Weighted Flow Diffusion for Local Graph Clustering with Node Attributes
An Algorithm and Statistical Guarantees

Shenghao Yang and Kimon Fountoulakis

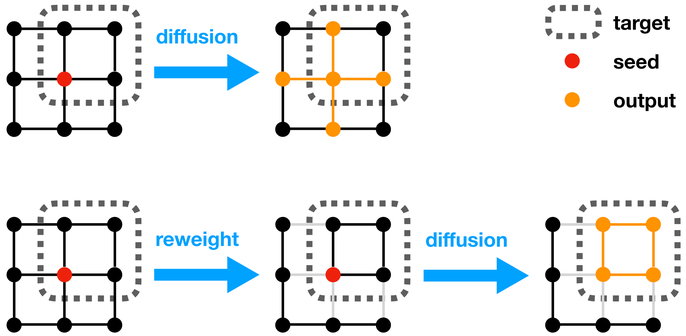
Introduction

TL; DR: We provide a very simple local clustering algorithm with provable guarantees for attributed graphs

- Graphs with node attributes which provide additional information are becoming increasingly available
- We study local clustering in attributed graphs, based on reweighing edges from a Gaussian kernel of node attributes and then locally diffusing mass in the graph
- We study how reweighing edges can help recover a target cluster from a contextual local random graph model
- We conduct experiments to illustrate the results

Intuition: why reweighing edges?

- If boundary edges (i.e. a cut) that connect a target cluster to the outside have very small weights, then this forces a diffusion to spread mass within the target
- Fewer mass leak can lead to better local clustering result
- The following toy example on a grid graph shows the difference between diffusion in unweighted and reweighed graphs



Local graph clustering with attributes

- **Input:** Graph $G = (V, E)$, seed node $s \in V$, attributes $X_i \in \mathbb{R}^d$
- **Algorithm** (informal)**:**
  - Define weighted graph $G' = (V, E, w)$ with edge weight
    $$w_{ij} = \exp(-\gamma \|X_i - X_j\|^2) \text{ if } (i,j) \in E$$
  - Run weighted local graph diffusion in G' starting from s
  - Check where and how the source mass spread within G'
  - Obtain an output cluster (by applying some rounding procedure)

The algorithm is local and does not require processing all data points

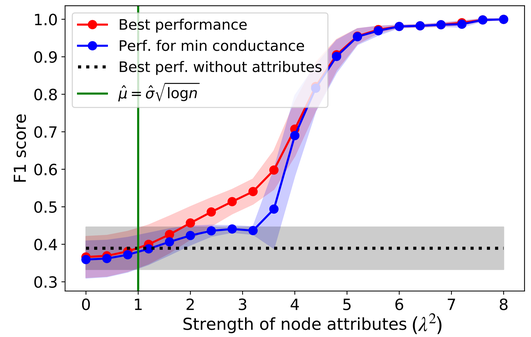
Contextual local random graph model

- Given a set of nodes V and a target cluster $K \subset V$
  - Draw an edge (i,j) with probability p if $i \in K, j \in K$
  - Draw an edge (i,j) with probability q if $i \in K, j \notin K$
  - Edges (i,j) where $i, j \notin K$ can be arbitrary
- Every node $i \in V$ has d-dimensional attributes $X_i = \mu_i + Z_i$
  - Signal $\mu_i = \mu_j$ for all $i, j \in K$
  - Noise Z_i has independent mean zero sub-Gaussian coordinates
  - Attribute-side cluster-wise signal and noise are
    $$\hat{\mu} = \min_{i \in K, j \notin K} \|\mu_i - \mu_j\| \text{ and } \hat{\sigma} = \max_{1 \le \ell \le d} \sigma_\ell$$
    where σ_ℓ is coordinate-wise sub-Gaussian variance proxy
  - **Assumption:** $\hat{\mu} = \omega(\hat{\sigma}\sqrt{\lambda \log |V|})$ for some λ

Performance guarantee

- **Very good node attributes:** local diffusion in the reweighed graph G' fully recovers K with zero false positives
- **Moderately good node attributes:** local diffusion in the reweighed graph G' fully recovers K with $O(1/\eta^2 - 1)|K|$ false positives, where
  $$\eta = \frac{p \cdot |K|}{p \cdot |K| + q \cdot |V \backslash K| \cdot e^{-\omega(\lambda)}}$$
  $p \cdot |K|$: internal connectivity; $q \cdot |V \backslash K|$: external connectivity; $\lambda \equiv$ attribute strength
  - This shows the recovery is jointly controlled by graph structure and node attributes

Experiments



- Plot shows recovery of a randomly chosen target cluster from 20-block SBM on 10,000 nodes and 100-dim Gaussian attributes
- Experiments on two real-world co-authorship networks show that, on average over 20 different target clusters, using node attributes lead to 4.3% increase in the F1 score
- Additional comparisons with global spectral and classification baselines are found in the paper