Catching the head, tail, and everything in between: a streaming algorithm for the degree distribution

ICDM 2015
Online and Incremental Learning Session
Atlantic City, New Jersey

OLIVIA SIMPSON, UC SAN DIEGO
C. SESHADHRI, UC SANTA CRUZ
ANDREW MCGREGOR, UMASS
as-Skitter (SNAP)
• Internet topology graph
  • Nodes are IP addresses
  • Edges are links
• Large: 1.7M nodes, 11M edges
Graphs in a streaming model

- Graph is an accumulation of a stream of edges
- No access to edges earlier in the stream
- Rather than computing over the fully accumulated graph, estimate properties using small space (a fraction of the number of vertices)
  - Limited memory, $M$
  - Update $M$ upon seeing an edge
Our problem: estimating the degree distribution

The complementary cumulative degree histogram (ccdh) is the sequence over all degrees $d$:

$$N(d) = \text{the number of nodes of degree } \geq d$$

Given a graph as a stream of edges, compute an approximation of the ccdh.
Estimating at all scales

Many networks are scale-free and heavy-tailed

- e.g. average degree is small (20) but maximum degree can be very large (~50K)
- There are nodes with degrees at all scales

Treat head and tail separately

Algorithms for counting frequent items:

- frequent [Demaine et al. ’02, Karp et al. ‘03, Berinde et al. ‘10]
- lossy counting [Manku and Motwani ‘02]
- space saving [Metwally et al. ‘05]
Measuring the quality of the estimator

\[ K_n \text{ VS. } K_{n+1} \]

- KS or \( L_1 \) distances are big
- Distributions are quite similar

\[ n\text{-STAR VS. MATCHING} \]

- KS or \( L_1 \) distances are small (distributions differ only at one point)
- Distributions are fundamentally different
Relative Hausdorff (RH) distance

\[
\inf\{\varepsilon | \forall d, \exists d' \in [(1-\varepsilon)d, (1+\varepsilon)d], \text{ such that } |F(d) - G(d')| \leq \varepsilon F(d)\}.
\]

Every point in one ccdh is close to some point in the other ccdh.
The **headtail** algorithm

| node140 | node2003 | node31 | node4990 | ...
|---------|----------|--------|-----------|---------|
| 1       | 2        | 14     | 1         | ...

**Sampling phase:**

- Sample a node with fixed probability and count its degree:
  - More likely to sample common degrees
- Sample an edge with fixed probability and count edges adjacent to endpoints after time of sampling:
  - More likely to sample high degrees

**Processing phase:**
Simple for the head, some tricks involved for the tail

---

**Single pass small-space!**
Results on a few SNAP graphs

as-Skitter
1.7M nodes, 11M edges
Storage: 31K nodes

com-LiveJournal
4M nodes, 34M edges
Storage: 200K nodes

com-Orkut
3M nodes, 117M edges
Storage: 150K nodes
## Results on a few SNAP graphs

<table>
<thead>
<tr>
<th>Graph</th>
<th>n</th>
<th>m</th>
<th>Space</th>
<th>RH distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>youtube</td>
<td>1.1M</td>
<td>3M</td>
<td>21K</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>90K</td>
<td>0.076</td>
</tr>
<tr>
<td>wiki-Talk</td>
<td>2.3M</td>
<td>5M</td>
<td>38K</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>74K</td>
<td>0.055</td>
</tr>
<tr>
<td>youtube-friendship</td>
<td>3M</td>
<td>9M</td>
<td>80K</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>196K</td>
<td>0.05</td>
</tr>
<tr>
<td>as-Skitter</td>
<td>1.7M</td>
<td>11M</td>
<td>31K</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>69K</td>
<td>0.073</td>
</tr>
<tr>
<td>soc-Pokec</td>
<td>1.6M</td>
<td>30M</td>
<td>75K</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>212K</td>
<td>0.14</td>
</tr>
<tr>
<td>com-LiveJournal</td>
<td>4M</td>
<td>34M</td>
<td>335K</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>467K</td>
<td>0.058</td>
</tr>
<tr>
<td>com-Orkut</td>
<td>3M</td>
<td>117M</td>
<td>273K</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>387K</td>
<td>0.13</td>
</tr>
</tbody>
</table>
Results: convergence

as-Skitter
1.7M nodes, 11M edges

1% of the edge stream
Results: compare to other methods

as-Skitter
1.7M nodes, 11M edges
Results: combining with existing methods

as-Skitter
1.7M nodes, 11M edges, Storage: 50K nodes (versus headtail, 31K)
Catching the head, tail, and everything in between: a streaming algorithm for the degree distribution

OLIVIA SIMPSON, UC SAN DIEGO
osimpson@ucsd.edu
Full paper linked from my webpage:
cseweb.ucsd.edu/~osimpson