Social Network Analysis

Lecture 5: Community Detection

Dr. Hung-Nghiep Tran
nghiepth@uit.edu.vn

University of Information Technology, VNU-HCM, Vietnam
2025

Mục tiêu buổi học

- Explain what “communities” mean in network analysis.
- Understand basic cut-based and modularity-based formulations.
- Explain and compare spectral clustering and Louvain methods.
- Interpret eigenvalues/eigenvectors of Laplacian in the context of clustering.
- Recognize higher-order approaches (motif clustering) for complex structures.

Câu hỏi kiểm tra trước

Quiz

1. What does “community” mean in a network context?
A. Group of nodes with equal degree B. Group of nodes that are densely connected internally
C. Group of nodes with no edges between them D. Randomly chosen set of nodes

2. Which measure is often optimized in community detection?
A. Degree B. Modularity C. Clustering coefficient D. Betweenness

3. What is the main idea behind spectral clustering?
A. Use eigenvectors of Laplacian matrix to find partitions B. Count triangles in the graph
C. Rank nodes by centrality D. Randomly cut the network

4. Which of the following methods is hierarchical and greedy?
A. PageRank B. Louvain C. Eigenvector centrality D. BFS clustering

5. The Laplacian matrix is defined as:
A. $L = A - D$ B. $L = D - A$ C. $L = A + D$ D. $L = A^T A$

Quiz

1. What does “community” mean in a network context?
A. Group of nodes with equal degree **B. Group of nodes that are densely connected internally**
C. Group of nodes with no edges between them D. Randomly chosen set of nodes

2. Which measure is often optimized in community detection?
A. Degree **B. Modularity** C. Clustering coefficient D. Betweenness

3. What is the main idea behind spectral clustering?
A. Use eigenvectors of Laplacian matrix to find partitions B. Count triangles in the graph
C. Rank nodes by centrality D. Randomly cut the network

4. Which of the following methods is hierarchical and greedy?
A. PageRank **B. Louvain** C. Eigenvector centrality D. BFS clustering

5. The Laplacian matrix is defined as:
A. $L = A - D$ **B. $L = D - A$** C. $L = A + D$ D. $L = A^T A$

Đọc và diễn giải

Tài liệu đọc

- Reading:
  - (Stanford CS224W L6) Community Structure
  - (Stanford CS224W L7) Spectral Clustering

Câu hỏi kiểm tra sau

Quiz

1. Why are modularity-based methods popular?
A. They are intuitive and scale to large networks B. They are guaranteed to find global optimum
C. They only work on trees D. They depend on node labels

2. What is the Fiedler vector?
A. The eigenvector with the largest eigenvalue B. The normalized degree vector
C. The eigenvector corresponding to the second smallest eigenvalue D. A random initialization vector

3. What limitation does modularity optimization face?
A. It cannot detect communities in disconnected graphs B. It suffers from resolution limit (small communities missed)
C. It only works for directed graphs D. It requires prior knowledge of k

4. How does spectral clustering find partitions?
A. Minimize cut directly B. Use degree distribution as threshold
C. Randomly split eigenvalues D. Relax the discrete optimization to continuous, then threshold eigenvector values

5. In motif-based clustering (higher-order spectral), the adjacency is redefined by:
A. Counting the number of edges between nodes B. Counting how often pairs of nodes co-occur in motifs
C. Counting shortest paths D. Using PageRank similarities

Quiz

1. Why are modularity-based methods popular?
A. They are intuitive and scale to large networks B. They are guaranteed to find global optimum
C. They only work on trees D. They depend on node labels

2. What is the Fiedler vector?
A. The eigenvector with the largest eigenvalue B. The normalized degree vector
C. The eigenvector corresponding to the second smallest eigenvalue D. A random initialization vector

3. What limitation does modularity optimization face?
A. It cannot detect communities in disconnected graphs **B. It suffers from resolution limit (small communities missed)**
C. It only works for directed graphs D. It requires prior knowledge of k

4. How does spectral clustering find partitions?
A. Minimize cut directly B. Use degree distribution as threshold
C. Randomly split eigenvalues **D. Relax the discrete optimization to continuous, then threshold eigenvector values**

5. In motif-based clustering (higher-order spectral), the adjacency is redefined by:
A. Counting the number of edges between nodes **B. Counting how often pairs of nodes co-occur in motifs**
C. Counting shortest paths D. Using PageRank similarities

Thảo luận

Chuẩn bị tuần trước

- Muddiest point: “Điểm nào mù mờ nhất sau khi đọc? (< 50 từ)”

Chuẩn bị tuần trước

Trả lời ngắn câu hỏi định hướng

1. What does “community” mean in a network context?
   - A community is a set of nodes that are densely connected internally but sparsely connected externally. Formally, good communities minimize the number of cut edges while maximizing internal connectivity or modularity.
2. Why are modularity-based methods popular for detecting communities?
   - They provide an intuitive, scalable, and unsupervised way to find partitions by maximizing the difference between observed and expected edge densities within groups. Louvain and Leiden algorithms can handle very large networks efficiently.
3. Compare Louvain and spectral clustering: main intuition behind each.
   - Louvain: Greedy optimization of modularity; hierarchical merging of small communities into larger ones. Fast and scalable.
   - Spectral clustering: Uses eigenvectors of Laplacian to find continuous relaxations of graph cuts, then discretizes to form clusters. Captures global structure but computationally heavier.

Chuẩn bị tuần trước

Trả lời ngắn câu hỏi định hướng

4. What kind of “ground truth” communities might exist in real datasets?
   - Social networks: friend groups, interest communities.
   - Biological networks: protein complexes, functional pathways.
   - Citation networks: research fields or topical clusters.
   - Infrastructure networks: regional clusters or subnetworks.
5. What challenges arise when evaluating community detection algorithms?
   - Lack of reliable ground truth in many datasets.
   - Resolution limit: small but meaningful communities missed.
   - Overlapping communities not captured by simple partitions.
   - Scalability and sensitivity to parameters.
   - Subjectivity of what counts as a “good” community (semantic vs structural).

Hands-on lab

Setup

- Install python
  - with Anaconda
- Install IDE:
  - (Microsoft) vscode
  - Jupyter Notebook
- Install networkx
  - with pip
- Install gephi

Chuẩn bị cho tuần tới

Chuẩn bị trước tuần sau

- Reading:
  - (Stanford CS224W L15) Network Centrality
  - (Stanford CS224W L3) PageRank

- Câu hỏi định hướng (trả lời ngắn):
  - What does “link prediction” mean in networks?
  - Why is link prediction important for understanding evolving graphs?
  - Describe two simple heuristic link predictors (e.g., Common Neighbors, Jaccard).
  - How does the preferential attachment model relate to link prediction?
  - What are some challenges when evaluating link prediction performance?

- Muddiest point: “Điểm nào mù mờ nhất sau khi đọc? (< 50 từ)”

Thank you for listening

Q & A