Diminishing Communities in Large Social and Information Networks

Jure Leskovec, CMU
Kevin Lang, Anirban Dasgupta and Michael Mahoney
Yahoo! Research
Network communities

- **Communities:**
  - Sets of nodes with **lots** of connections **inside** and **few** to **outside** (the rest of the network)

- **Assumption:**
  - Networks are (hierarchically) composed of **communities** (modules)

**Question:** Are large networks really like this?

Communities, clusters, groups, modules

Jure Leskovec, NetSci '08
Community score (quality)

- How community like is a set of nodes?
- Need a natural intuitive measure

**Conductance** (normalized cut)

\[ \Phi(S) = \frac{\# \text{ edges cut}}{\# \text{ edges inside}} \]

- Small \( \Phi(S) \) corresponds to more community-like sets of nodes

Jure Leskovec, NetSci '08
Community score (quality)

What is “best” community of 5 nodes?

Score: $\Phi(S) = \# \text{ edges cut} / \# \text{ edges inside}$

Jure Leskovec, NetSci ‘08
What is “best” community of 5 nodes?

Score: \( \Phi(S) = \frac{\# \text{ edges cut}}{\# \text{ edges inside}} \)

Bad community

\[ \Phi = \frac{5}{6} = 0.83 \]
Community score (quality)

What is “best” community of 5 nodes?

Better community
\[ \Phi = 2/5 = 0.4 \]

Bad community
\[ \Phi = 5/7 = 0.7 \]

Score: \( \Phi(S) = \text{# edges cut} / \text{# edges inside} \)

Jure Leskovec, NetSci '08
What is “best” community of 5 nodes?

Score: $\Phi(S) = \frac{\text{# edges cut}}{\text{# edges inside}}$

Better community
$\Phi = \frac{2}{5} = 0.4$

Bad community
$\Phi = \frac{5}{7} = 0.7$

Best community
$\Phi = \frac{2}{8} = 0.25$

Jure Leskovec, NetSci ’08
We define: Network community profile (NCP) plot

Plot the score of best community of size $k$

$$\Phi(k) = \min_{S \subseteq V, |S| = k} \phi(S)$$

$\Phi(5) = 0.25$

$\Phi(7) = 0.18$
Dolphin social network [Lusseau et al. 2005]
- Two communities of dolphins

Network

NCP plot

Jure Leskovec, NetSci '08
NCP plot: Zachary’s karate club

- Zachary’s university karate club social network
  - During the study club split into 2
  - The split (squares vs. circles) corresponds to cut B

Jure Leskovec, NetSci ’08
Collaborations between scientists in Networks
[Newman, 2005]
Geometric and Hierarchical graphs

Geometric (grid-like) network

Hierarchical network

[Ravasz&Barabasi, 2003]

Jure Leskovec, NetSci '08
Geometric and Hierarchical graphs

- Small social networks
- Geometric and
- Hierarchical network have downward NCP plot
Previously researchers examined community structure of small networks (~100 nodes)

We examined more than 70 different large networks

Some of the networks:

<table>
<thead>
<tr>
<th>Type</th>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social nets</td>
<td>LIVEJOURNAL</td>
<td>4,843,953</td>
<td>42,845,684</td>
<td>Blog friendships [5]</td>
</tr>
<tr>
<td></td>
<td>EPINIONS</td>
<td>75,877</td>
<td>405,739</td>
<td>Trust network [28]</td>
</tr>
<tr>
<td></td>
<td>CA-DBLP</td>
<td>317,080</td>
<td>1,049,866</td>
<td>Co-authorship [5]</td>
</tr>
<tr>
<td>Information (citation) nets</td>
<td>CIT-hep-th</td>
<td>27,400</td>
<td>352,021</td>
<td>Arxiv hep-th [14]</td>
</tr>
<tr>
<td></td>
<td>AMAZONPROD</td>
<td>524,371</td>
<td>1,491,793</td>
<td>Amazon products [8]</td>
</tr>
<tr>
<td>Web graphs</td>
<td>WEB-GOOGLE</td>
<td>855,802</td>
<td>4,291,352</td>
<td>Google web graph</td>
</tr>
<tr>
<td></td>
<td>WEB-WT10G</td>
<td>1,458,316</td>
<td>6,225,033</td>
<td>TREC WT10G</td>
</tr>
<tr>
<td>Bipartite affiliation</td>
<td>ATP-DBLP</td>
<td>615,678</td>
<td>944,456</td>
<td>DBLP [21]</td>
</tr>
<tr>
<td>(authors-to-papers) networks</td>
<td>AT-M-IMDB</td>
<td>2,076,978</td>
<td>5,847,693</td>
<td>Actors-to-movies</td>
</tr>
<tr>
<td>Internet networks</td>
<td>ASSKITTER</td>
<td>1,719,037</td>
<td>12,814,089</td>
<td>Autonom. sys.</td>
</tr>
<tr>
<td></td>
<td>GNUTELLA</td>
<td>62,561</td>
<td>147,878</td>
<td>P2P network [29]</td>
</tr>
</tbody>
</table>
Previously researchers examined community structure of small networks (~100 nodes)

We examined more than 70 different large networks

Large real-world networks look very different!
Example of our findings

- **Typical example:**
  General relativity collaboration network
  (4,158 nodes, 13,422 edges)

Jure Leskovec, NetSci '08
NCP: LiveJournal (N=5M, E=42M)

Better and better communities

Best communities get worse and worse

Best community has 100 nodes

Jure Leskovec, NetSci '08
- **Whiskers** are responsible for downward slope of NCP plot

**Explanations:**

- **Whisker** is a set of nodes connected to the network by a **single** edge.

**NCP plot**

*Largest whisker*

-Jure Leskovec, NetSci '08*
Whisker shapes

Whiskers in real networks are non-trivial (richer than trees)

Jure Leskovec, NetSci '08
Whiskers in real networks are larger than expected.
Each new edge inside the community costs more

\[ \Phi = \frac{1}{3} = 0.33 \]

\[ \Phi = \frac{2}{4} = 0.5 \]

\[ \Phi = \frac{8}{6} = 1.3 \]

\[ \Phi = \frac{64}{14} = 4.5 \]

Each node has twice as many children.

Jure Leskovec, NetSci'08
Network structure: Core-periphery (jellyfish, octopus)

Core contains 60% node and 80% edges

Whiskers are responsible for good communities

Denser and denser core of the network

Whiskers
Caveat: Bag of whiskers

What if we allow cuts that give disconnected communities?

- Cut all whiskers
- Compose communities out of whiskers
- How good “community” do we get?
Communities made of whiskers

We get better community scores when composing disconnected sets of whiskers.
What if we remove whiskers?

Nothing happens! Now we have 2-edge connected whiskers to deal with.

Jure Leskovec, NetSci '08
Comparison to a rewired network

Rewired network: random network with same degree distribution
What is a good model that explains such network structure?

None of the existing models work
Forest Fire model works

- **Forest Fire** [Leskovec, Kleinberg, Faloutsos, KDD ‘05]: connections spread like a fire
  - New node joins the network
  - Selects a seed node
  - Connects to some of its neighbors
  - Continue recursively

As community grows it blends into the core of the network
Forest Fire NCP plot

 rewired
network
Bag of whiskers
Conclusion and connections

- **Whiskers:**
  - Largest whisker has \(~100\) nodes
  - Whisker size is independent of network size
  - Dunbar number: a person can maintain social relationship to at most 150 people

- **Core:**
  - 60% of the nodes, 80% edges
  - Core has little structure (hard to cut)
  - Still more structure than the random network
NCP plot is a way to analyze network community structure.

Our results agree with previous work on small networks (people did not hit the Dunbar’s limit).

But large networks are different:
- Whiskers + Core (core-periphery) structure
- Small well isolated communities blend into the core of the networks as they grow.