Unsupervised Clustering on Networked Data

★ KEY CONCEPTS

1 What is networked data?

Interconnected data points represented as a network; such as a group of individuals (nodes) and whether they know each other (edges).

2 What is community detection?

Finding clusters of densely connected nodes in networked data such that clusters can be gently connected to each other.

3 What are overlapping communities?

Some nodes can naturally belong to more than one cluster. So, the clusters can overlap with each other.

4 Why is this task challenging?

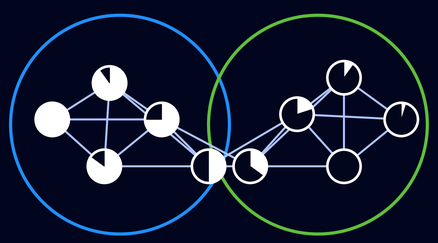
We only have the network topology as input (no features). Real networks are large and noisy. Overlaps add further difficulty.

5 Why does it matter?

Community detection is a foundational data science task with wide applications ranging from market segmentation to drug discovery.

6 Aim of this study

Developing *Highway*, a scalable algorithm that takes a network as input and return overlapping clusters as output using a sparse network backbone.



Assignment of each node to the two clusters (white vs. dark) is a pie chart.

Overlapping communities allow each node to belong to more than one cluster.

- **Zihe Zhou, MIE Department**
- **Supervisor: Prof. Samin Aref, MIE Department**

Algorithm for Scalable Clustering with Overlaps

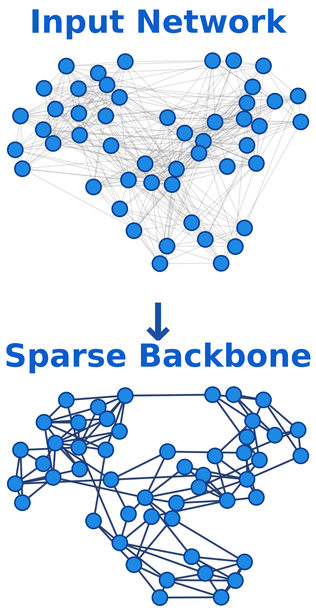
Our community detection algorithm beats most existing methods in performance and scalability.

Mechanical & Industrial Engineering
UNIVERSITY OF TORONTO

Our proposed algorithm described in 4 steps

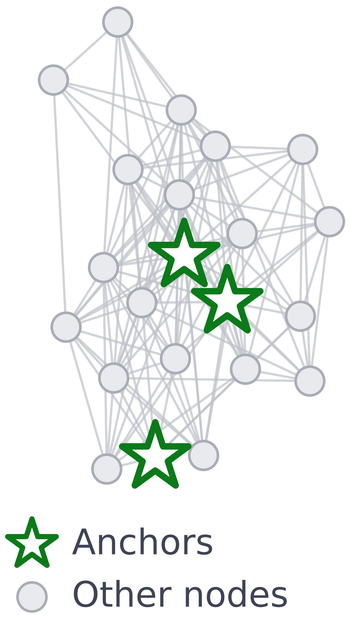
1 Sparse Backbone Extraction

Keep informative edges and remove redundant edges.



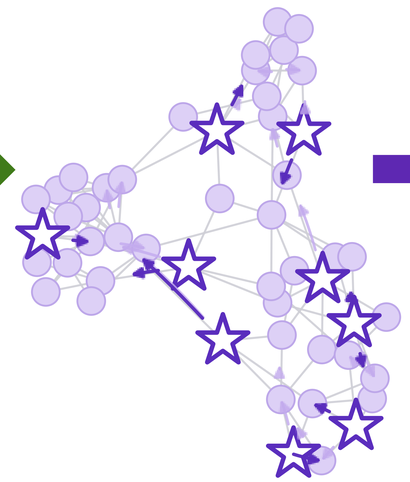
2 Anchor Selection

Choose representative nodes as effective starting points.



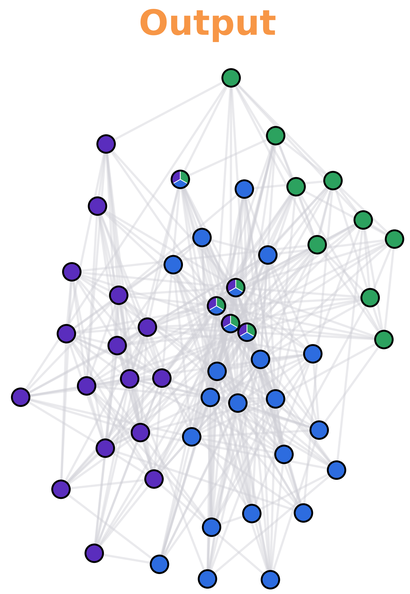
3 Neighbor-only Propagation

Propagate anchor-indexed memberships on the backbone.

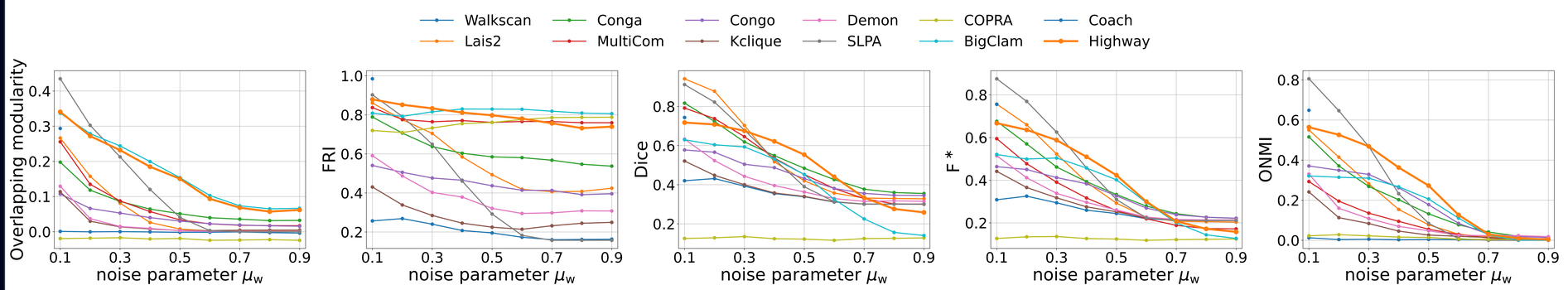


4 Overlapping Community Decoding

Decode the anchor-indexed memberships as community memberships



Comparing against 11 methods on standard benchmarks (the higher the better)



Comparative performance results

- **Highway** algorithm achieves strong **overlapping modularity**.
- **Performance** remains **competitive** as the noise parameter μ_w varies.
- **Backbone propagation** improves efficiency while preserving the structure.
- **Highway** algorithm scales better than most alternatives as instance size increases.
- **Sparse backbone construction** reduces unnecessary propagation cost.
- **Our runtime** remains among the **fastest** while returning **accurate** results.

Preprint: will be made available on arXiv. **Code:** Will be made available on GitHub.