MU5IN075
Network Analysis and Mining
14. Structure of the Internet

Esteban Bautista-Ruiz, Lionel Tabourier

LIP6 – CNRS and Sorbonne Université

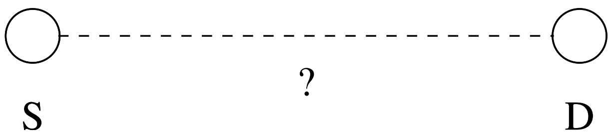
first_name.last_name@lip6.fr

January 11, 2022

Outline

Topology of the internet

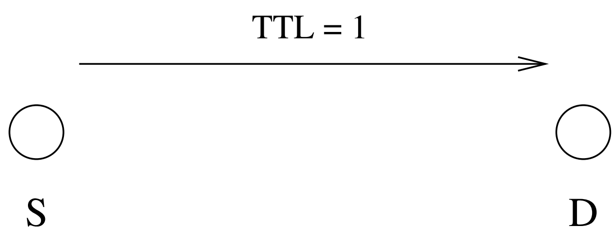
Measurement: exploration using `traceroute`



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

Introduction: traceroute measurement
Metrology
Correcting the bias

Topology of the internet

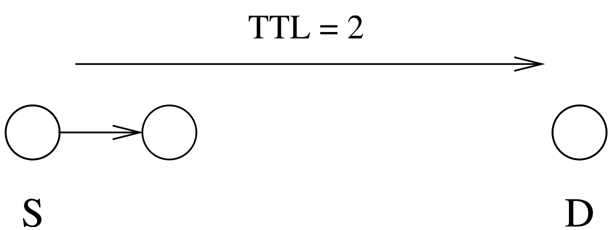
Measurement: exploration using `traceroute`

Principle: packets with same destination and increasing TTL

Introduction: traceroute measurement
Metrology
Correcting the bias

Topology of the internet

Measurement: exploration using `traceroute`

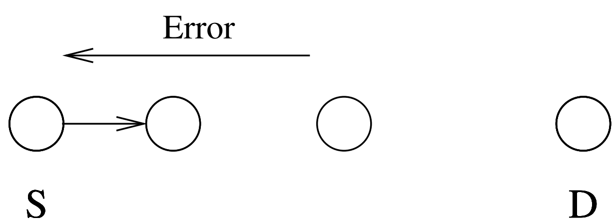


Principle: packets with same destination and increasing TTL

Introduction: traceroute measurement
Metrology
Correcting the bias

Topology of the internet

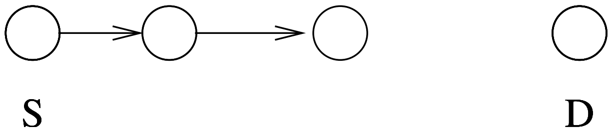
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

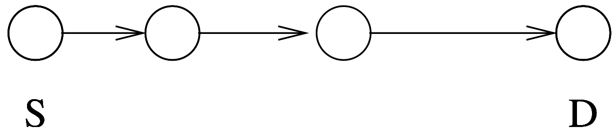
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

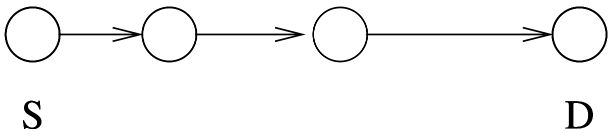
Measurement: exploration using `traceroute`



Principle: packets with same destination and increasing TTL

Topology of the internet

Measurement: exploration using `traceroute`

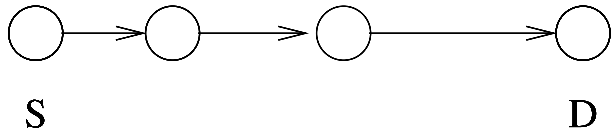


If no answer: *

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*

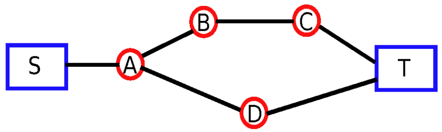
example: traceroute measurement from source S to target T

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*

example: traceroute measurement from source S to target T



Problematic

Information collection

Practically, 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

Degree distribution of the Internet:
we know it is heterogeneous, close to a power-law

On routes and multicast trees in the Internet - *Pansiot and Grad, 1998*
On power-law relationships of the internet topology - *Faloutsos, Faloutsos and Faloutsos, 1999*

Problematic

Information collection

Practically, 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

Degree distribution of the Internet:
we know it is heterogeneous, close to a power-law

On routes and multicast trees in the Internet - *Pansiot and Grad, 1998*
On power-law relationships of the internet topology - *Faloutsos, Faloutsos and Faloutsos, 1999*

Surprising degree distribution observed → bias?

How to procede?

- Experimental: measure from a large number of sources
- Also calls for theoretical studies

Lecture goal
understand and analyze research results
topic: IP maps of the Internet

Surprising degree distribution observed → bias?

How to procede?

- Experimental: measure from a large number of sources
- Also calls for theoretical studies

Lecture goal
understand and analyze research results
topic: IP maps of the Internet

Outline

Volume of information

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

When using more sources and destinations...

→ ...does it increase the volume of information?
→ ...does it decrease the bias?

Volume of information

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

When 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

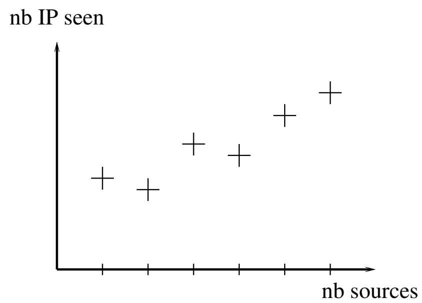
s sources, d destinations → $s \times d$ combination of values

A lot of possibilities...

→ how to choose?

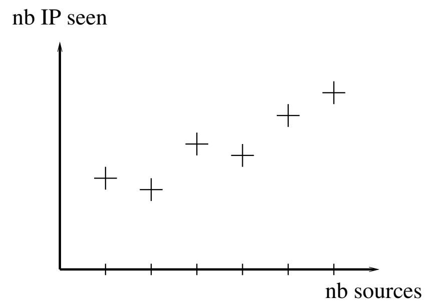
Methodology

What do we want?



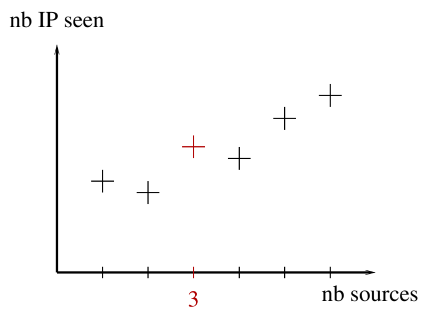
Methodology

What do we want?



same thing with destinations

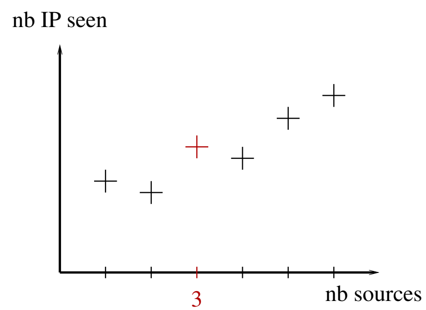
Problem



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

Looking for a strategy which maximizes the IPs seen

Problem



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

Looking for a strategy which maximizes the IPs seen

Example

One source → set of IPs seen

Example

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

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

Depends on how complementary the sources are
→ no obvious choice

Example

One source → set of IPs seen

Example

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

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

Depends on how complementary the sources are
→ no obvious choice

Example

One source → set of IPs seen

Example

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

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

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_4 : $\{g, h\}$
s_2 : $\{a, b, c, d, f\}$ | s_5 : $\{i, j, k\}$
s_3 : $\{a, b\}$ | 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

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: $s_1 s_5$

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: $s_1 s_5 s_4$

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 of this processing

Complexity of the union of two sets

Complexity of step 2

compute $n-1$ unions
(supposing that all sets approximately of the same size for simplicity)

Complexity of step i

compute $n-(i-1)$ unions

Complexity of this processing

Complexity of the union of two sets

$\rightarrow$ proportional to size of the smallest
(at least, it depends on the implementation)

Complexity of step 2

compute $n - 1$ unions
(supposing that all sets approximately of the same size for simplicity)

Complexity of step i

compute $n - (i - 1)$ unions

Complexity of this processing

Complexity of the union of two sets

$\rightarrow$ proportional to size of the smallest
(at least, it depends on the implementation)

Complexity of step 2

compute $n - 1$ unions
(supposing that all sets approximately of the same size for simplicity)
$\rightarrow (n - 1) \times k$ if all sets are of size k

Complexity of step i

compute $n - (i - 1)$ unions

Complexity of this processing

Complexity of the union of two sets

$\rightarrow$ proportional to size of the smallest
(at least, it depends on the implementation)

Complexity of step 2

compute $n - 1$ unions
(supposing that all sets approximately of the same size for simplicity)
$\rightarrow (n - 1) \times k$ if all sets are of size k

Complexity of step i

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

Complexity of this processing

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

so, from step $i = 2$ to step $i = n$
$\rightarrow k((n-1) + (n-2) + \ldots + 2 + 1) = \frac{kn(n-1)}{2}$

total: $\mathcal{O}(kn^2)$
long if large number of sources (n)

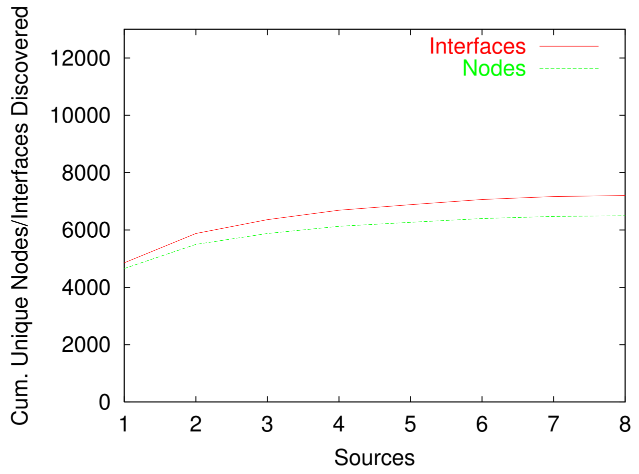
Complexity of this processing

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

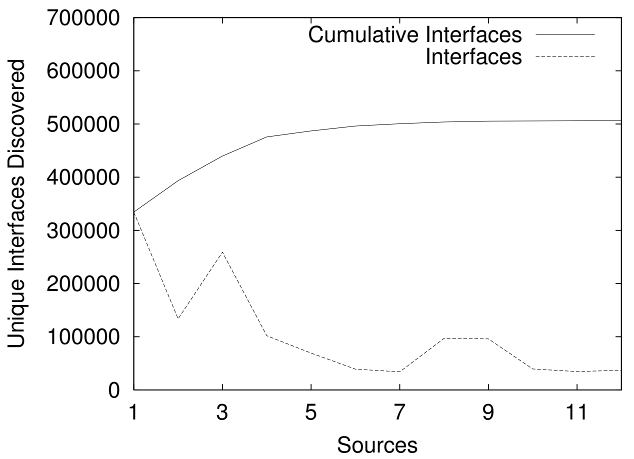
so, from step $i=2$ to step $i=n$
$\rightarrow k((n-1)+(n-2)+\ldots+2+1)=\frac{kn(n-1)}{2}$

total: $\mathcal{O}(kn^2)$
long if large number of sources (n)

Results



Results



Observations

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

to be discussed later...

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, similar approach:
for every destination → set of IPs seen

but greedy strategy is expensive → random strategy

Random strategy

For one source, at each step:

- add randomly a destination

Compare curves for all sources

Destinations utility

In the ideal case, similar approach:
for every destination → set of IPs seen

but greedy strategy is expensive → random strategy

Random strategy

For one source, at each step:

- add randomly a destination

Compare curves for all sources

Destinations utility

In the ideal case, similar approach:
for every destination → set of IPs seen

but greedy strategy is expensive → random strategy

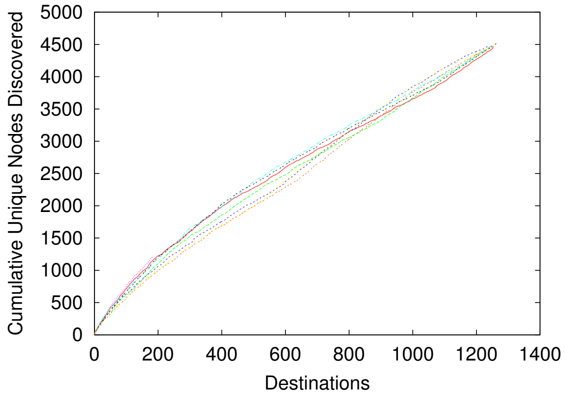
Random strategy

For one source, at each step:

- add randomly a destination

Compare curves for all sources

Results



Observation: roughly linear increase
same benefit for all destinations

Comparison sources and destinations

→ relevant to increase number of destinations rather than number of sources

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

We observe difference between curves
→ why such difference between sources and destinations?

while the intuition is
s sources, *d* destinations $\Longleftrightarrow$ *d* sources, *s* destinations

→ importance of the strategy used
greedy vs random

Comparison sources and destinations

→ relevant to increase number of destinations rather than number of sources

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

We observe difference between curves
→ why such difference between sources and destinations?

while the intuition is
s sources, *d* destinations $\Longleftrightarrow$ *d* sources, *s* destinations

→ importance of the strategy used
greedy vs random

Comparison sources and destinations

→ relevant to increase number of destinations rather than number of sources

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

We observe difference between curves
→ why such difference between sources and destinations?

while the intuition is
s sources, *d* destinations $\Longleftrightarrow$ *d* sources, *s* destinations

→ importance of the strategy used
greedy vs random

Comparison sources and destinations

→ relevant to increase number of destinations rather than number of sources

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

We observe difference between curves
→ why such difference between sources and destinations?

while the intuition is
s sources, d destinations $\iff$ d sources, s destinations

→ importance of the strategy used
greedy vs random

Critical look at the study

In spite of its interest ...

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

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

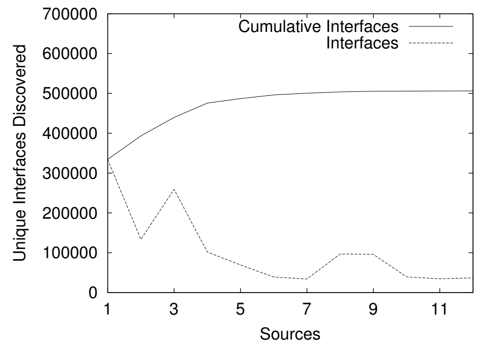
Critical look at the study

In spite of its interest ...

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

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

Critical look at the study



Last sources: bring few information
but the greedy strategy induces 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
- $\sim$ 2 months

Difference between sources

Number of IPs seen per sources

Vary between:

- $\sim$ 16,500
- $\sim$ 26,500

$\rightarrow$ Every sources are not equivalent
(although more homogeneous than in *Barford et al.*)

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

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

Influence of sources and destinations

Greedy strategy $\neq$ maximum possible with k sources

Example

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

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_2 : $\{a, b, e, f\}$
s_3 : $\{a, c, d, g\}$

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 IPs
$\Rightarrow s_2 + s_3$ gives more info than $s_1 + s_2$

Influence of sources and destinations

Greedy strategy $\neq$ maximum possible with k sources

Example

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

3 sources : $s_1 s_2 s_3$

$s_2 + s_3$: 7 IPs
$\Rightarrow s_2 + s_3$ gives more info than $s_1 + s_2$

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

Influence of sources and destinations

Other strategies

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

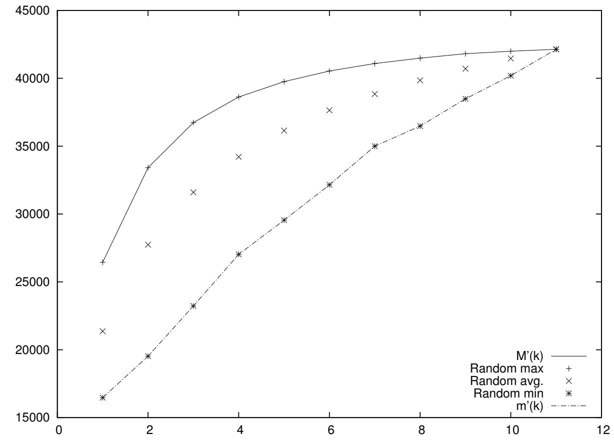
Influence of sources and destinations

Example

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

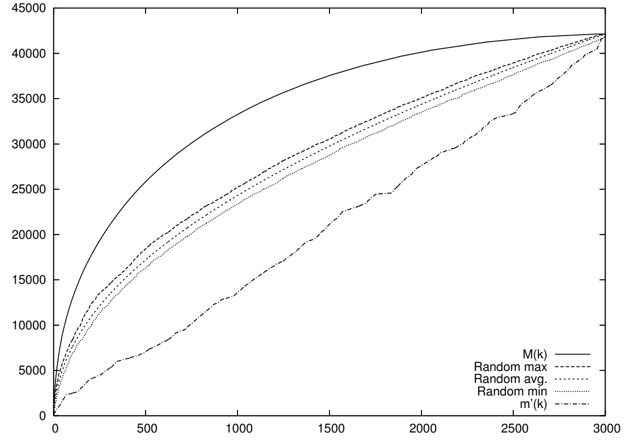
random 1	s_3	s_4	s_6	s_5	s_2	s_1
	2	4	6	7	10	11
random 2	s_5	s_6	s_2	s_4	s_3	s_1
	3	3	7	9	10	11
random min	2	3	6	7	10	11
random max	3	4	7	9	10	11
random average	2.5	3.5	6.5	8	10	11

Results (number of IPs seen)



Influence of sources
rk: M′ is greedy-max, m’ is greedy-min

Results (number of IPs seen)



Influence of destinations

rk: M′ is greedy-max, m’ is greedy-min

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

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

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

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 decreases, but not null
→ choice of sources might be more important than number

Ouédraogo, Magnien - *Computer Communications, 2011*

Exploration bias

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

Principle of the article: simulation-based

- Generate artificial graphs → topology
- Simulate traceroutes → measure
- Observe and analyze results

Explore the explicative dimension of modelling

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

Exploration bias

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

Principle of the article: simulation-based

- Generate artificial graphs → topology
- Simulate traceroutes → measure
- Observe and analyze results

Explore the explicative dimension of modelling

30/54

Implementation - graph models

Basic graph models

- Erdős-Rényi
- Fixed degree distribution → configuration model

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?

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?

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

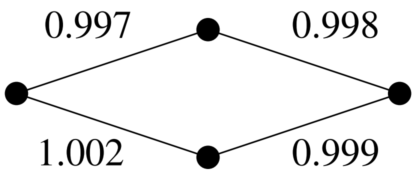
Give a weight to each link ($\rightarrow$ weighted graph)
$1 + \epsilon$, with a random $\epsilon \ll 1$

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

The authors' choice

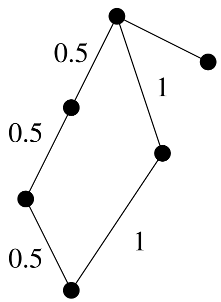
Give a weight to each link ($\rightarrow$ weighted graph)
$1 + \epsilon$, with a random $\epsilon \ll 1$

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



Computation of the shortest weighted path

Remark: BFS not suited for weighted networks



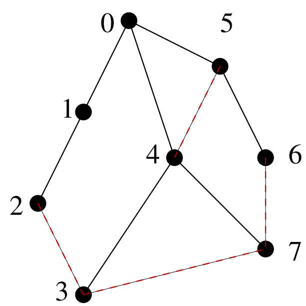
shortest paths from one node in weighted graph (weights>0)
$\rightarrow$ Dijkstra algorithm (not detailed here)

Our choice: restricted BFS

- no weight
- distances computed with a BFS
- storage of the output of the BFS $\rightarrow$ table of fathers

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)

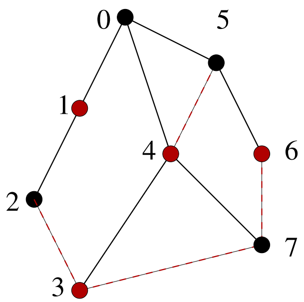
- While `AR[d] == -1`
  - AR[d] = A[d]
  - d = A[d]

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 using table of fathers:

Degree computation

Degree of a node in the BFS tree using table of fathers:
- 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:
$\rightarrow$ 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

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

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

39/54

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

Connectedness

Problem if the graph is not connected...

Authors' choice:

Restrict to the largest connected component

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

Simulations

Two cases under study:

Erdős-Rényi graphs (homogeneous degree)

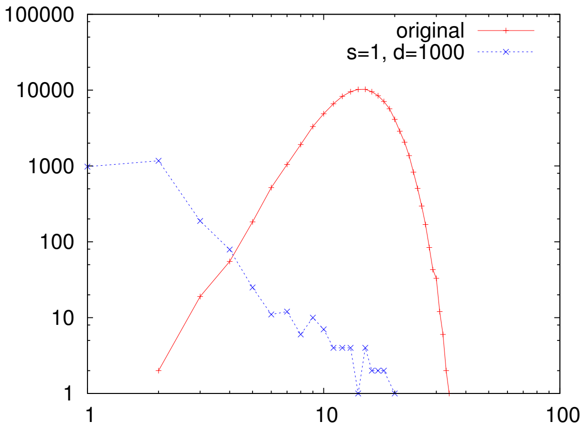
- $n = 100\,000$
- $m = 750\,000$ ($d^{\circ}(G) = 15$)
- sources: 1, 5, 10
- destinations: 1000, chosen randomly

Fixed degree distribution (heterogeneous degree)

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

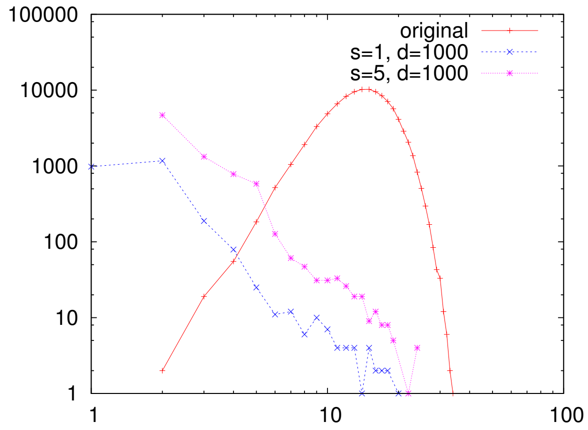
Results

Erdős-Rényi graphs



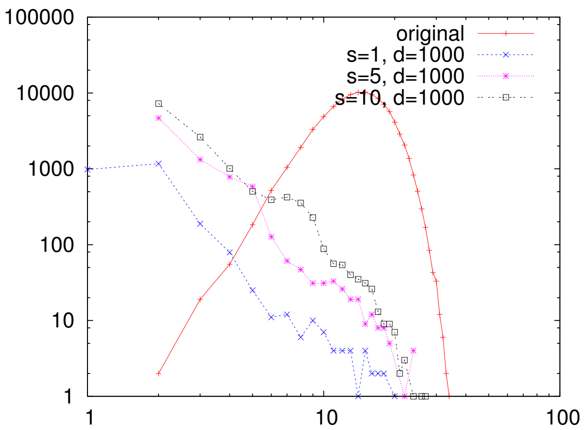
Results

Erdős-Rényi graphs



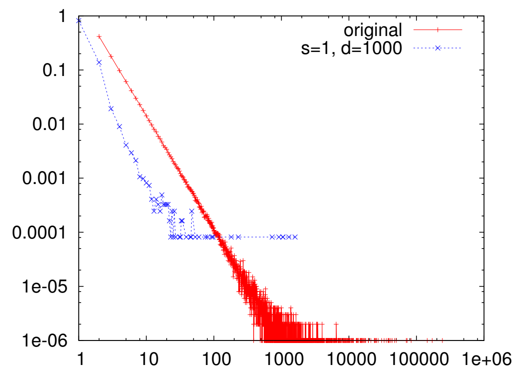
Results

Erdős-Rényi graphs



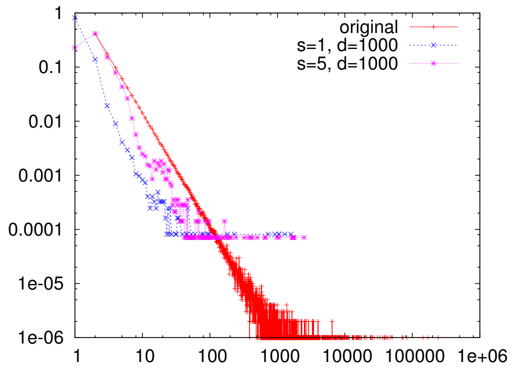
Results

Graphs with fixed heterogeneous degree



Results

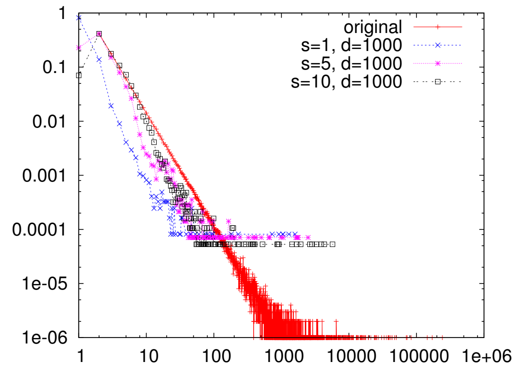
Graphs with fixed heterogeneous degree



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

Results

Graphs with fixed heterogeneous degree



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

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
$\rightarrow$power-law models are not reliable at this scale

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

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
$\rightarrow$power-law models are not reliable at this scale

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

$\rightarrow$ is the real distribution homogeneous?
$\rightarrow$ is the 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.

Experimentally, maximum degree observed > 1000
$\rightarrow$ ER graph with average degree $\simeq$ 1000? not realistic

Discussion (2/2)

Case of ER graphs

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

Experimentally, maximum degree observed > 1000
$\rightarrow$ ER graph with average degree $\simeq$ 1000? not realistic

$\rightarrow$real distribution probably heterogeneous. . .
need more work

Causes of the bias: first hypothesis

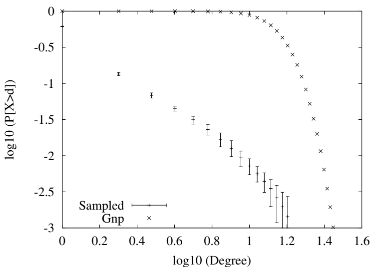
Hyp: Bias in the node sample?

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

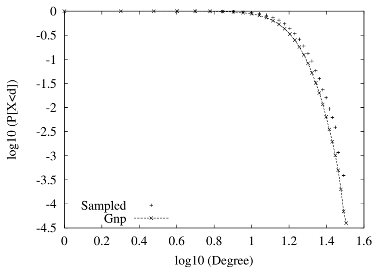
Causes of the bias: first hypothesis

Hyp: Bias in the node sample?

observed deg vs original deg



real deg vs original deg

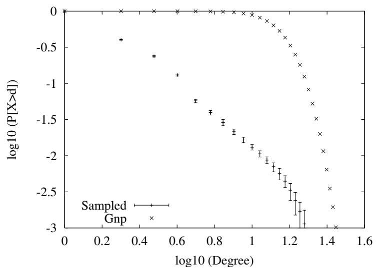


With 1 source

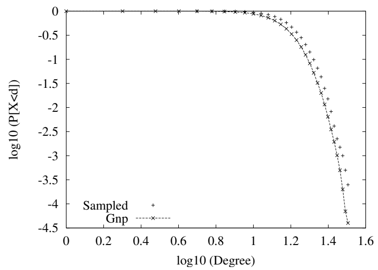
Causes of the bias: first hypothesis

Hyp: Bias in the node sample?

observed deg vs original deg



real deg vs original deg

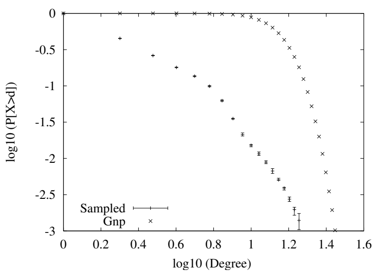


With 5 sources

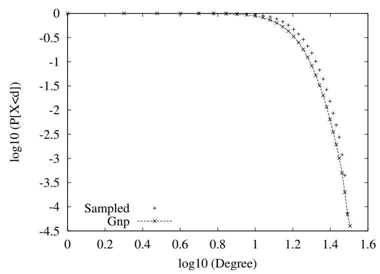
Causes of the bias: first hypothesis

Hyp: Bias in the node sample?

observed deg vs original deg



real deg vs original deg

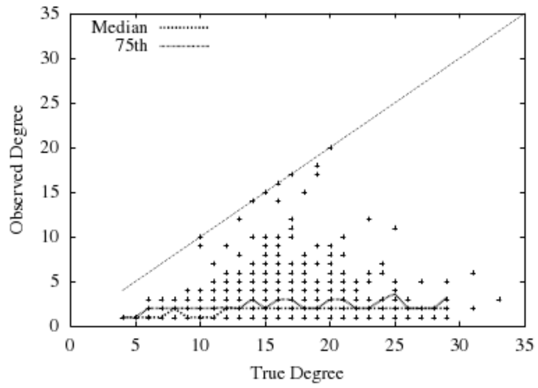


With 10 sources
Nodes are chosen without bias on the degree

Causes of the bias: second hypothesis

Hyp: Bias in the link sample?

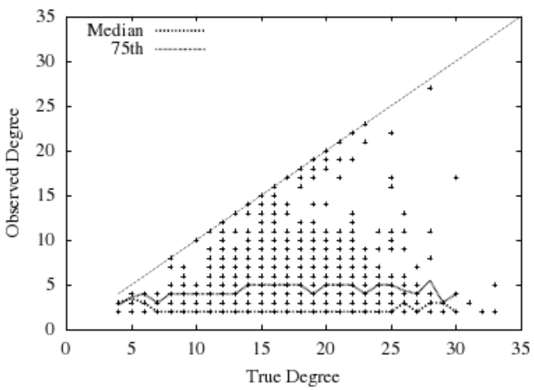
degree observed vs true degree
With 1 source



Causes of the bias: second hypothesis

Hyp: Bias in the link sample?

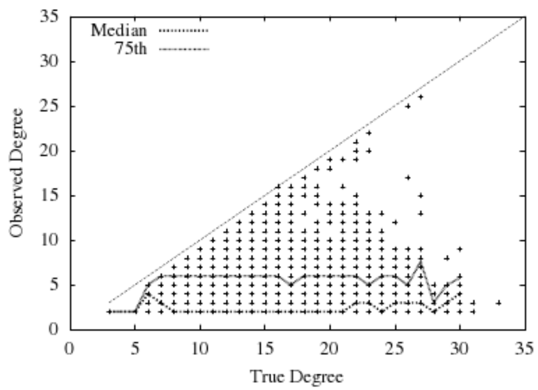
degree observed vs true degree
With 5 sources



Causes of the bias: second hypothesis

Hyp: Bias in the link sample?

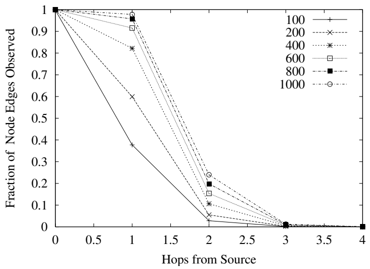
degree observed vs true degree
With 10 sources



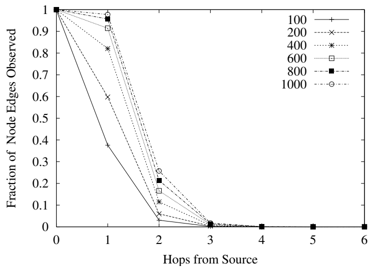
Bias sources

Link visibility as a function of their distance to the source

10,000 nodes



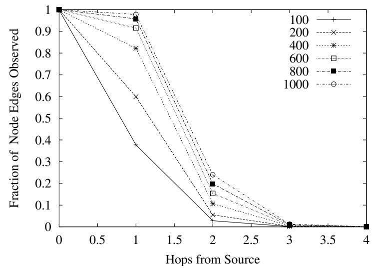
1,000,000 nodes



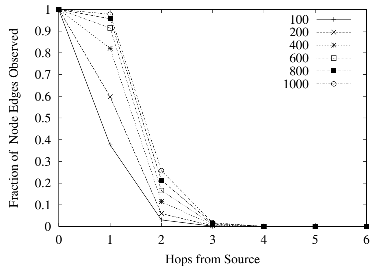
Bias sources

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

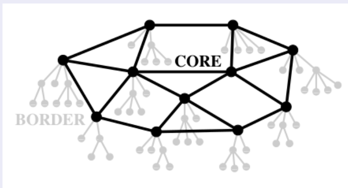
Outline

How to make an unbiased measurement?

Rigorous Measurement of the Internet Degree Distribution - *Latapy et al., 2017*

Core routers vs. border routers

- core: non-trivial routing
- border: routing in a tree

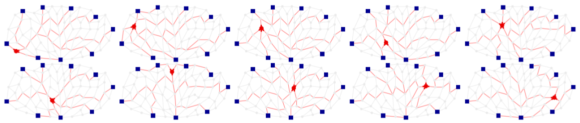


→ focus on core routers degree distribution

Uniform sampling of core routers: protocol

Measurement protocol basis

- 700 sources (PlanetLab monitors)
- 3 million destinations (IP addresses)
- UDP ping from all monitors to all destinations on unallocated port



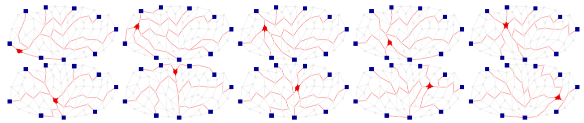
Information

- one of target interfaces sends back error message
  ⇒ one interface address of the router
- enough monitors ⇒ all core interfaces = router core-degree

Uniform sampling of core routers: protocol

Measurement protocol basis

- 700 sources (PlanetLab monitors)
- 3 million destinations (IP addresses)
- UDP ping from all monitors to all destinations on unallocated port



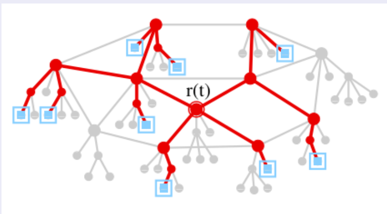
Information

- one of target interfaces sends back error message
  ⇒ one interface address of the router
- enough monitors ⇒ all core interfaces = router core-degree

Unbiased core degree distribution

Nodes filtering

- core nodes: all interfaces in the core are seen
- border nodes: monitors only detect the interface towards the core

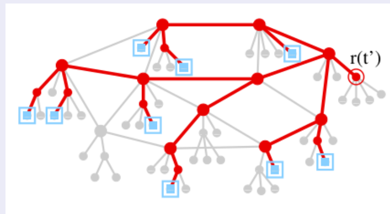


→ 5600 interfaces of reliable core routers

Unbiased core degree distribution

Nodes filtering

- core nodes: all interfaces in the core are seen
- border nodes: monitors only detect the interface towards the core

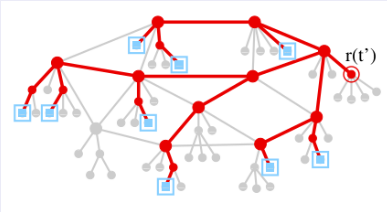


→ 5600 interfaces of reliable core routers

Unbiased core degree distribution

Nodes filtering

- core nodes: all interfaces in the core are seen
- border nodes: monitors only detect the interface towards the core



$\rightarrow$ 5600 interfaces of reliable core routers

Unbiased core degree distribution

During all measurements, we register the interfaces of a reliable router that send an error message

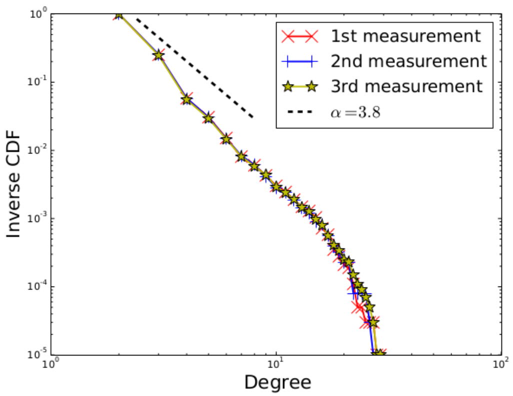
From the observed to the real distribution

Other biases to eliminate:

1. Discard border interfaces of core routers
2. Correct probability distribution as the probability to sample a router is proportional to its number of interfaces
   - router core-degree k: probability observed p'_k
   - real degree distribution $p_k \propto \frac{p'_k}{k}$

After all that...

(Core) degree distribution of core routers



Heterogeneous, but not a power-law