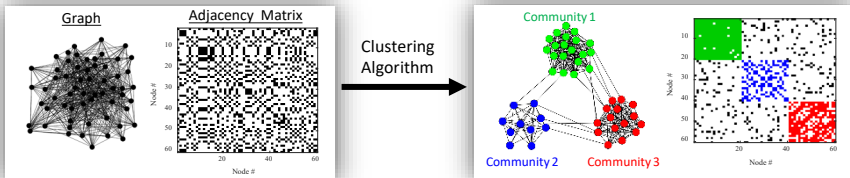


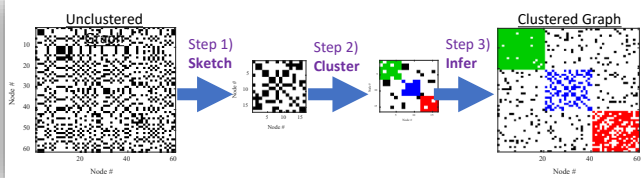
Abstract

This paper proposes a sketch-based approach to the community detection problem which clusters the full graph through the use of an informative and concise sketch. The reduced sketch is built through an effective sampling approach which selects few nodes that best represent the complete graph and operates on a pairwise node similarity measure based on the average commute time. After sampling, the proposed algorithm clusters the nodes in the sketch, and then infers the cluster membership of the remaining nodes in the full graph based on their aggregate similarity to nodes in the partitioned sketch. By sampling nodes with strong representation power, our approach can improve the success rates over full graph clustering. In challenging cases with large node degree variation, our approach not only maintains competitive accuracy with full graph clustering despite using a small sketch, but also outperforms existing sampling methods. The use of a small sketch allows considerable storage savings, and computational and timing improvements for further analysis such as clustering and visualization. We provide numerical results on synthetic data based on the homogeneous, heterogeneous and degree corrected versions of the stochastic block model, as well as experimental results on real-world data.

Traditional Community Detection



Our Approach: Sketch-based Clustering



We minimize the clustering bottleneck by using a small sketch.

We can incorporate any existing clustering algorithm.

Sampling

Representative Node Sampling

Obtain a reduced sketch which best “represents” the full graph.

- Built upon a positive-definite similarity measure
- Aims at maximizing the information gain, while minimizing redundancy

Modeling: Encode the representation power of each sample in a vector

Find an optimal representation power encoding for all samples aggregated in $\mathbf{R} \in \mathbb{R}^{N \times N}$

$$\min_{\mathbf{R}} \text{tr} \{ \mathbf{R}^T \mathbf{S} \mathbf{R} - 2 \mathbf{S} \mathbf{R} \} + \lambda \| \mathbf{R}^T \|_{1,2}$$

- Rewards representing the other sample
- Penalizes similarity among the chosen samples
- Obtains a reduced subset using structured sparsity
- Handles complex inter-relations through non-linear formulation

Similarity via Average Commute Time

Average commute time $\mathcal{C}(i, j)$:
Time for a random walk from node i to node j and then back again.

Why is it useful? 1) It reflects community structure
2) We can form embedding where distance between nodes is $\sqrt{\mathcal{C}(i, j)}$

The pseudoinverse of the Laplacian \mathbf{L}^+ is the gram matrix for the embedded nodes. Our similarity matrix \mathbf{S} is the cosine similarity derived from \mathbf{L}^+ :

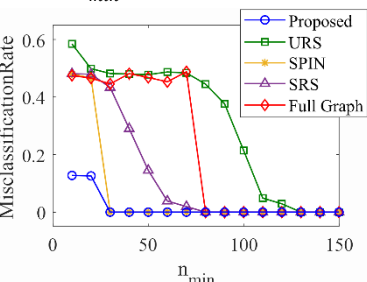
$$s_{ij} = \frac{l_{ij}^+}{\sqrt{l_{ii}^+ l_{jj}^+}}$$

The non-zero row-norms of the matrix \mathbf{R} identifies the representative nodes.

Experiments

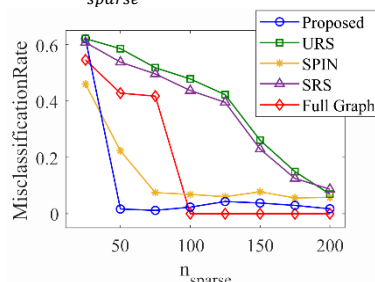
Homogeneous SBM

Uniform intra-cluster edge density for graph. Three clusters - n_{min} is size of smallest cluster. Smaller n_{min} indicates more imbalance.



Heterogeneous SBM

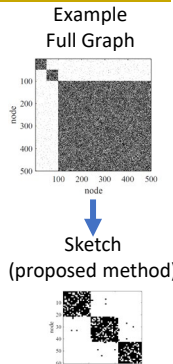
Intra-cluster edge density varies for each cluster. Three clusters. Smaller n_{sparse} indicates more imbalance.



Clustering Time

Time in seconds - homogeneous SBM

N	Sketch-based	Full Graph
400	4×10^{-2}	3.0×10^0
800	1.8×10^{-1}	1.4×10^1
1600	7.2×10^{-1}	9.3×10^1
3200	1.9×10^0	7.8×10^2
6400	9.7×10^0	6.3×10^3
12800	6.7×10^1	4.6×10^4



DCSBM

p	Full Graph	Sketch-based Algorithm			
		Proposed	URS	SPIN	SRS
0.05	0.064	0.052	0.326	0.432	0.362
0.10	0.004	0.004	0.122	0.333	0.150
0.15	0.002	0.002	0.034	0.214	0.058
0.20	0	0	0.011	0.090	0.019

Political Blogs Dataset

Dataset	Full Graph	Sketch-based Algorithm			
		Proposed	URS	SPIN	SRS
Full	0.050	0.052	0.178	0.438	0.218
Unbalanced	0.437	0.142	0.224	0.334	0.289

Misclassification rate

Misclassification rate