Large Graph Mining: Power Tools and a Practitioner’s guide

Christos Faloutsos
Gary Miller
Charalampos (Babis) Tsourakakis
CMU
Outline

• Introduction – Motivation
• Task 1: Node importance
• Task 2: Community detection
• Task 3: Recommendations
• Task 4: Connection sub-graphs
• Task 5: Mining graphs over time
• Task 6: Virus/influence propagation
• Task 7: Spectral graph theory
• Task 8: Tera/peta graph mining: hadoop

Observations – patterns of real graphs
• Conclusions
Observations – ‘laws’ of real graphs

- Observation #1: small and SHRINKING diameter
- Observation #2: power law / skewed degree distributions
- Observation #3: power laws in several aspects
- Observation #4: communities
Observation 1 – diameter

• Small diameter – ‘six degrees’
• … and the diameter SHRINKS as the graph grows (!)
Diameter – “Patents”

- Patent citation network
- 25 years of data
Observation 1 – diameter

• Small diameter – ‘six degrees’
• … and the diameter SHRINKS as the graph grows (!)

Practical implication: BFS may die:
   – 3-step-away neighbors => half of the graph!
Observations  2 – degree distribution

Skewed degree distribution

• Most nodes have degree 1 or 2
• … but they probably have a neighbor with degree 100,000 or so (!)
Degree distributions

- web hit counts [w/ A. Montgomery]

Web Site Traffic

\[ \log(\text{count}) \]

\[ \log(\text{in-degree}) \]

Zipf

``ebay''

users

sites

KDD’09 Faloutsos, Miller, Tsourakakis

P9-8
epinions.com

- who-trusts-whom
  [Richardson + Domingos, KDD 2001]

(out) degree

<table>
<thead>
<tr>
<th>Count</th>
<th>Out-degree</th>
</tr>
</thead>
<tbody>
<tr>
<td>10000</td>
<td>10000</td>
</tr>
<tr>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

trusts-2000-people user
Observation 2 – degree distributions

Skewed degree distribution

• Most nodes have degree 1 or 2
• … but they probably have a neighbor with degree 100,000 or so (!)

Practical implications:

• May need to delete/ignore those high degree nodes
• Could probably also trim the 1-degree nodes, saving significant space and time
Observation 3 – power laws

Power-laws / skewed distributions in everything:

– Most pairs: within 2-3 steps; but, some pair: ~20 or more steps away
– Triangles: power laws [Tsourakakis’08]
– # of cliques: ditto [Du+’09]
– Weight vs degree: ditto [McGlohon+’08]
Observation 4 – communities

• ‘Negative dimensionality’ paradox [Chakrabarti+’04]

Practical implication:
• Graphs may have no good cuts
Conclusions

0) Graphs appear in numerous settings
1) Singular / eigenvalue analysis: valuable
   – Fixed points – random walks – importance
   – Eigenvalue and epidemic threshold
   – Laplacians -> communities
Conclusions – cont’d

2) Random walks -> proximity
   – Recommendations, auto-captioning, etc
   – Fast algo’s, through Sherman-Morrison

3) Tera-byte scale graphs: hadoop

4) Beware: counter-intuitive properties
   – small diameters; power-laws; possible lack of good cuts
Acknowledgements

Funding:

IIS-0705359, IIS-0534205, DBI-0640543, CNS-0721736

PITA (PA Inf. Tech. Alliance)
Acknowledgements - foils

- Chakrabarti, Deepay (cross-associations)
- Kolda, Tamara (tensors)
- Papadimitriou, Spiros (cross-associations)
- Sun, Jimeng (tensors)
- Tong, Hanghang (proximity)
THANK YOU!

Christos Faloutsos
www.cs.cmu.edu/~christos

Gary Miller
www.cs.cmu.edu/~glmiller

Charalampos (Babis) Tsourakakis
www.cs.cmu.edu/~ctsourak

www.cs.cmu.edu/~christos/TALKS/09-KDD-tutorial/