2	s_1
	2	4	6	7	10	11
	s_5	s_6	s_2	s_4	s_3	s_1
	3	3	7	9	10	11
Min	2	3	6	7	10	11
Max	3	4	7	9	10	11
Average	2.5	3.5	6.5	8	10	11

Results



Influence of sources

Results



Influence of destinations

Observations

- Every curves ends at point n
- Random max (min) = Greedy max (min) for sources only
- Greedy max (averaged)
- In practice, larger variability with sources

Observations

- Every curves ends at point n
  because every node discovered
- Random max (min) = Greedy max (min) for sources only
  because few sources
- Greedy max (averaged)
  similar qualitative behaviors for sources and destinations
- In practice, larger variability with sources
  because few sources

Conclusion

Utility decrease, but not null

Choice of sources might be more important than number

Exploration bias

Lakhina, Byers, Crovella, Xie - *Sampling Biases in IP Topology Measurements, 2003*

Principle of the article: simulation-based

- Generate artificial graphs $\rightarrow$ topology
- Simulate traceroutes $\rightarrow$ measure
- Observe and analyze results

Explore the explicative dimension of modelling

Implementation - graph models

Basic graph models

- Erdős-Rényi
- Fixed degree distribution $\rightarrow$ configuration model

Implementation – traceroute simulation

How to simulate traceroute?
...several possibilities

Implementation – traceroute simulation

How to simulate traceroute?
...several possibilities

Usual choice

- route = shortest path (not true but default choice)

Shortest path

- One/every shortest paths?
- If one, which one?

The authors' choice

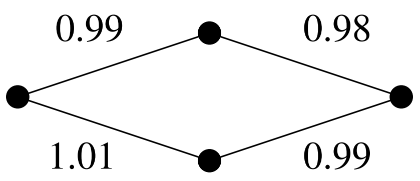
Give a weight to each link ($\rightarrow$ weighted graph)
$1 + \epsilon$, with a random $\epsilon \ll 1$

Length of a path: sum of the weights of the links
$\rightarrow$Every paths have different weights

The authors' choice

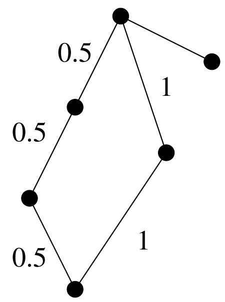
Give a weight to each link ($\rightarrow$ weighted graph)
$1 + \epsilon$, with a random $\epsilon \ll 1$

Length of a path: sum of the weights of the links
$\rightarrow$Every paths have different weights



Computation of the shortest weighted path

BFS not suited for weighted networks



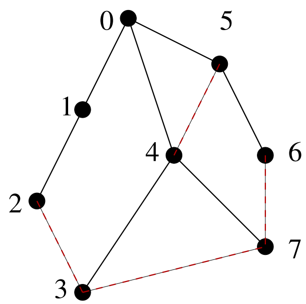
shortest paths from one node in weighted graph (weights>0)
$\rightarrow$ Dijkstra algorithm (not detailed here)

Our choice: restricted BFS

No weight
Distances computed with a BFS
Storage of the output of the BFS → table

Value i: father of i
Value root: root itself

0	0	1	4	0	0	5	4
0	1	2	3	4	5	6	7



Restiction to destinations

Table initialized at -1

For each destination `d` : (here : `d` = 3, 4, 6, 1)

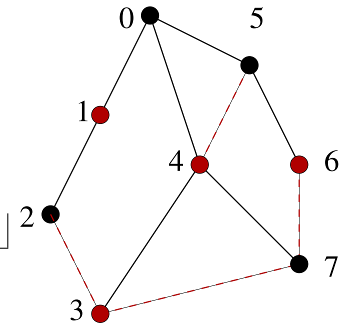
- While `AR[d] == -1`
  - AR[d] = A[d]
  - d = A[d]

A

0	0	1	4	0	0	5	4
0	1	2	3	4	5	6	7

AR

0	0	-1	4	0	0	5	-1
0	1	2	3	4	5	6	7



Degree computation

Degree of a node in the BFS tree:

Degree computation

Degree of a node in the BFS tree:
- number of times it appears +1
- except for the **root** : number of times -1

0	0	-1	4	0	0	5	-1
0	1	2	3	4	5	6	7

(boxes with -1: nodes which are not in the BFS tree)

Several sources

Several sources:
→ one BFS per source

How to compute the degree of the nodes?
mark links as present or absent

Connectedness

Problem if the graph is not connected...

Several solutions

- Choose sources and destinations in the same connected component
- Use only connected graphs
- ...

No ideal solution

Connectedness

Problem if the graph is not connected...

Authors' choice:

Restrict to the largest connected component

Simulations

Two cases under study:

Erdős-Rényi graphs (homogeneous degree)

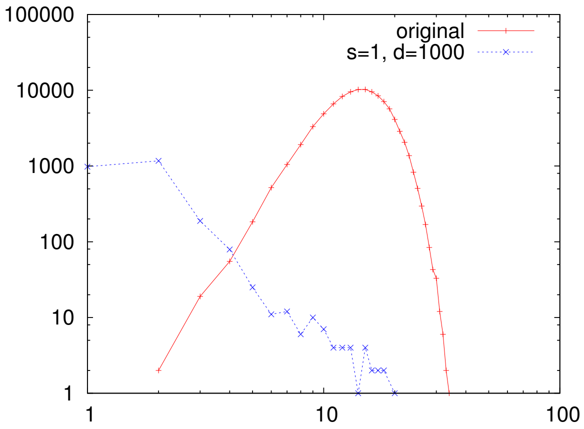
- $n = 100\,000$
- $m = 750\,000$ ($d^\circ(G) = 15$)
- sources: 1, 5, 10
- destinations: 1000, chosen randomly

Fixed degree distribution (heterogeneous)

- $n \sim 100\,000$
- $m \sim 190\,000$
- power-law, $\alpha \sim 2.1$

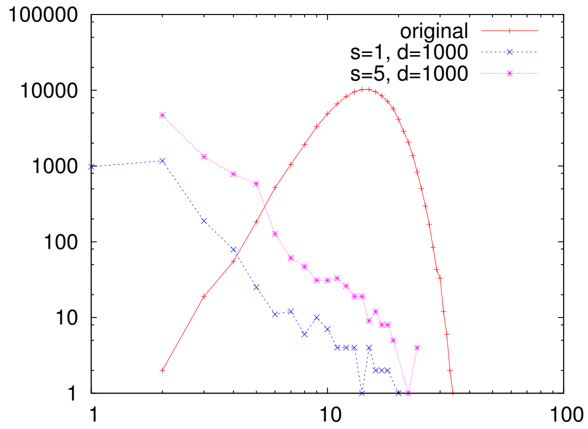
Results

Erdős-Rényi graphs



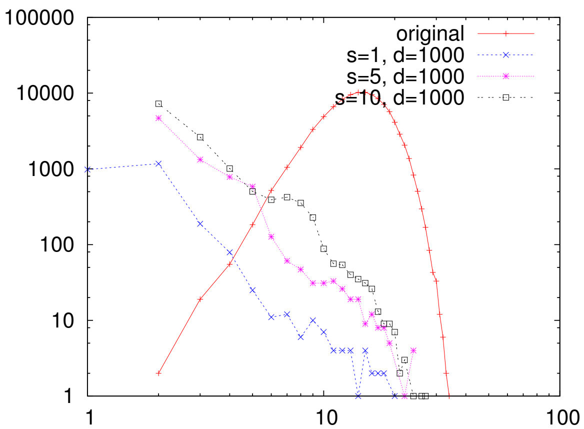
Results

Erdős-Rényi graphs



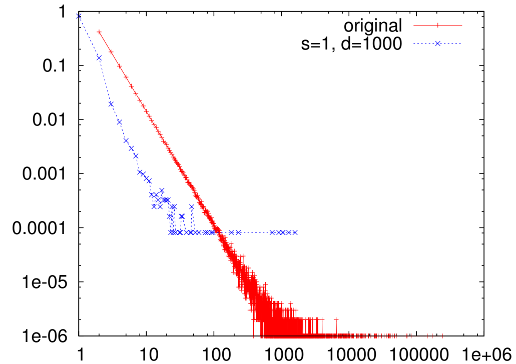
Results

Erdős-Rényi graphs



Results

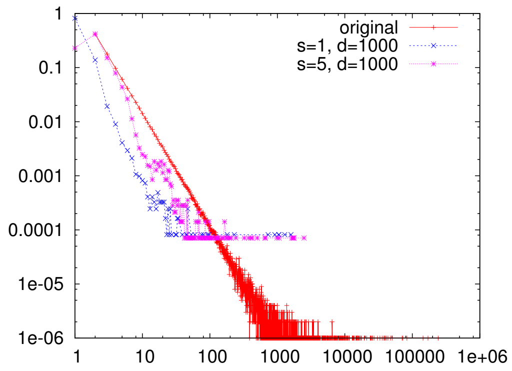
Graphs with fixed heterogeneous degree



Remark: notice the 1/N floor

Results

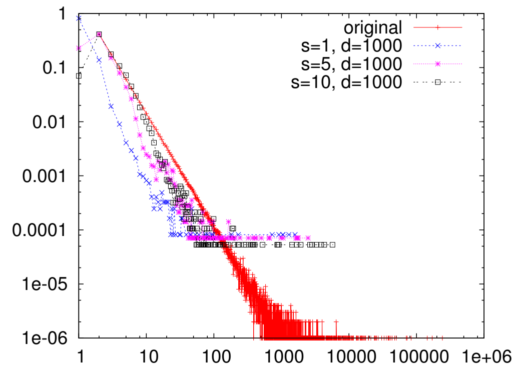
Graphs with fixed heterogeneous degree



Remark: notice the 1/N floor

Results

Graphs with fixed heterogeneous degree



Remark: notice the 1/N floor

Observations

- Distribution observed $\neq$ real distribution
- Erdős-Rényi: qualitative difference
  homogeneous appears as heterogeneous
- Graphs with fixed degree: quantitative difference
  slope, max degree, ...

Warning:
ER graphs: Maximum degree observed $\sim$ 30
$\rightarrow$impossible to conclude on heterogeneity

Observations

- Distribution observed $\neq$ real distribution
- Erdős-Rényi: qualitative difference
  homogeneous appears as heterogeneous
- Graphs with fixed degree: quantitative difference
  slope, max degree, ...

Warning:
ER graphs: Maximum degree observed $\sim$ 30
$\rightarrow$impossible to conclude on heterogeneity

Conclusion of the study

Observing heterogeneous distrib $\not\Rightarrow$ Real heterogeneous distrib

No conclusion on the real distribution

Discussion (1/2)

Important result

- From a theoretical point of view
- Need to be careful about conclusions in practice

What conclusions can we draw from this?

Observed distribution heterogeneous

$\rightarrow$ Real distribution homogeneous?
$\rightarrow$ Real distribution heterogeneous?

Discussion (2/2)

Case of ER graphs

Maximal degree observed:

Discussion (2/2)

Case of ER graphs

Maximal degree observed:
close to average degree of the graph.

Practically, maximum degree observed > 1000
$\rightarrow$ random graph with average degree = 1000?

$\rightarrow$real distribution probably heterogeneous...
Need more studies

Sources of the bias

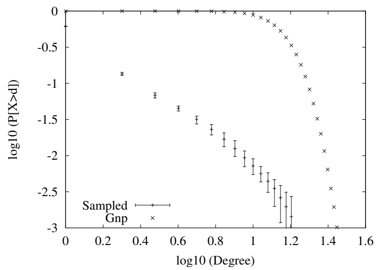
Hyp: Bias in the node sample?

For each node: compare the degree observed to its real degree

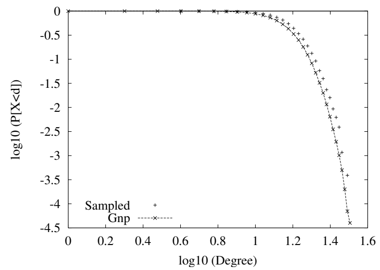
Sources of the bias

Hyp: Bias in the node sample?

observed deg vs original deg



real deg vs original deg

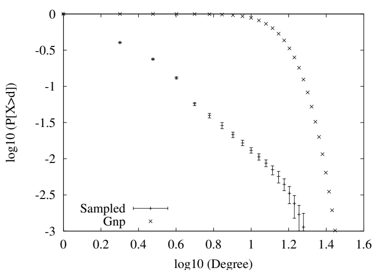


With 1 source

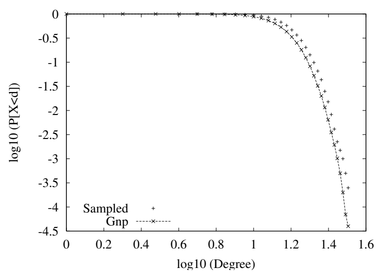
Sources of the bias

Hyp: Bias in the node sample?

observed deg vs original deg



real deg vs original deg

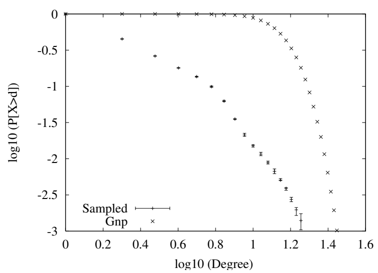


With 5 sources

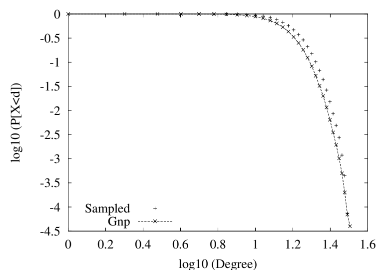
Sources of the bias

Hyp: Bias in the node sample?

observed deg vs original deg



real deg vs original deg

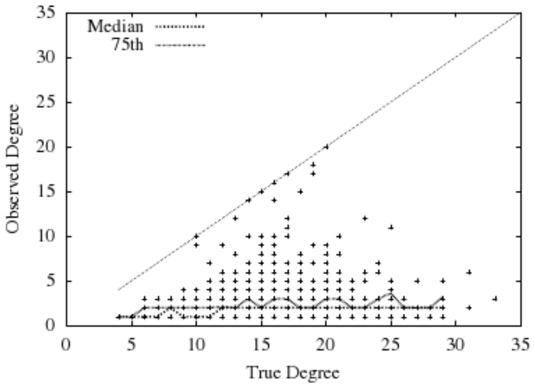


With 10 sources
Nodes are chosen without bias on the degree

Bias sources

Hyp: Bias in the link sample?

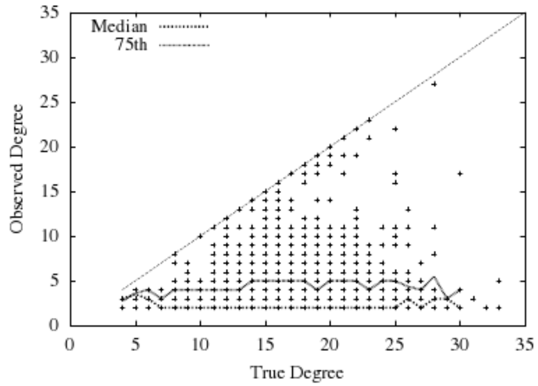
degree observed vs original degree
With 1 source



Bias sources

Hyp: Bias in the link sample?

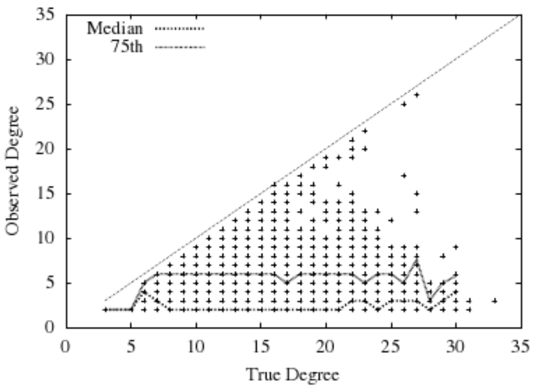
degree observed vs original degree
With 5 sources



Bias sources

Hyp: Bias in the link sample?

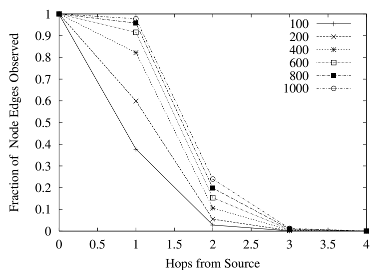
degree observed vs original degree
With 10 sources



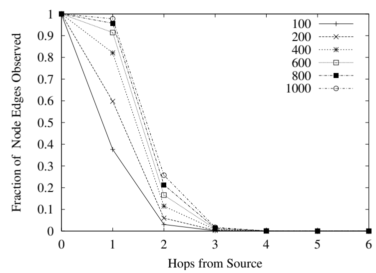
Bias sources

Link visibility as a function of their distance to the source

10,000 nodes



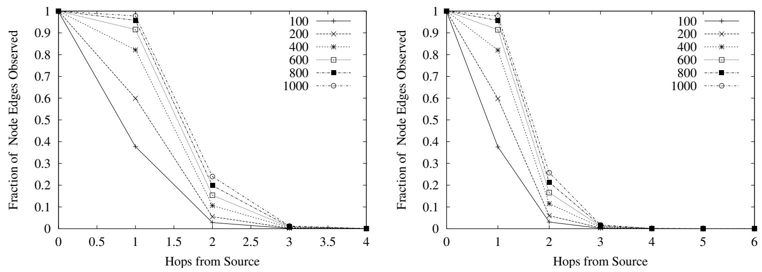
1,000,000 nodes



Bias sources

Link visibility as a function of their distance to the source

10,000 nodes 1,000,000 nodes



The further an edge is from the source,
the less are its chances to be seen

Given sample → bias?

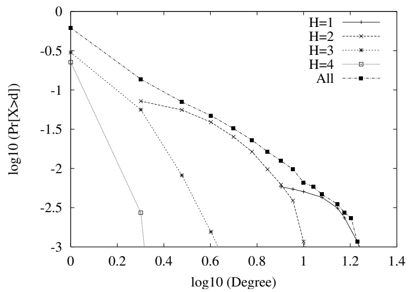
Given a sample (but not the original graph),
can we know if there is some bias?

Introduction: traceroute measurement
Metrology
Influence of sources and destinations
Bias on degree

Given sample → bias?

Given a sample (but not the original graph),
can we know if there is some bias?

Measure the probability to observe both degree d *and* distance h



The most distant are the nodes, the weaker is the degree