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

Shenghao Yang, Kimon Fountoulakis
Local graph clustering

Setting: Given a graph $G = (V, E)$, and a seed node $s \in V$

Goal: Find a good cluster that contains $s$, without necessarily exploring the whole graph
Local graph clustering

**Setting:** Given a graph $G = (V, E)$, and a seed node $s \in V$

**Goal:** Find a good cluster that contains $s$, without necessarily exploring the whole graph

- Random walk [Spielman & Teng 2013]
- PageRank [ACL 2006]
- Heat kernel [Chung 2007]
- Evolving sets [Andersen & Peres 2008]
- Capacity releasing diffusion [Di et al 2017]
- Flow diffusion [Fountoulakis et al 2020]
and many more…
Local graph clustering

Setting (this work): Given a graph $G = (V, E)$ with node attributes, and a seed node $s \in V$

Goal: Find a good cluster that contains $s$, without necessarily exploring the whole graph.
Contributions

• A simple algorithm for local clustering in attributed graphs based on reweighing edges from a Gaussian kernel of node attributes and then locally diffusing mass in the graph

• A theoretical analysis on the recovery of an unknown target cluster in a contextual random graph model

• Experiments over synthetic and real-world data to illustrate our results
Local graph diffusion

- Generic process to spread mass from a seed node to nearby nodes via edges in the graph
- PageRank, random walk: spread probability mass
- Capacity releasing diffusion, flow diffusion: spread source mass
- Mass tend to spread within well-connected clusters
Local graph clustering

- **Input:** Graph $G = (V, E)$, seed node $s \in V$

- **Algorithm** (informal):
  - Run local graph diffusion in $G$ starting from $s$
  - Check where and how the mass spread within $G$ around $s$
  - Obtain an output cluster (by applying rounding/post-processing)
Local graph clustering

- **Input:** Graph $G = (V, E)$, seed node $s \in V$, node attributes $X_i \in \mathbb{R}^d$, $\forall i$

- **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 mass spread within $G'$ around $s$

  - Obtain an output cluster (by applying rounding/post-processing)
How does reweighing edges help exactly?
Example: how edge weights can help
Example: how edge weights can help
Contextual local random 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** $Z_i \sim N(0, \sigma^2 I)$ for all $i$
  - $\hat{\mu} := \min_{i \in K, j \notin K} \|\mu_i - \mu_j\|$  
    - **Assumption:** $\hat{\mu} = \omega(\sigma \sqrt{\lambda \log |V|})$ for some $\lambda$
Recovery guarantees

- Given a seed node $s \in K$, the goal is to recover $K$

- If we have **very good node attributes**: local diffusion on the reweighed graph fully recovers $K$ with no false positives, as long as $K$ is connected

- If we have **moderately good node attributes**: local diffusion on the reweighed graph fully recovers $K$ with $O(1/\eta^2 - 1)|K|$ false positives, where

$$
\eta = \frac{p \cdot |K|}{p \cdot |K| + q \cdot |V\setminus K| \cdot e^{-\omega(\lambda)}}
$$

- $\lambda = \text{node attribute signal}$

**Internal connectivity**

$\eta = \frac{p \cdot |K|}{p \cdot |K| + q \cdot |V\setminus K| \cdot e^{-\omega(\lambda)}}$

**External connectivity**
Recovery guarantees

- If we have **moderately good node attributes**:
  local diffusion on the reweighed graph fully recovers $K$ with $O(1/\eta^2 - 1) \cdot |K|$ false positives, where

  $\eta = \frac{p \cdot |K|}{p \cdot |K| + q \cdot |V\backslash K|} \cdot e^{-\omega(\lambda)}$  \quad \lambda = \text{node attribute signal}

- [Ha et al., 2021] Approximate Personalized PageRank on an **unweighted** graph fully recovers $K$ with $O(1/\eta^2 - 1) \cdot |K|$ false positives, where

  $\eta = \frac{p \cdot |K|}{p \cdot |K| + q \cdot |V\backslash K|}$
Experiments on real-world data

- Co-authorship networks
- Target clusters are ground-truth communities based on authors’ primary research area
- Average F1 score over 100 trials for each target cluster
- Overall 4.3% increase in F1 over 20 clusters

<table>
<thead>
<tr>
<th>Network</th>
<th>Cluster</th>
<th>No attr.</th>
<th>Use attr.</th>
<th>Improv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bioinformatics</td>
<td>32.1</td>
<td>39.3</td>
<td>7.2</td>
<td></td>
</tr>
<tr>
<td>Machine Learning</td>
<td>30.9</td>
<td>37.3</td>
<td>6.4</td>
<td></td>
</tr>
<tr>
<td>Computer Vision</td>
<td>37.6</td>
<td>35.5</td>
<td>-2.1</td>
<td></td>
</tr>
<tr>
<td>NLP</td>
<td>45.2</td>
<td>52.3</td>
<td>7.1</td>
<td></td>
</tr>
<tr>
<td>Graphics</td>
<td>38.6</td>
<td>49.2</td>
<td>10.6</td>
<td></td>
</tr>
<tr>
<td>Networks</td>
<td>44.1</td>
<td>47.0</td>
<td>2.9</td>
<td></td>
</tr>
<tr>
<td>Security</td>
<td>29.9</td>
<td>35.7</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>Databases</td>
<td>48.5</td>
<td>58.1</td>
<td>9.6</td>
<td></td>
</tr>
<tr>
<td>Data Mining</td>
<td>27.5</td>
<td>28.8</td>
<td>1.3</td>
<td></td>
</tr>
<tr>
<td>Game Theory</td>
<td>60.6</td>
<td>66.0</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>HCI</td>
<td>70.0</td>
<td>77.6</td>
<td>7.6</td>
<td></td>
</tr>
<tr>
<td>Information Theory</td>
<td>47.4</td>
<td>46.9</td>
<td>-0.5</td>
<td></td>
</tr>
<tr>
<td>Medical Informatics</td>
<td>65.7</td>
<td>70.3</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>Robotics</td>
<td>59.9</td>
<td>59.9</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Theoretical CS</td>
<td>66.3</td>
<td>70.7</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>Phys. Rev. A</td>
<td>69.4</td>
<td>70.9</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Phys. Rev. B</td>
<td>41.4</td>
<td>42.3</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Phys. Rev. C</td>
<td>79.3</td>
<td>82.1</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>Phys. Rev. D</td>
<td>62.3</td>
<td>68.9</td>
<td>6.6</td>
<td></td>
</tr>
<tr>
<td>Phys. Rev. E</td>
<td>49.5</td>
<td>53.7</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>AVERAGE</td>
<td>50.3</td>
<td>54.6</td>
<td>4.3</td>
<td></td>
</tr>
</tbody>
</table>
Thank you!