On the bias of traceroute sampling

Aaron Clauset
with Dimitris Achlioptas, David Kempe and Cristopher Moore

at MSRI Real-World Random Networks Workshop
April 18 2004
Internet Maps

Nov 23 2003, Opte Project
Internet Maps

apparent power-law degree distribution

\[ P(k) \sim k^{-\alpha} \]

\[ 2 < \alpha < 3 \]

excerpted from Govindan et al. (2000)
Some Problems

Topology must be inferred

- router links are not directly visible (unlike the Web)
- links are queried indirectly, e.g., by traceroutes
- Faloutsos et al. (1999) uses a single source

But, traceroutes are biased

- yields power laws even in homogeneous networks
- for $G(n,p)$ empirically (Lakhina et al., 2003), analytically (Clauset and Moore, 2005)
- traceroutes overlook distant edges
Some Results

• Idealize single-source traceroute study as a breadth-first spanning (BFS) tree; assume shortest path.

• Degree distribution of BFS tree = network’s observed degree distribution.

• Model BFS as a continuous-time queue process.

• For (nearly) arbitrary random graphs, can derive observed degree sequence \( \{a^\text{obs}_k\} \) from the underlying degree sequence \( \{a_k\} \).
Some Results

• **Expressed using generating functions**

Input: \( g(z) = \sum_{j=0}^{\infty} a_j z^j \)

Output: \( g^{\text{obs}}(z) = z \int_{0}^{1} g'[t - \frac{(1 - z)}{g'(1)} g'(\frac{g'(t)}{g'(1)})] \, dt \)

• **Examples**: random graphs, and \( \delta \)-regular graphs (both yield power laws \( P(k) \sim k^{-1} \))
This Talk

- Sketch expectation results (concentration results in paper (STOC 2005))
- Example networks
- Discuss implications and future work
Continuous-time Model

- Given a “reasonable” degree sequence \( \{a_k\} \)

\[
\begin{align*}
C &> 0 \\
\alpha &> 2 \\
a_k &< C \cdot k^{-\alpha} \\
\delta &= \sum k \cdot a_k
\end{align*}
\]

we build a random multigraph (configuration model) “on the fly,” as in Kim’s Poisson cloning model:

- give each vertex copy a real-valued index \( x \) chosen uniformly on \([0,1]\), and

- use copy-level BFS algorithm
Some Terminology

- *Copies of the same vertex are siblings*

- *Exposed copies have known partners*
Some Terminology

- **Enqueued copies**: at boundary of the BFS tree $T$; at least one sibling has been exposed.

- **Untouched copies**: no sibling has been exposed; entire vertex is outside $T$

- An edge $(u, v)$ is visible only if copy $u$ is matched with an untouched copy $v$
Copy-level BFS

while Q is non-empty do

Pop a copy u from the head of Q

Match u to the unexposed v with largest index t

if v is untouched then

Add the edge (u,v) to T

Append v’s siblings to Q

else

Remove v from Q
Some Comments

- View as continuous-time process, with $t$ decreasing from 1 to 0
- At time $t$, copy at head of $Q$ is matched to copy with index $t$
- So, indices of unmatched copies are independent and uniform on $[0,t)$ (including the copy at the head)
Observed degree distribution

- Given by

\[ E[a_{m+1}^{\text{obs}}] = \sum_i a_i \rho_{i,m} \]

- Let \( \rho_{i,m}(t) \) be the prob. that we observe \( m \) edges of a vertex with degree \( i \), if its maximum index is \( t \)

- So, as a function of time \( t \),

\[ \rho_{i,m} = \int_0^1 it^{i-1} \rho_{i,m}(t) dt \]
Edge Visibility

- To calculate $\rho_{i,m}(t)$, need to know prob. $p_{\text{vis}}$ that a single edge appears in $T$, i.e., it is visible.

- Given $p_{\text{vis}}$, and assuming independence

$$
\rho_{i,m}(t) = \binom{i-1}{m} p_{\text{vis}}(t)^m (1 - p_{\text{vis}}(t))^{i-1-m}
$$
Edge Visibility

- Given $v$ with index $t$, what is prob. that its sibling $v'$ will result in the visible edge $(v',w)$?

- This happens iff
  1. $v'$ gets to head of $Q$, and
  2. its partner $w$ is still unmatched

- These are same as, at time $t$
  1. $w$ is untouched, and
  2. each of $w$’s siblings’ partners untouched
Edge Visibility

- Prob. that $w$ is untouched at time $t$:

$$p_{unto}(t) = \frac{c_{unto}(t)}{c_{unex}(t)} = \frac{\sum_j j a_j t^j}{\delta t^2}$$

and prob. that $w$'s vertex has degree $k$

$$p_{unto,k}(t) = \frac{k a_k t^k}{c_{unex}(t)} = \frac{k a_k t^k}{\sum_j j a_j t^j}$$
• Requiring that all $k$ siblings are untouched yields (assuming local tree-structure)

\[ p_{\text{vis}}(t) = \sum_k p_{\text{unto},k}(t) p_{\text{unto}}(t)^k \]
Observed degree distribution

• Thus, putting everything together yields:

\[ a_{m+1}^{\text{obs}} = \sum_i a_i \left[ \int_0^1 i t^{i-1} \binom{i-1}{m} p_{\text{vis}}(t)^m (1 - p_{\text{vis}}(t))^{i-1-m} \, dt \right] \]
As generating functions

- Input: gen. func. for \( \{a_k\} \)

\[
g(z) = \sum_{k=0}^{\infty} a_j z^j
\]

- Output: gen. func. for \( \{a_{k}^{\text{obs}}\} \)

\[
g^{\text{obs}}(z) = z \int_0^1 \left[ t - \frac{(1 - z)}{g'(1)} g'\left( \frac{g'(t)}{g'(1)} \right) \right] \, dt
\]
Example: $\delta$-Regular Graphs

- **Given:** $g(z) = z^\delta$, 

$$g^{\text{obs}}(z) = z\delta \cdot \int_0^1 t^{\delta-1}(1 - (1 - z)t^{\delta(\delta-2)})^{\delta-1} dt$$

$$= z \cdot \frac{\Gamma(\delta)}{\Gamma(\delta - \frac{1}{\delta-2})(\delta - 2)} \sum_{m=0}^{\infty} \frac{\Gamma\left(m + \frac{1}{\delta - 2}\right)}{m!} z^m$$

- **In limit of** $\delta \to \infty$, the coefficients become

$$a_{m+1}^{\text{obs}} \sim m^{-1}$$
Example: \( G(n, p = \delta/n) \)

- **Given**: \( g(z) = e^{-\delta(1-z)} \),

\[
g^{\text{obs}}(z) = z\delta \cdot \int_{t_0}^{1} e^{-\delta(1-t)} e^{-\delta(1-z)} e^{-\delta(1-e^{-\delta(1-t)})} \, dt
\]

\[
= \sum_{m=0}^{\infty} \frac{z^{m+1}}{\delta m!} \left( \Gamma(m, \delta e^{-\delta(1-t_0)}) - \Gamma(m, \delta) \right)
\]

- **In limit of** \( \delta \to \infty \), the coefficients become

\[
a^{\text{obs}}_{m+1} \sim \frac{1}{\delta m}
\]
Conclusions

We have

- developed fully analytic framework for single-source traceroute studies on (nearly) arbitrary random graphs
- framed results using generating functions
- given examples of regular and \(G(n,p)\) graphs
Future Directions

• *Power-law degree distributions*
  
  • *given underlying exponent, want observed exponent*
  
  • *difficult to get $\alpha^{obs}$ out of integral*

• *Inverting relationship: given $\{a^{obs}_k\}$, get $\{a_k\}$*

• *Multi-source traceroute studies*

• *More realistic analysis: geographic structure of Internet*
Acknowledgements

• Cristopher Moore, David Kempe, Dimitris Achlioptas, Jon Kleinberg, André Allavena and Tracy Conrad

• NSF grants PHY-0200909, ITR-0324845, CCR-0220070, EIA-0218563

• Accepted to ACM Symp. Theory of Comp. (STOC)

• Preprint at cond-mat/0503087