

Outline of the Seminar Paper

Modularity-based Community Detection in Bacterial and Viral Networks

Your Name

January 31, 2026

Proposed Structure of the Paper (15–20 pages)

1. Introduction (2–3 pages)

(a) Motivation

- Why are networks important?
- Biological networks (protein interactions, gene regulation)
- Social networks
- Information networks
- Why community detection?
- Functional modules
- Organization of complex systems

(b) Definition of a Network

- Graph as a mathematical object
- Nodes and edges
- Directed vs. undirected networks
- Weighted vs. unweighted networks

(c) Mathematical Representation

- Adjacency list
- Adjacency matrix
- Formal definition: $G = (V, E)$

(d) What Can We Do with Networks?

- Descriptive statistics
- Structural analysis
- Prediction
- Clustering and communities

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
 - Edge betweenness
 - Closeness and betweenness (brief)

3. Similarity Measures for Community Comparison (2 pages)

- (a) Rand Index
 - Definition and intuition
- (b) Adjusted Rand Index (ARI)
 - Correction for chance
- (c) Adjusted Mutual Information (AMI)
 - Information-theoretic perspective
- (d) 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
- (d) Leiden Method

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

5. Data Description (2 pages)

(a) Data Source

- Bacterial network
- Viral network
- Protein homology networks

(b) Network Structure

- Nodes represent proteins
- Edges represent sequence similarity
- Weights: bit score or e-value

(c) Preprocessing

- Filtering
- Removal of self-loops
- Largest connected component

6. Experimental Setup (2 pages)

(a) Implementation

- Python (NetworkX)
- R (igraph)

(b) Parameters

- Louvain: resolution parameter
- Leiden: resolution parameter
- Edge betweenness: number of cuts

(c) Evaluation Criteria

- Number of communities
- Community sizes
- Modularity
- ARI
- AMI

7. Results (3–4 pages)

(a) Descriptive Comparison of Bac and Vir

- Size
- Density
- Degree distributions

(b) Method Comparison

- Tables: number of communities

- Modularity values
- Runtime
- (c) Hyperparameter Sensitivity Analysis
 - Plots: resolution vs. number of communities
 - ARI and AMI vs. resolution
- (d) Interpretation
 - Stability of communities
 - Biological plausibility

8. Discussion (2 pages)

- Why do Bac and Vir differ?
- Which method is more stable?
- When does modularity fail?
- Resolution limit problem

9. Conclusion (1 page)

- Summary of findings
- Main insights
- Methodological implications
- Outlook and future work