Announcements:
Project milestones graded
Keep up the good work!

Community Detection: Overlapping Communities
Non-overlapping Communities

Network

Adjacency matrix

Nodes
Overlapping Communities

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

- A node can belong to many social “circles”
What if communities overlap?

High school

Company

Stanford (Basketball)

Stanford (Squash)
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:

- Adjacent 4-cliques
- Non-adjacent 4-cliques
- Communities for $k=4$
Clique Percolation Method:

- Find maximal-cliques
  - Def: Clique is maximal if no superset is a clique

- Clique overlap super-graph:
  - Each clique is a super-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
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

(1) Graph

(2) Clique overlap matrix

(3) Thresholded matrix at 3

(4) Communities (connected components)
Example: Phone-Call Network

Communities in a “tiny” part of a phone call network of 4 million users
[Palla et al., ’07]
How to Find Maximal Cliques?

- **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

$Q = \emptyset$

$R = \emptyset$

$\Gamma(u) \ldots$ 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

$$Q = \{a\} \quad \{a, b\} \quad \{a, b, c\} \quad \text{bktrack} \quad \{a, b, d\}$$

$$R = \{b, c, d\} \quad \{b, c, d\} \quad \{d\} \cap \Gamma(c) = \emptyset \quad \{c\} \cap \Gamma(d) = \emptyset$$

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 $Q_p$
  - $R = R - \{p\}$
How to Model Networks with Communities?
How should we think about large scale organization of clusters in networks?

Finding: Community Structure
How should we think about large scale organization of clusters in networks?

**Finding:** Core-periphery structure
How do we reconcile these two views?
(and still do community detection)

Community structure vs. Core-periphery
How community-like is a set of nodes?

A good cluster $S$ has
- Many edges internally
- Few edges pointing outside

What’s a good metric: Conductance

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

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
NCP Plot: Meshes

- Meshes, grids, dense random graphs:

  ![Graphs showing conductance vs. number of nodes in the cluster]

  - Random graph, $1/d \approx 0$
  - Cube, $1/d \approx 0.33$
  - Grid, $1/d \approx 0.50$
  - Chain, $1/d \approx 1.0$

  - d-dimensional meshes
  - California road network

11/18/2014
Collaborations between scientists in networks

[Newman, 2005]

Dips in the conductance graph correspond to the "good" clusters we can visually detect
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) AT&T-DBLP

(d) CIT-HEP-TH

(e) WEB-GOOGLE

(f) AMAZON ALL

-- Rewired graph
-- Real graph
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.

\[ \Phi = \frac{1}{7} = 0.14 \]

\[ \Phi = \frac{2}{10} = 0.2 \]

\[ \Phi = \frac{8}{20} = 0.4 \]

\[ \Phi = \frac{64}{92} = 0.69 \]
Empirically we note that **best clusters** (corresponding to **green nodes**) are **barely connected** to the network

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

Nothing happens!⇒ Nestedness of the core-periphery structure
Suggested Network Structure

Denser and denser core of the network

Core contains 60% node and 80% edges

Nested Core-Periphery (jellyfish, octopus)

Whiskers are responsible for good communities
Part 2: Explanation

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)

![Graph for Amazon product network](image2)
Communities as Tiles!

- 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?

Granovetter 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 $p_c$
AGM: Generative Process

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
  - If they "miss" the first time, they get another chance through the next community

\[
p(u, v) = 1 - \prod_{c \in M_u \cap M_v} (1 - p_c)
\]
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, \text{Community memberships}

\( d, \text{Farness Centrality} \)

LiveJournal social network
Primary & Secondary Cores

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