Data Integration Course Outline

Aaron Brooks

Section: Cluster detection, analysis and visualization

130 min lecture / 110 min lab (4 hrs total)

1. Intro to clustering (20 min lecture, 15 min lab)

Lecture

- What is clustering/community detection? Why cluster? Relationship to graph/network structure? Role in data integration.
- Simple biological and non-biological examples (motivation, e.g. karate club, regulatory "modules")
- Approaches to clustering: Node vs edge clustering, Hard vs soft clustering
- Overview of topics covered / Lab data set

Lab

- Brief exercises to begin thinking algorithmically about the process of clustering
- GGobi?
- Generate heatmaps for individual and combined data sets from Jean Karim

2. Cluster detection (60 min lecture, 45 min lab)

Lecture

- Clustering in a nutshell: some way to find elements on a graph that are more related to each other than to everything else
- Distance/Similarity metrics (10 min)
- Node-based clustering (35 min)
    - Hierarchical clustering

    - K-means
    - Spectral clustering / clustering with dimensionality reduction
- Link-based clustering (15 min)
    - Link-community detection

Lab

- Perform spectral clustering and link-community detection on integrated kernel network from Jean-Karim.
- Understand how use and interpret the results

3. Cluster analysis (30 min lecture, 30 min lab)

Lecture

- How to evaluate the quality of clustering
- Cluster quality metrics
- Cluster (biological) interpretation

Lab

4. Cluster visualization and extensions (20 min lecture, 20 min lab)

Lecture

- Basic visualizations: Heatmap and cluster dendrogram
- More advanced visualization: Network-based visualization (link-community detection), Gephi for dynamic graphs?
- Biological recap. Why are we doing this?
- Preview of advanced methods for data clustering. Hypergraphs, "Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes", etc.

Lab

- Gephi for highlighting communities and visualizing their dynamics