Outline of the Seminar Paper

Community Detection in Bacterial and Viral Networks

Jonas Ziegler

February 19, 2026

Proposed Structure of the Paper (15–20 pages)

1. **Introduction (2–3 pages)**
   (a) Motivation
      - Why statistical network analysis is important
      - Biological networks (protein interactions, gene regulation)
      - Why community detection?

   (b) Definition of a Network
      - Graph as a mathematical object
      - Nodes and edges
      - Adjacency list
      - Adjacency matrix
      - Directed vs. undirected networks
      - Weighted vs. unweighted networks
      - Weighted undirected network: $G = (V, E, W)$

   (c) What can we do with Networks?
      - Descriptive statistics
      - Structural analysis
      - Clustering and communities
      - Prediction of Future Community Behavior

2. **Network Summary Statistics (2–3 pages)**
   (a) Basic Metrics
      - Number of nodes
      - Number of edges
      - Density
      - Average degree

   (b) Degree Distribution
      - Scale-free networks
      - Power-law behavior (if relevant)

(c) Centrality Measures

- Degree centrality
- Betweenness centrality
  - Node betweenness
  - Edge betweenness
- Closeness and betweenness (brief)

3. **Similarity Measures for Community Comparison (2 pages)**

(a) Adjusted Rand Index (ARI)

(b) Adjusted Mutual Information (AMI)

(c) Comparison of ARI and AMI

- When to use which measure

4. **Community Detection Methods (4–5 pages)**

(a) Modularity

- Basic idea
- Modularity formula:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(c_i, c_j)$$

(b) Edge Betweenness (Girvan–Newman)

- Main focus of the paper
- Idea: edges with high betweenness separate communities
- Algorithm:
  i. Compute edge betweenness
  ii. Remove the strongest edge
  iii. Repeat
- Advantages and disadvantages

(c) Louvain Method

- Greedy optimization of modularity
- Very fast
- Hierarchical structure
- Limitations of Louvain
- Motivation for improvement -> Leiden

(d) Leiden Method

- Improvement over Louvain
- Guarantees well-connected communities
- State-of-the-art method

5. Data Description (2 pages)
   (a) Data Source
      - Bacterial protein similarity networks
      - Viral protein similarity networks
   (b) Descriptive Comparison of Bac and Vir
      - Size
      - Density
      - Degree distributions
   (c) Network Construction
      - Nodes represent proteins
      - Edges represent sequence similarity
      - Weights: bit score
      - bit score formula:
$$S' = \frac{\lambda S - \ln K}{\ln 2}$$
   (d) Preprocessing
      - Filtering
      - Removal of self-loops
      - Largest connected component
6. Experimental Setup (2 pages)
   (a) Implementation
      - Python (NetworkX)
      - `https://github.com/j0nascz/Paper_SMNCA.git`
   (b) Parameters
      - Louvain: resolution parameter
      - Leiden: resolution parameter
      - Edge betweenness: number of cuts
   (c) Evaluation Criteria
      - Pairwise comparison between methods and parameter settings
        – Changing parameters(e.g resolution parameter) in Leiden and Lovain and comparing results with ARI and AMI
7. Results (3–4 pages)
   (a) Method Comparison
      - Tables: number of communities
      - Modularity values
      - Runtime

(b) Hyperparameter Sensitivity Analysis
    - Plots: resolution vs. number of communities
    - ARI and AMI vs. resolution

(c) Interpretation
    - Stability of communities

8. Discussion (2 pages)
    - Why do Bac and Vir differ?
        - Structural differences between the networks
        - Effect on detected communities
    - Which method is more stable?
        - Interpretation of stability based on ARI and AMI
        - Intra-method vs. inter-method differences
    - Interpretation of ARI/AMI results
    - Resolution limit problem

9. Conclusion (1 page)
    - Summary of findings
    - Main insights
    - Methodological implications
    - Outlook and future work