Sampling Biases in IP Topology Measurements

Anukool Lakhina

with **John Byers, Mark Crovella** and **Peng Xie**
Department of Computer Science
Boston University

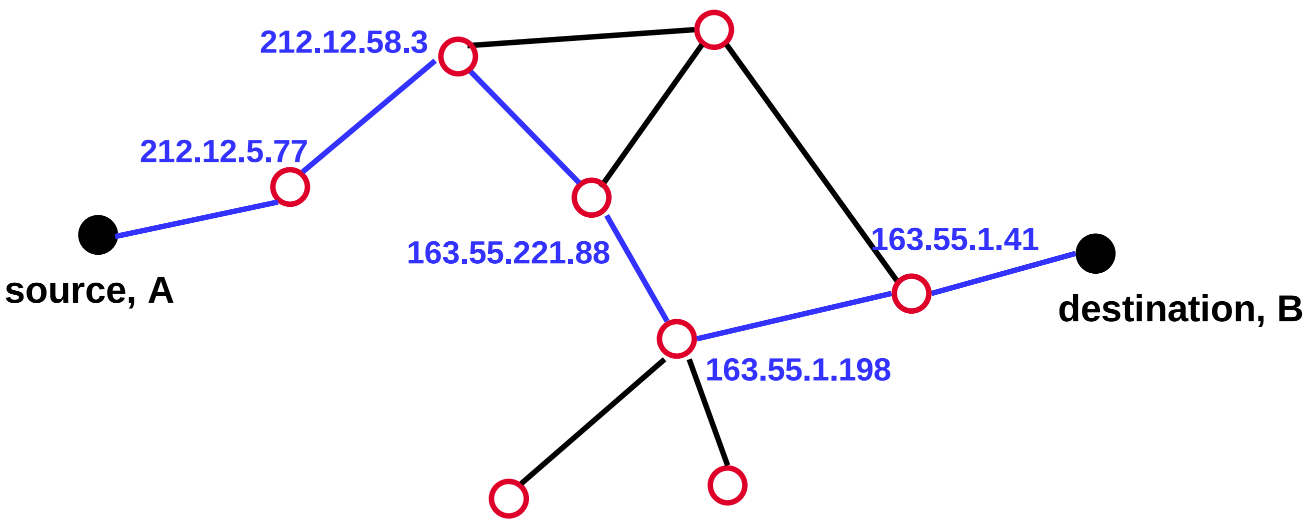
Computer Science

Discovering the Internet topology

Boston University
Computer Science

Goal: Discover the Internet Router Graph

- Vertices represent routers,
- Edges connect routers that are one IP hop apart

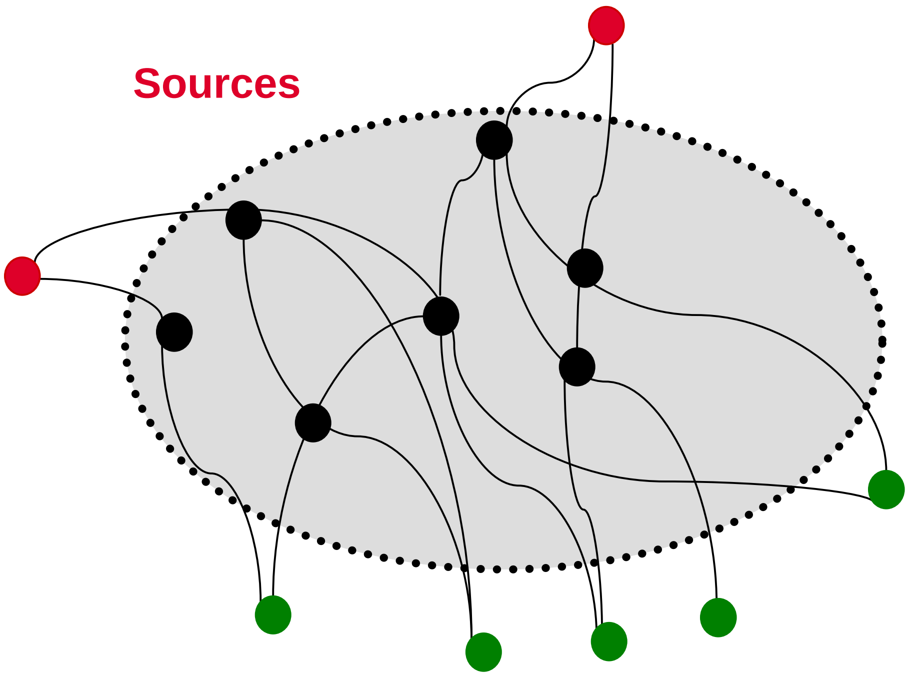


Measurement Primitive: `traceroute`

Reports the IP path from A to B i.e., how IP paths are overlaid on the router graph

Traceroute studies today

Boston University
Computer Science



- *k sources*: Few active sources, strategically located.
- *m destinations*: Many passive destinations, globally dispersed.
- Union of many traceroute paths.

(k,m)-traceroute study

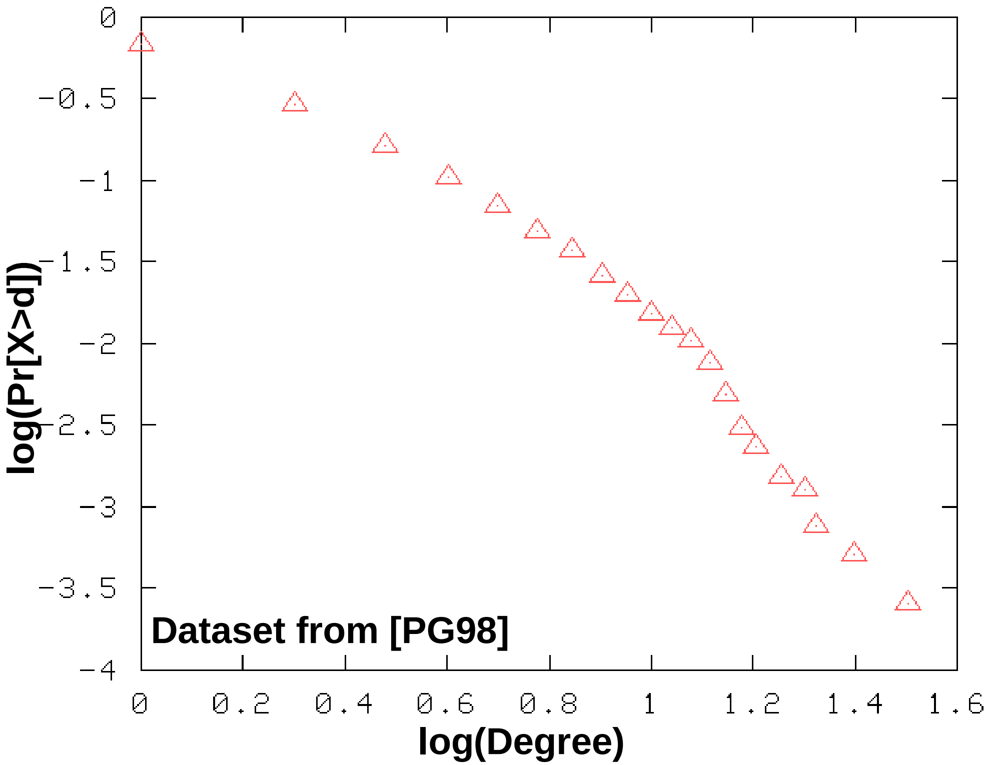
High Variability in node degrees

Computer Science

Degree distribution of routers found to be *highly variable* (degrees span several orders of magnitude).

Various studies have even concluded that the degree distribution has a power law tail,

$$\Pr[X > d] \propto d^{c}$$



[FFF99, GT00, BC01, …]

Our Question

- How reliable are *(k,m)-traceroute* methods in sampling graphs?

- We show that as a tool for measuring degree distribution, *(k,m)-traceroute* methods exhibit significant bias.

A thought experiment

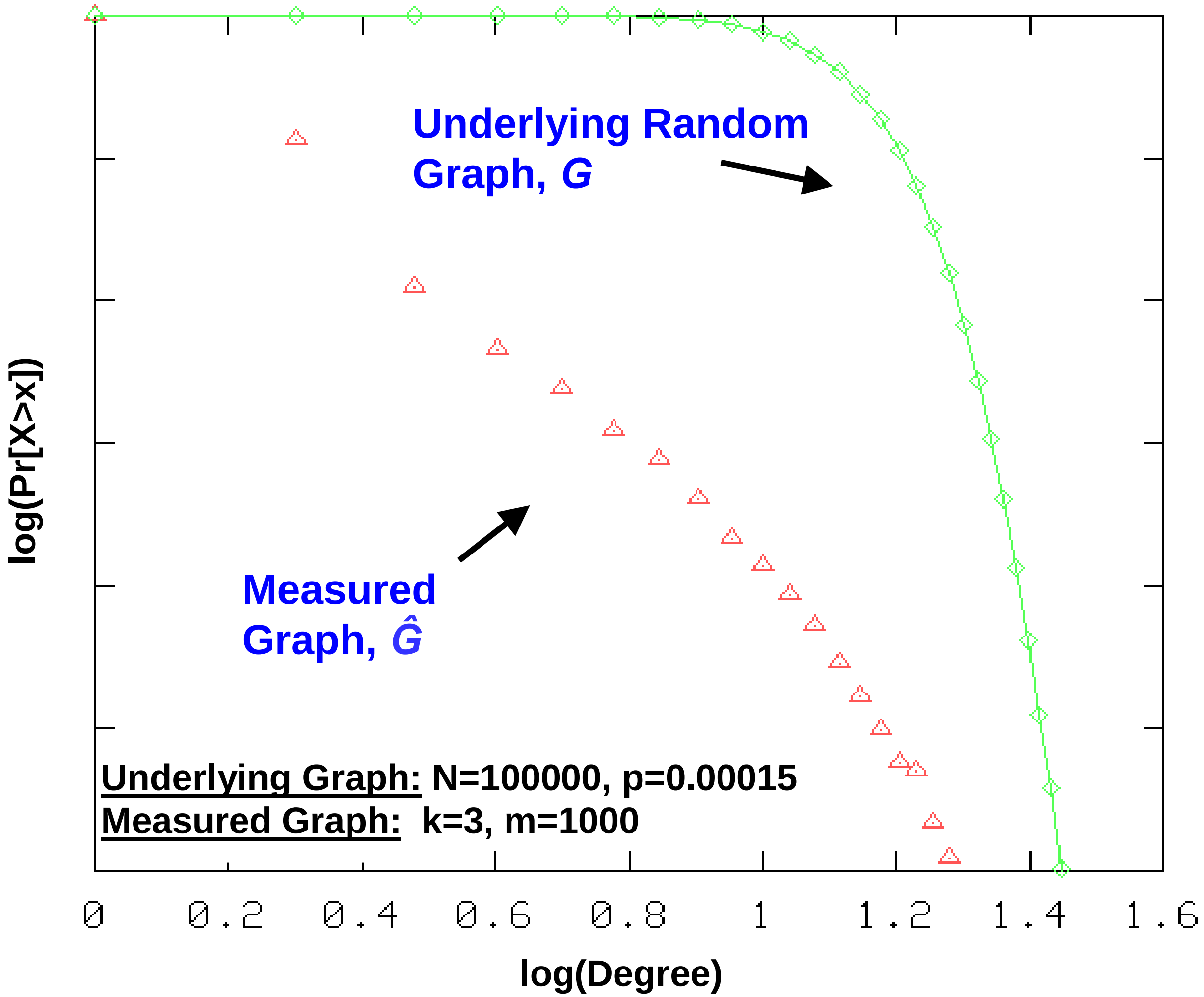
Computer Science

Idea: Simulate topology measurements on a random graph.

1. Generate a sparse Erdös-Rényi random graph, *G=(V,E).*
   Each edge present independently with probability *p*
   Assign weights: $w(e) = 1 + \varepsilon$, where ε in $\left[\frac{-1}{|V|}, \frac{1}{|V|}\right]$
2. Pick *k* unique source nodes, uniformly at random
3. Pick *m* unique destination nodes, uniformly at random
4. Simulate traceroute from *k* sources to *m* destinations, i.e. learn shortest paths between *k* sources and *m* destinations.
5. Let $\hat{G}$ be union of shortest paths.

Ask: How does $\hat{G}$ compare with *G* ?

Computer Science



Ĝ is a *biased sample* of *G* with a dramatically different degree distribution.

Can "high variability" be a *measurement artifact*?

Outline

- Motivation and Thought Experiments
- Understanding Bias on Simulated Topologies
- Detecting Bias in Simulated Scenarios
  - Statistical hypotheses to infer presence of bias
- Examining Internet Maps

Understanding Bias

(k,m)-traceroute sampling of graphs is biased

An intuitive explanation:

When traces are run from few sources to large destinations, some portions of underlying graph are explored more than others.

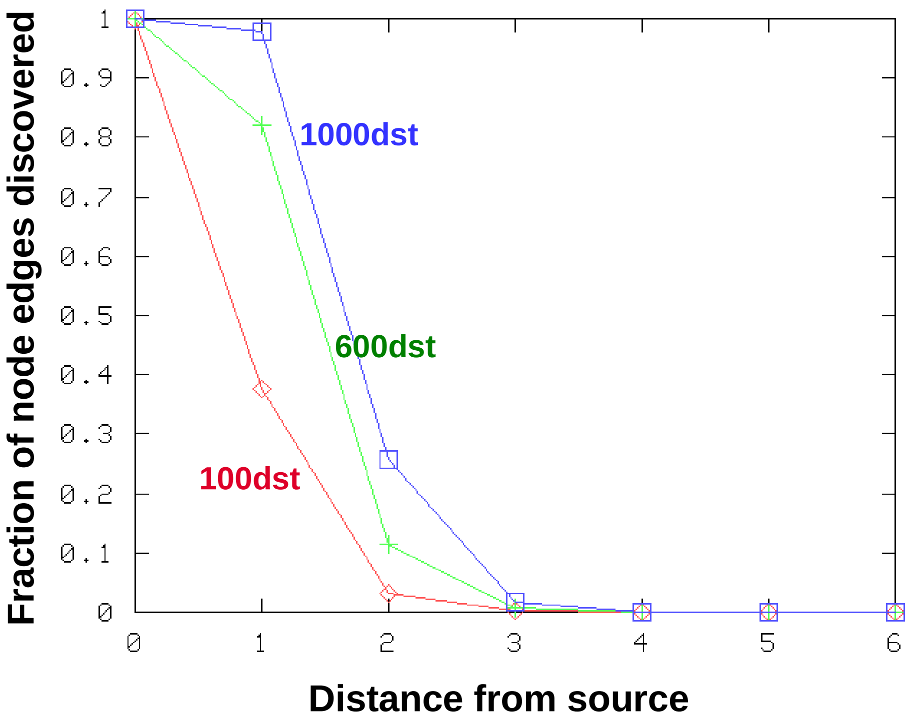
Edges incident to a node in $\hat{G}$ are sampled disproportionately.

Analyzing nonuniform edge sampling

Boston University
Computer Science

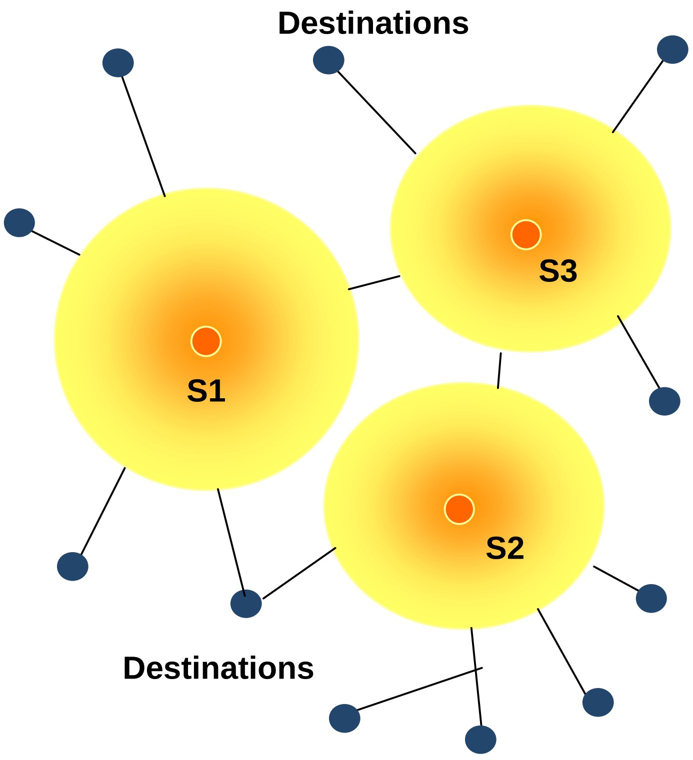
- Question:
  Given some vertex in $\hat{G}$ that is h hops from the source, what fraction of its true edges are contained in $\hat{G}$?

- Analysis reveals that:
  As h increases, fraction of edges discovered falls off sharply.



What does this suggest?

Edges close to the source are sampled more often than edges further away.



Intuitive Picture:

Neighborhood near sources is well explored but, this visibility falls with hop distance from sources.

Inferring Bias

Computer Science

Goal:

Given a measured $\hat{G}$, is it a biased sample?

Why this is difficult:

Don’t have underlying graph.

Don’t have criteria for checking bias.

General Approach:

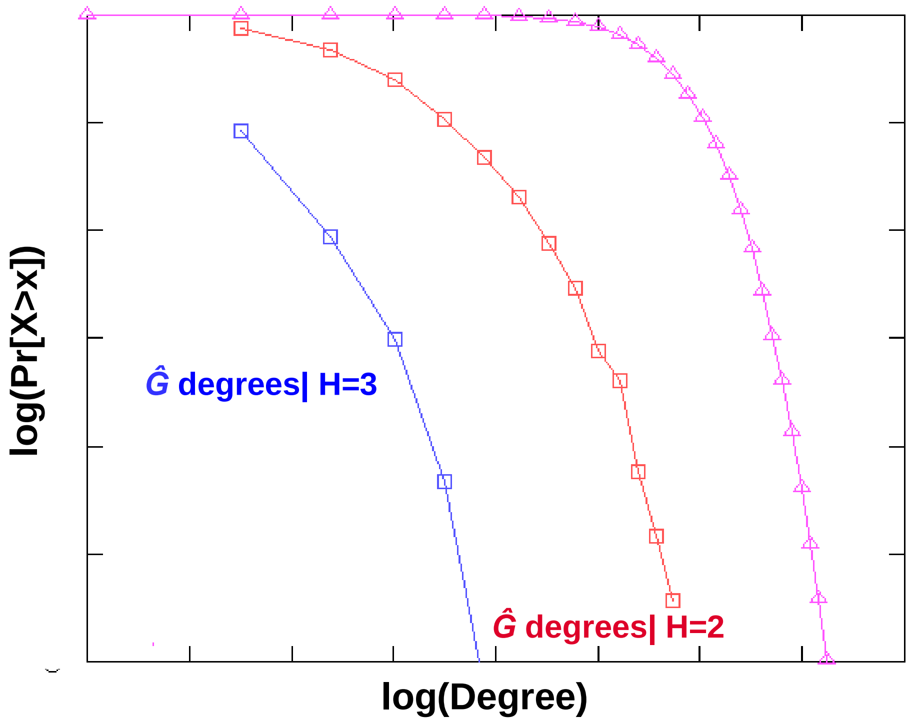
Examine statistical properties as a function of distance from nearest source.

- Unbiased sample → No change
- Change → Bias

Towards Detecting Bias

Computer Science

Examine *Pr[D|H]*, the conditional probability that a node has degree *d*, given that it is at distance *h* from the source.



Two observations:

1. Highest degree nodes are near the source.
2. Degree distribution of nodes near the source differs from those further away.

A Statistical Test for C1

C1: Are the highest-degree nodes near the source?
If so, then consistent with bias.

H_0^{C1} The 1% highest degree nodes occur at random with distance to nearest source.

Cut vertex set in half: ***N*** (near) and ***F*** (far), by distance from nearest source.

Let ***v*** : (0.01) |V|
k : fraction of ***v*** highest-degree nodes that lie in ***N***

Can bound likelihood ***k*** deviates from *1/2* using *Chernoff-bounds*:

$$\Pr[k > \tfrac{(1+\delta)}{2}] < \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\frac{v}{2}}$$

Reject null hypothesis with confidence *1-*α if:

$$\alpha \geq \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\frac{v}{2}}$$

A Statistical Test for C2

C2: Is the degree distribution of nodes near the source different from those further away? If so, consistent with bias.

H_0^{C2} Degree distribution of nodes near the source is consistent with that of all nodes.

Compare degree distribution of nodes in ***N*** and ***Ĝ***, using the *Chi-Square Test*:

$$\chi^2 = \sum_{i=1}^{l} (O_i - E_i)^2 / E_i$$

where *O* and *E* are observed and expected degree frequencies and *l* is histogram bin size.

Reject hypothesis with confidence *1-α* if:

$$\chi^2 > \chi^2_{[\alpha, l-1]}$$

Our Definition of Bias

Computer Science

- Bias (Definition):
  Failure of a sampled graph to meet statistical tests for randomness associated with *C1* and *C2*.

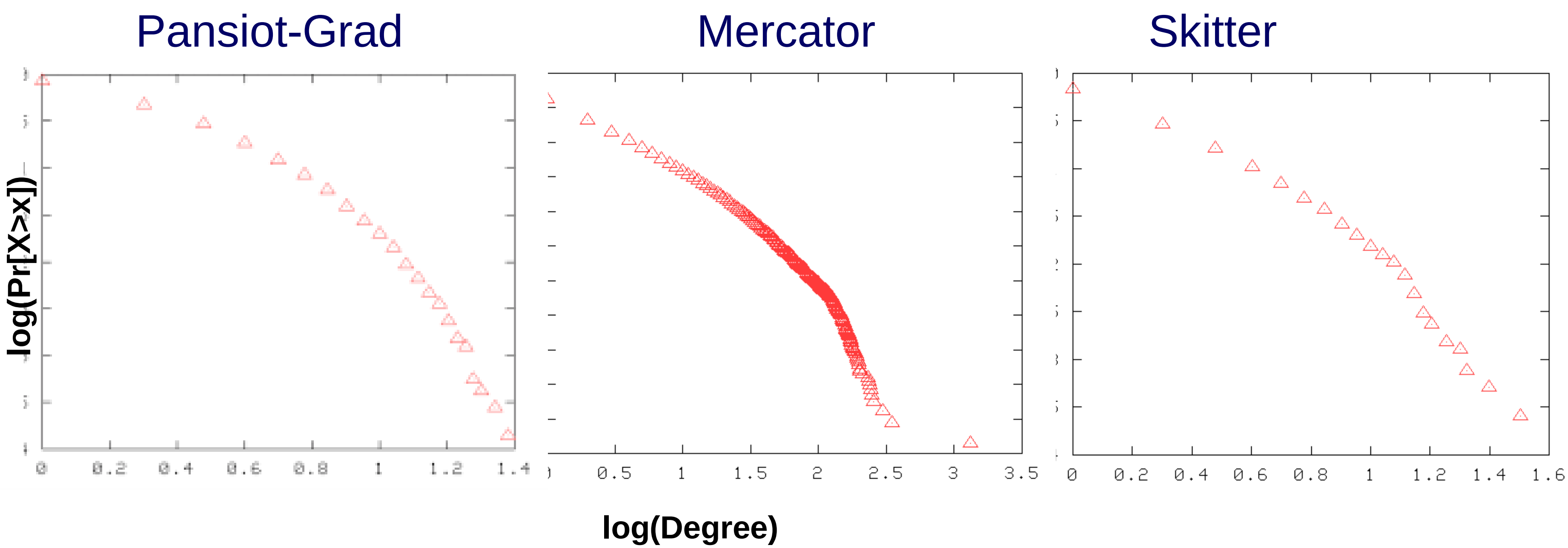
- Disclaimer:
  Tests are binary and don't tell us *how* biased datasets are.

- A dataset that fails both tests is a poor choice for making generalizations about underlying graph.

Introducing datasets

Computer Science

Dataset Name	Date	# Nodes	# Links	# Srcs	# Dsts	Reference
Pansiot-Grad	1995	3,888	4,857	12	1270	PG98
Mercator	1999	228,263	320,149	1	NA	GT00
Skitter	2000	7,202	11,575	8	1277	BBBC01



Testing C1

H_0^{C1} The 1% highest degree nodes occur at random with distance to source.

Dataset	ν	κ	Chernoff Bound	$\mathcal{H}_0^{C1}$
Pansiot-Grad	41	38	2×10^{-4}	Reject
Mercator Routers	2,290	2,065	10^{-172}	Reject
Skitter Routers	104	87	9×10^{-7}	Reject

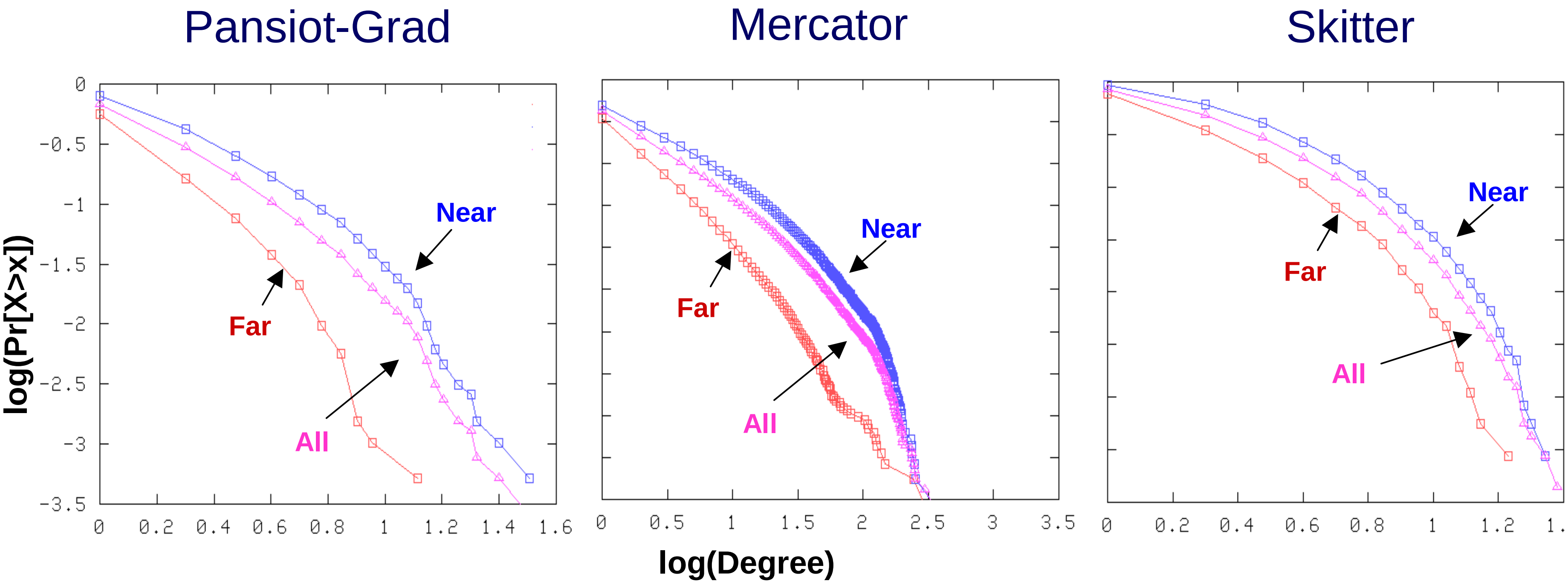
Pansiot-Grad: 93% of the highest degree nodes are in ***N***
Mercator: 90% of the highest degree nodes are in ***N***
Skitter: 84% of the highest degree nodes are in ***N***

Testing C2

Computer Science

H_0^{C2} Degree distribution of nodes near the source is consistent with that of all nodes.

Dataset	ℓ	α	$\chi^2_{[1-\alpha;\ell-1]}$	χ^2	$\mathcal{H}_0^{C2}$
Pansiot-Grad	17	0.005	35.72	1082.0	Reject
Mercator Routers	123	0.005	167.4	59729	Reject
Skitter Routers	19	0.005	23.59	1965	Reject



Summary of Statistical Tests

For all datasets, we reject both null hypotheses of “no bias”.

We conclude that it is likely that *true* degree distribution of sampled routers is different than what is shown in these datasets.

Final Remarks

- Using (k,m)-traceroute methods to discover Internet topology yields biased samples.
- Rocketfuel [SMW:02] may avoid some pitfalls of (k,m)-traceroute studies but is *limited-scale*
- One open question: How to sample the degree of a router at random?