Bias reduction for traceroute sampling: towards a more accurate map of the internet

Abraham D. Flaxman,
Microsoft Research

Juan Vera,
University of Waterloo

December 11, 2007

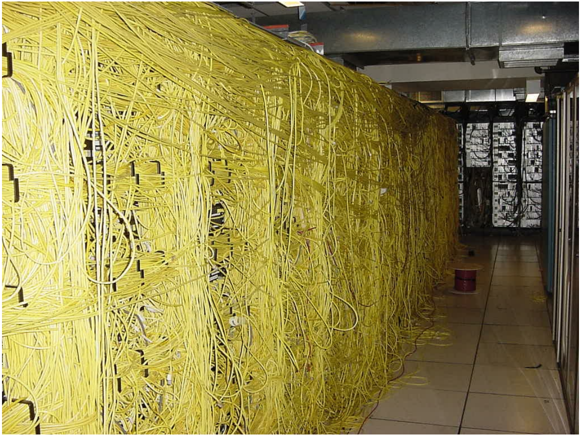
Outline

Networks in the real world

- Real-world networks are complex

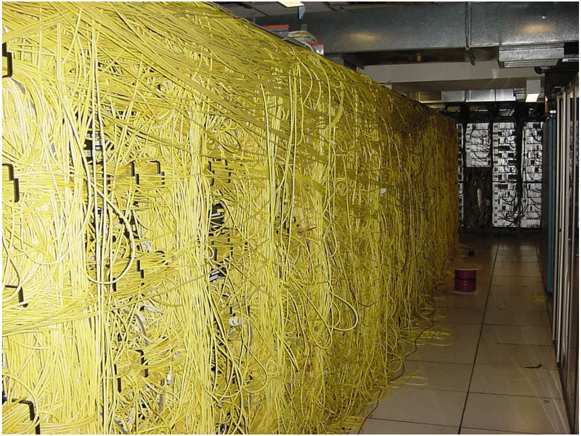
Networks in the real world

- Real-world networks are complex



Networks in the real world

- Real-world networks are complex



- So complex, the structure has not been recorded.

A famous example

- The Internet.

A famous example

- The Internet.

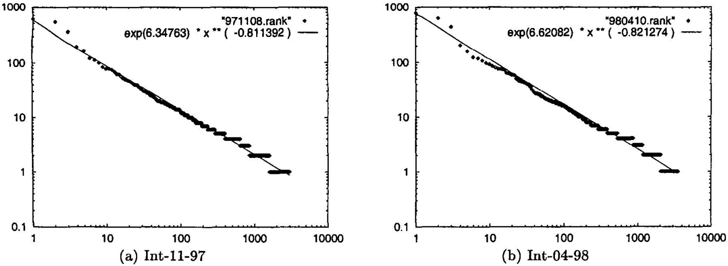
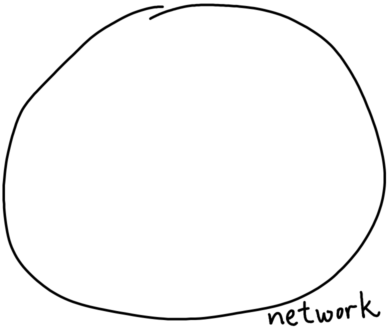


Figure 3: The rank plots. Log-log plot of the outdegree d_v versus the rank r_v in the sequence of decreasing outdegree.

- Measurements of the autonomous systems (AS) graph of the Internet in 1998 showed that the degree distribution follows a power law.
  [M. Faloutsos, P. Faloutsos, C. Faloutsos, 1999]

Where did the data come from?

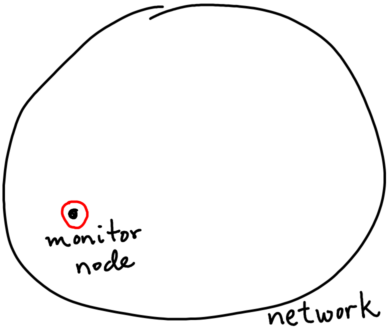
A theorist's sketch:



(Note: This is certainly simplified. I would love a theorist-friendly lesson in what is actually going on.)

Where did the data come from?

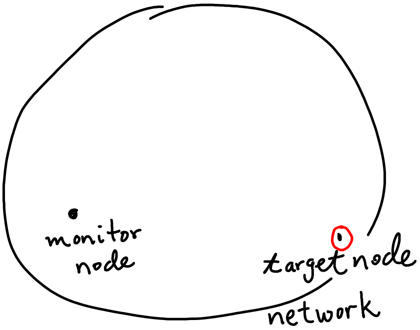
A theorist's sketch:



(Note: This is certainly simplified. I would love a theorist-friendly lesson in what is actually going on.)

Where did the data come from?

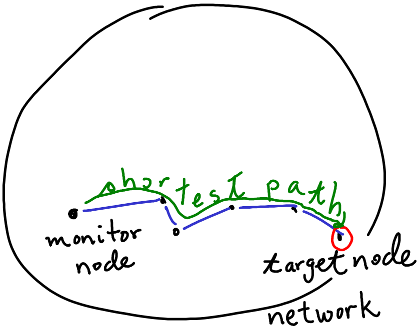
A theorist's sketch:



(Note: This is certainly simplified. I would love a theorist-friendly lesson in what is actually going on.)

Where did the data come from?

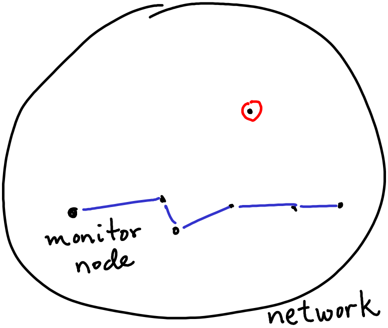
A theorist's sketch:



(Note: This is certainly simplified. I would love a theorist-friendly lesson in what is actually going on.)

Where did the data come from?

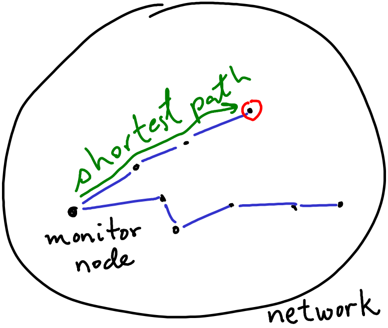
A theorist's sketch:



(Note: This is certainly simplified. I would love a theorist-friendly lesson in what is actually going on.)

Where did the data come from?

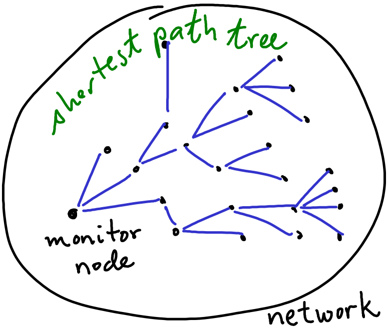
A theorist's sketch:



(Note: This is certainly simplified. I would love a theorist-friendly lesson in what is actually going on.)

Where did the data come from?

A theorist's sketch:



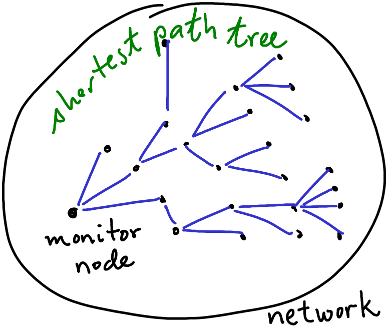
(Note: This is certainly simplified. I would love a theorist-friendly lesson in what is actually going on.)

Traceroute sampling has a problem: bias

What bias is introduced during traceroute sampling, and can we correct it?

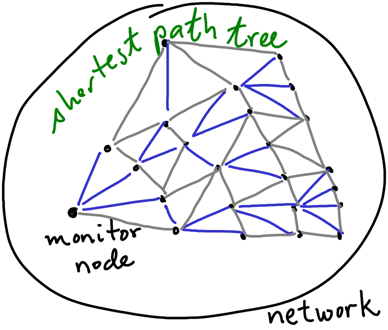
Example: Counting triangles

How many triangles are in this graph?



Example: Counting triangles

How many triangles are in this graph?



- Many present, none seen.

Rest of this talk: Degree distribution

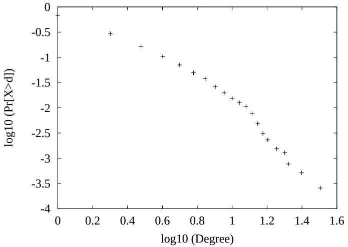
Rest of this talk: Degree distribution

- A formal definition: The ccdf of the *degree distribution*, is given by

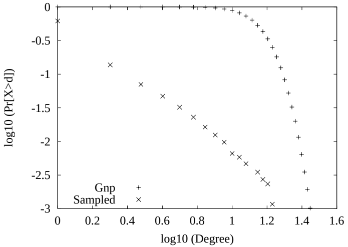
$$\overline{F}(k) = \Pr[\deg(u) > k] = \frac{\#\{v \in V : \deg(v) > k\}}{|V|},$$

where the vertex u is chosen uniformly at random from V.

Prior work: degree distribution



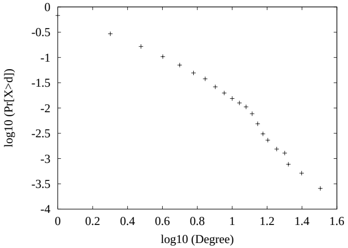
(a) Pansiot-Grad



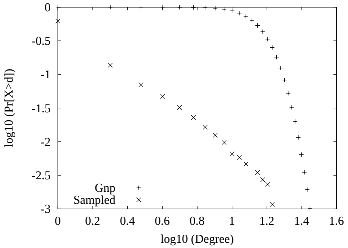
(b) Subgraph sampled from $G_{N,p}$

- A. Lakhina, J. W. Byers, M. Crovella, P. Xie, Sampling Biases in IP Topology Measurements, INFOCOMM 2003.

Prior work: degree distribution



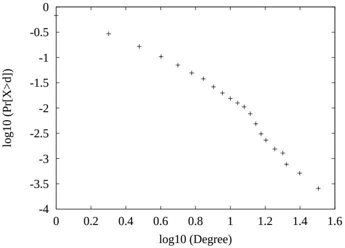
(a) Pansiot-Grad



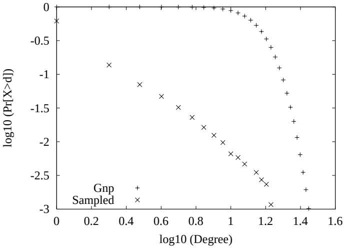
(b) Subgraph sampled from $G_{N,p}$

- A. Lakhina, J. W. Byers, M. Crovella, P. Xie, Sampling Biases in IP Topology Measurements, INFOCOMM 2003.
- A. Clauset and C. Moore, Phys Review Letters 2005
  Petermann and de los Rios, Euro Phys Journal 2004

Prior work: degree distribution



(a) Pansiot-Grad



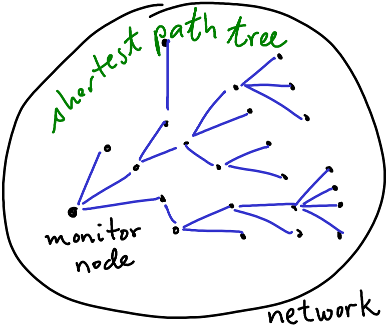
(b) Subgraph sampled from $G_{N,p}$

- A. Lakhina, J. W. Byers, M. Crovella, P. Xie, Sampling Biases in IP Topology Measurements, INFOCOMM 2003.
- A. Clauset and C. Moore, Phys Review Letters 2005
  Petermann and de los Rios, Euro Phys Journal 2004
- D. Achlioptas, A. Clauset, D. Kempe, C. Moore, On the Bias of Traceroute Sampling, STOC 2005.

Does the Internet really have a power-law?

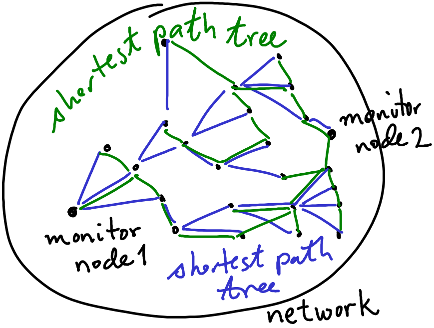
Perhaps we shouldn't have been so certain.

Use more than one monitor node



This is what the experimentalists have been doing for years.

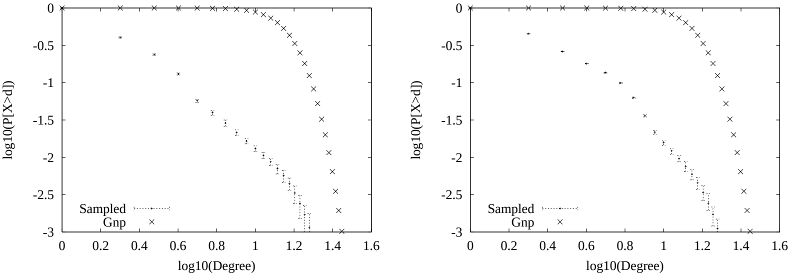
Use more than one monitor node



This is what the experimentalists have been doing for years.

Using more monitors

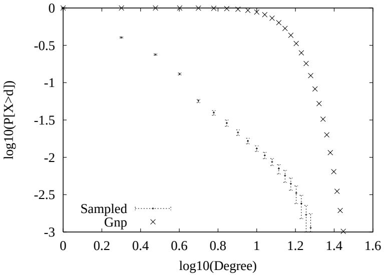
- Lakhina *et al* show, by computer experiment, union of edges from more monitors may not help in degree distribution estimation.



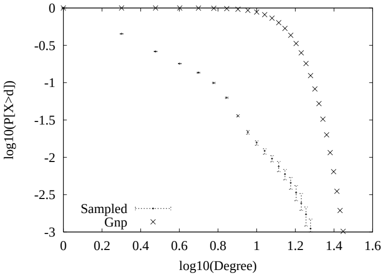
(b) 5 sources, 1000 destinations　　(c) 10 sources, 1000 destinations

Using more monitors

- Lakhina *et al* show, by computer experiment, union of edges from more monitors may not help in degree distribution estimation.



(b) 5 sources, 1000 destinations



(c) 10 sources, 1000 destinations

- There is something better than taking the union of the edges.

Multiple-recapture population estimates

How many fish are in the sea?

Multiple-recapture population estimates

How many fish are in the sea?

- Go out one day, catch all the fish you can, tag and release.

Multiple-recapture population estimates

How many fish are in the sea?

- Go out one day, catch all the fish you can, tag and release.
- Next day, go out again, catch fish again.

Multiple-recapture population estimates

How many fish are in the sea?

- Go out one day, catch all the fish you can, tag and release.
- Next day, go out again, catch fish again.
- Record number of fish:
  - caught first day, A;
  - caught second day, B;
  - caught first day and again second day, C.

Multiple-recapture population estimates

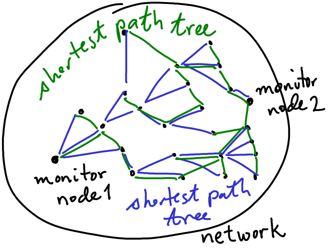
How many fish are in the sea?

- Go out one day, catch all the fish you can, tag and release.
- Next day, go out again, catch fish again.
- Record number of fish:
  - caught first day, A;
  - caught second day, B;
  - caught first day and again second day, C.
- Estimate total number of fish: $\widehat{N} = \frac{A \cdot B}{C}$ (*Petersen estimate*).
  [C. J. G. Petersen, The yearly immigration of young plaice into the Limfjord from the German sea, 1896.]

Application to bias reduction in traceroute sampling

Apply Petersen estimate repeatedly to estimate the degree of each node:

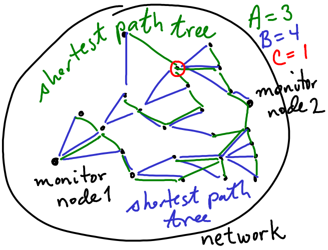
- Population to count is number of edges incident to a fixed vertex v
- Fish are edges
- Days are monitor nodes
- We catch an edge if it is on shortest-path tree rooted at the day



Application to bias reduction in traceroute sampling

Apply Petersen estimate repeatedly to estimate the degree of each node:

- Population to count is number of edges incident to a fixed vertex *v*
- Fish are edges
- Days are monitor nodes
- We catch an edge if it is on shortest-path tree rooted at the day



When is this a good estimate?

- If the animals caught on a given day are i.i.d., this is the maximum likelihood estimator, and it is asymptotically unbiased.

When is this a good estimate?

- If the animals caught on a given day are i.i.d., this is the maximum likelihood estimator, and it is asymptotically unbiased.
- In zoology (and in the United States Census), the validity of this assumption is debated.

When is this a good estimate?

- If the animals caught on a given day are i.i.d., this is the maximum likelihood estimator, and it is asymptotically unbiased.
- In zoology (and in the United States Census), the validity of this assumption is debated.
- In traceroute sampling, there is no debate; the assumption does not hold.

Never surrender

- We won’t give up, though.

Never surrender

- We won't give up, though.
- Experiments and theory led us to estimator

$$\widehat{\deg}_{s,t}(u) = \begin{cases} \frac{|N_s(u)|\cdot|N_t(u)|}{|N_s(u)\cap N_t(u)|}, & \text{if } |N_s(u) \cap N_t(u)| > 2; \\ \infty, & \text{otherwise;} \end{cases}$$

where $N_s(u)$ is the neighborhood of u in the shortest-path tree rooted at s.

Never surrender

- We won't give up, though.
- Experiments and theory led us to estimator

$$\widehat{\deg}_{s,t}(u) = \begin{cases} \frac{|N_s(u)|\cdot|N_t(u)|}{|N_s(u)\cap N_t(u)|}, & \text{if } |N_s(u) \cap N_t(u)| > 2; \\ \infty, & \text{otherwise;} \end{cases}$$

  where $N_s(u)$ is the neighborhood of u in the shortest-path tree rooted at s.
- Rigorous proof: the estimator is asymptotically unbiased on $G_{n,p}$, for $p > \frac{\log n}{n}$.

Never surrender

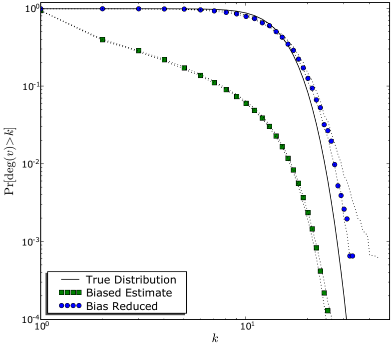
- We won't give up, though.
- Experiments and theory led us to estimator

$$\widehat{\deg}_{s,t}(u) = \begin{cases} \frac{|N_s(u)|\cdot|N_t(u)|}{|N_s(u)\cap N_t(u)|}, & \text{if } |N_s(u)\cap N_t(u)| > 2; \\ \infty, & \text{otherwise;} \end{cases}$$

  where $N_s(u)$ is the neighborhood of u in the shortest-path tree rooted at s.
- Rigorous proof: the estimator is asymptotically unbiased on $G_{n,p}$, for $p > \frac{\log n}{n}$.
- Therefore, can reject a null hypothesis that the sampled graph is an Erdős-Rényi graph.

Experimental results

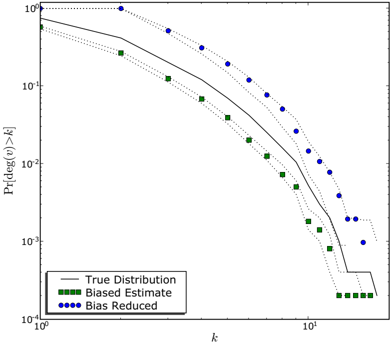
It's hard to prove things about 2 different shortest path trees on the same graph, so we also used simulations.



(a) $G_{n,m}$ with $n = 100,000$, $d = 2m/n = 15$.

Experimental results

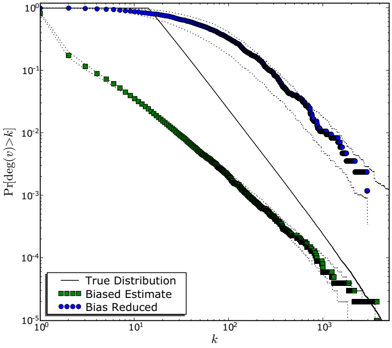
It's hard to prove things about 2 different shortest path trees on the same graph, so we also used simulations.



(d) Western states power graph from [22].

Note, in particular, the PA graph

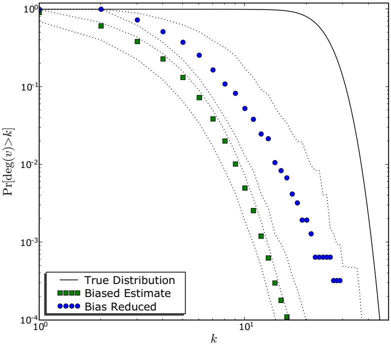
It certainly changes the shape of the degree distribution if you're dealing with a preferential attachment graph.



(b) PA graph with $n = 100,000$, $m = 15$.

Conjectures

When does this work?
When neighborhoods are "sufficiently random". When there are not many triangles or other small cycles?



(c) $G(\mathcal{X}; r)$ with $n = 100,000$, $d = \pi r^2 = 25$.

Should bias reduction work on the Internet?

- The point of this talk is, be skeptical of anyone who says they really know.
- It works well on the Western States Power Grid, which is another network of “things connected with wires”.

Should bias reduction work on the Internet?

- The point of this talk is, be skeptical of anyone who says they really know.
- It works well on the Western States Power Grid, which is another network of “things connected with wires”.
- Let’s see what happens.

Going out on a limb

Effects of bias reduction in the AS graph, and the possible changes in degree distribution following technological trends.

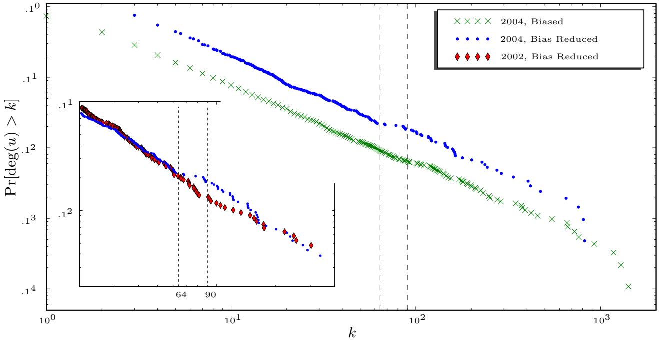


Fig. 1. Degree sequence ccdf estimates for the AS graph (from CAIDA skitter). Main panel: March, 2004, with and without bias reduction. Inset: a portion of ccdf for March, 2004 and March, 2002, both with bias reduction. The nodes with degree between 65 and 90 in 2002 have disappeared in 2004.

Conclusion and future work

- Be careful, know your data.
- Statisticians have developed all kinds of techniques for going beyond assumptions of i.i.d. fishes. Can they be applied here?
- Find ways to apply bias reduction to other network statistics and other network sampling methods, e.g. PageRank computation or other centrality measures.