Community Detection: Overlapping Communities

CS224W: Social and Information Network Analysis
Jure Leskovec, Stanford University
http://cs224w.stanford.edu
Non-overlapping Communities

Network Adjacency matrix

Network

Adjacency matrix

Nodes

Nodes
Overlapping Communities

- Non-overlapping vs. overlapping communities
Overlaps of Social Circles

- A node belongs to many social “circles’
What if communities overlap?

High school

Company

Stanford (Squash)

Stanford (Basketball)
Two nodes belong to the same community if they can be connected through adjacent $k$-cliques:

- **$k$-clique:**
  - Fully connected graph on $k$ nodes

- **Adjacent $k$-cliques:**
  - overlap in $k-1$ nodes

**$k$-clique community**

- Set of nodes that can be reached through a sequence of adjacent $k$-cliques
Two nodes belong to the same community if they can be connected through adjacent $k$-cliques:
Clique Percolation Method:

- Find maximal-cliques (not $k$-cliques!)

Clique overlap graph:

- Each clique is a node
- Connect two cliques if they overlap in at least $k-1$ nodes

Communities:

- Connected components of the clique overlap matrix

How to set $k$?

- Set $k$ so that we get the “richest” (most widely distributed cluster sizes) community structure
CPM method: Example

- Start with graph
- Find maximal cliques
- Create clique overlap matrix
- Threshold the matrix at value $k-1$
  - If $a_{ij} < k - 1$ set 0
- Communities are the connected components of the thresholded matrix
Communities in a “tiny” part of a phone call network of 4 million users
[Palla et al., ‘07]
No nice way, hard combinatorial problem

Maximal clique: Clique that can’t be extended
- \{a, b, c\} is a clique but not maximal clique
- \{a, b, c, d\} is maximal clique

Algorithm: Sketch
- Start with a seed node
- Expand the clique around the seed
- Once the clique cannot be further expanded we found the maximal clique

Note:
- This will generate the same clique multiple times
How to Find Maximal Cliques?

- Start with a seed vertex $a$
- **Goal**: Find the max clique $Q$ that $a$ belongs to
  - **Observation**:
    - If some $x$ belongs to $Q$ then it is a neighbor of $a$
    - Why? If $a, x \in Q$ but edge $(a, x)$ does not exist, $Q$ is not a clique!
- **Recursive algorithm**:
  - $Q$ ... current clique
  - $R$ ... candidate vertices to expand the clique to
- **Example**: Start with $a$ and expand around it

Steps of the recursive algorithm

$\Gamma(u)$...neighbor set of $u$
How to Find Maximal Cliques?

- Start with a seed vertex \( a \)
- **Goal:** Find the max clique \( Q \) that \( a \) belongs to
  - **Observation:**
    - If some \( x \) belongs to \( Q \) then it is a neighbor of \( a \)
    - **Why?** If \( a, x \in Q \) but edge \((a, x)\) does not exist, \( Q \) is not a clique!
- **Recursive algorithm:**
  - \( Q \) ... current clique
  - \( R \) ... candidate vertices to expand the clique to
- **Example:** Start with \( a \) and expand around it

\[
\begin{align*}
Q &= \{a\} \quad \{a,b\} \quad \{a,b,c\} \quad \text{bktrack} \quad \{a,b,d\} \\
R &= \{b,c,d\} \quad \{b,c,d\} \quad \{d\} \cap \Gamma(c) = \{} \quad \{c\} \cap \Gamma(d) = \{} \\
\cap \Gamma(b) &= \{c,d\}
\end{align*}
\]

Steps of the recursive algorithm

\( \Gamma(u) \)...neighbor set of \( u \)
How to Find Maximal Cliques?

- \( Q \) ... current clique
- \( R \) ... candidate vertices

**Expand** \((R, Q)\)

- **while** \( R \neq \{\} \)
  - \( p = \text{vertex in } R \)
  - \( Q_p = Q \cup \{p\} \)
  - \( R_p = R \cap \Gamma(p) \)
  - **if** \( R_p \neq \{\} \): Expand \((R_p, Q_p)\)
  - **else**: output \( Q_p \)
  - \( R = R - \{p\} \)

---

Start: Expand\((V, \{\})\)

\( R = \{a...f\}, Q = \{\} \)

- \( p = \{a\} \)
- \( Q_p = \{a\} \)
- \( R_p = \{b,d\} \)

Expand\((R_p, Q)\):

- \( R = \{b,d\}, Q = \{a\} \)
- \( p = \{b\} \)
- \( Q_p = \{a,b\} \)
- \( R_p = \{d\} \)

Expand\((R_p, Q)\):

- \( R = \{d\}, Q = \{a,b\} \)
- \( p = \{d\} \)
- \( Q_p = \{a,b,d\} \)
- \( R_p = \{\} \): output \( \{a,b,d\} \)
- \( p = \{d\} \)
- \( Q_p = \{a,d\} \)
- \( R_p = \{b\} \)

Expand\((R_p, Q)\):

- \( R = \{b\}, Q = \{a,d\} \)
- \( p = \{b\} \)
- \( Q_p = \{a,d\} \)
- \( R_p = \{\} \): output \( \{a,d,b\} \)
How to Model Networks with Communities?

Announcement: On Thu we will have Lars Backstrom from Facebook Data Science team give a guest lecture!
How should we think about large scale organization of clusters in networks?

**Finding:** There is a lack of (traditional) clustering.
How do we reconcile these two views? (and still do community detection)
Community Score

- How community like is a set of nodes?
- A good cluster $S$ has
  - Many edges internally
  - Few edges pointing outside
- Simplest objective function: Conductance

$$\phi(S) = \frac{\left| \{(i, j) \in E; i \in S, j \notin S\} \right|}{\sum_{s \in S} d_s}$$

Small conductance corresponds to good clusters
(Note $|S| < |V|/2$)
Define: Network community profile (NCP) plot

Plot the score of best community of size $k$

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

(Note $|S| < |V|/2$)
How to (Really) Compute NCP?

- Run the favorite clustering method
- Each dot represents a cluster
- For each size find “best” cluster

Cluster size, log $k$

Cluster score, log $\Phi(k)$
NCP Plot: Meshes

Meshes, grids, dense random graphs:

- d-dimensional meshes
- California road network
Collaborations between scientists in networks

[Newman, 2005]
Natural hypothesis about NCP:

- NCP of real networks slopes downward
- Slope of the NCP corresponds to the “dimensionality” of the network

What about large networks?
Typical example: General Relativity collaborations (n=4,158, m=13,422)
More NCP Plots of Networks

(a) LIVEJOURNAL01

(b) MESSENGER-DE

(c) ATP-DBLP

(d) CIT-HEP-TH

(e) WEB-GOOGLE

(f) AMAZONALL
NCP: LiveJournal \((n=5m, m=42m)\)

Better and better clusters

Clusters get worse and worse

Best cluster has ~100 nodes
As clusters grow the number of edges inside grows **slower** than the number crossing.

Each node has twice as many children.
Empirically we note that **best clusters are barely connected** to the network.

⇒ Core-periphery structure
What If We Remove Good Clusters?

Nothing happens!
⇒ Nestedness of the core-periphery structure
Denser and denser core of the network

Core contains 60% node and 80% edges

Whiskers are responsible for good communities

Nested Core-Periphery (jellyfish, octopus)
How do we reconcile these two views?
Many methods for overlapping communities

- Clique percolation [Palla et al. ‘05]
- Link clustering [Ahn et al. ‘10] [Evans et al. ‘09]
- Clique expansion [Lee et al. ‘10]
- Mixed membership stochastic block models [Airoldi et al. ‘08]
- Bayesian matrix factorization [Psorakis et al. ‘11]

What do these methods assume about community overlaps?
Many overlapping community detection methods make an implicit assumption:

- Edge probability decreases with the number of shared communities

Is this true?
Ground-truth Communities

- Basic question: nodes $u$, $v$ share $k$ communities
- What’s the edge probability?

![Graph for LiveJournal social network](image1)

- $P(k)$, Edge probability
- $k$, Number of shared communities
- LiveJournal social network

![Graph for Amazon product network](image2)

- $P(k)$, Edge probability
- $k$, Number of shared communities
- Amazon product network
Edge density in the overlaps is higher!

“The more different foci (communities) that two individuals share, the more likely it is that they will be tied”

- S. Feld, 1981

Communities as “tiles”
Communities as overlapping tiles

Web of affiliations [Simmel ‘64]
What does this mean?

Ganovetter and all non-overlapping methods

Palla et al., MMSB and other overlapping methods as well
Many methods fail to detect dense overlaps:

- Clique percolation, ...
Generative model: How is a network generated from community affiliations?

Model parameters:
- Nodes $V$, Communities $C$, Memberships $M$
- Each community $c$ has a single probability $p_c$
Given parameters \((V, C, M, \{p_c\})\)

- Nodes in community \(c\) connect to each other by flipping a coin with probability \(p_c\).
- Nodes that belong to multiple communities have multiple coin flips: Dense community overlaps

\[
p(u, v) = 1 - \prod_{c \in M_u \cap M_v} (1 - p_c)
\]
AGM: Dense Overlaps

Model

Network
AGM is flexible and can express variety of network structures: Non-overlapping, Nested, Overlapping
Connections: Core-Periphery
Test of the Conjecture

\( \langle m \rangle \), Community memberships vs. \( d \), Farness Centrality

LiveJournal social network
Primary & Secondary Cores

- **Primary core**: Foodwebs, Web-graph, Social
- **Secondary cores**: PPI, Products