Ph.D. Research Proposal

Title: Enhancing Spectral Clustering: Robust Graph Laplacians and Scalable Eigen-Computation

1. Problem Statement

Spectral clustering is a widely used and theoretically grounded unsupervised learning technique that utilizes the eigenstructure of graph Laplacians to identify meaningful clusters in data. Despite its popularity and empirical success, classical spectral clustering faces several limitations that hinder its applicability to real-world scenarios:

- Sensitivity to Noise: Small perturbations in the input affinity matrix (e.g., due to measurement error or adversarial noise) can significantly alter the resulting eigenvectors, leading to unreliable cluster assignments.
- Scalability Bottleneck: The reliance on exact eigendecomposition, which scales as O(n^3), makes spectral clustering computationally prohibitive for large-scale datasets.
- Dependence on Affinity Construction: The performance is highly sensitive to the choice of similarity measure (e.g., Gaussian kernels), which may not generalize across domains or reflect true data structure.

Research Objective

This research aims to develop theoretically grounded, robust, and scalable variants of spectral clustering by focusing on:

- Graph Laplacian Stability: Designing noise-resilient Laplacian formulations.
- Efficient Eigen-Decomposition: Creating faster approximate eigensolvers with provable guarantees.
- Data-Adaptive Affinity Learning: Learning optimal similarity metrics from data using self-supervised or neural approaches.

2. Significance and Expected Benefits

Scientific and Practical Impact

- Improved Robustness: Enables reliable clustering of noisy or corrupted datasets in fields such as bioinformatics, finance, and social sciences.
- Scalability: Makes spectral clustering applicable to modern big data problems.
- Versatile Applications: Particularly useful for domains requiring precision, such as fraud detection and single-cell RNA sequencing.

Theoretical Contributions

- Advances in matrix perturbation theory tailored to spectral clustering.
- New connections between spectral methods, deep learning, and graph neural networks (GNNs).
- Theoretical bounds on approximation error and cluster stability under noise.

3. Related Work and Existing Gaps

Robust Spectral Clustering:

- Ng et al. (2002): Pioneering work but lacks robustness under noise.
- Zhang & Rohe (2018): Studied robustness using stochastic block models.
- Cucuringu et al. (2020): Proposed matrix completion to mitigate noise.

Scalable Eigen-Computation:

- Nyström Approximation (Williams & Seeger, 2001): Subsampling for speedup, but lacks robustness.
- Randomized SVD (Halko et al., 2011): Efficient low-rank approximations.
- Neural Approaches (Dong et al., 2019): Promising but not yet well understood.

Adaptive Affinity Learning:

- Zelnik-Manor & Perona (2004): Introduced local scaling.
- Deep Spectral Clustering (Tian et al., 2021): Uses deep embeddings.

Identified Gaps:

- Lack of a unified framework addressing both robustness and scalability.
- Insufficient theoretical guarantees for approximate eigen-decomposition methods.
- Limited exploration of learned affinity functions.

4. Key References

- Von Luxburg (2007). A Tutorial on Spectral Clustering.
- Damle et al. (2019). Fast Spectral Clustering via the Nyström Method.
- Lei & Rinaldo (2015). Consistency of Spectral Clustering in Stochastic Block Models.

5. Research Objectives and Methodology

Objective	Methodology	Expected Outcome
Robust Laplacian Design	Develop regularized and noise-resilient Laplacians using perturbation theory.	Provable stability under noisy input.
Scalable Eigen-Solvers	Leverage randomized numerical linear algebra (e.g., Nyström, Krylov subspace methods).	Near-linear time spectral clustering.

Adaptive Affinity Learning	Apply self-supervised deep learning to learn domain-specific similarity functions.	More accurate and context-sensitive clustering.
Theoretical Guarantees	Derive bounds on spectral gap and clustering accuracy under approximation.	Rigorous performance guarantees.

6. Research Plan

Step 1: Conduct robustness analysis on synthetic and real datasets. Develop new formulations of Laplacians.
Step 2: Implement and benchmark scalable eigen-solvers; explore GPU acceleration.
Step 3: Integrate robustness and scalability into a unified framework. Extend to learned affinity metrics.

7. Evaluation Metrics

- Clustering Accuracy: Adjusted Rand Index (ARI), Normalized Mutual Information (NMI).
- Computational Efficiency: Wall-clock runtime comparison with classical methods.
- Noise Robustness: Accuracy under adversarial and stochastic perturbations.
- Scalability: Performance on datasets exceeding one million samples.

8. Expected Contributions

- A new class of spectral clustering algorithms with theoretical and empirical robustness.
- Fast eigen-computation pipelines suitable for real-world large-scale datasets.
- A modular Python/PyTorch implementation to facilitate community use and reproducibility.
- Peer-reviewed publications and potential applications in computational biology and cybersecurity.

9. Next Steps

- Theoretical Development: Collaborate with domain experts in numerical linear algebra.
- Experimental Validation: Apply the proposed framework to single-cell RNA-sequencing and transaction fraud datasets.
- Open Science: Release datasets, code, and benchmarks to support reproducibility.