Research Talk

Graph Partitioning Clustering with User-Specified Relative Density

February 1, 2023

Ms. Rohi Tariq (Ph.D student)
Research Advisors: Assoc. Prof. Dr. Kittichai Lavangnananda, Assoc. Prof. Dr. Pascal Bouvry, Assoc. Prof. Dr. Pornchai Mongkolnam

School of Information Technology
King Mongkut’s University of Technology Thonburi
Outline

• Graph Clustering Algorithms Categorization
• Graph Clustering Quality Evaluation Metrics Categorization Based on Internal and External Connectivity

• Graph Clustering Algorithm Based on Clique Strategy
  1. Algorithm-1 Cliques Determination
  2. Algorithm-2 Clusters Determination

• Metric Mean Relative Density Deviation Coefficient (MRDDC)
• Experimental Results (Real-world Networks)
• Algorithm Effectiveness and Process Evaluation
# Graph Clustering Algorithms Categorization

## Parameter-Free Algorithms
- **Fast-Greedy** (Modularity Optimization)
- **Info-Map** (Greedy Approach) (Information theory)
- **Label propagation** (Labels)
- **Walktrap** (Random Walk)
- **Spinglass** (Potts Model)
- **Leading eigenvector** (Modularity Matrix)
- **Graph-based k-means**

## Parameterizable Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Parameter/Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Girvan-Newman (G)</td>
<td>(Edge-Betweenness, Modularity Optimization)</td>
</tr>
<tr>
<td>Level-Cut</td>
<td>Tunable parameters</td>
</tr>
<tr>
<td>Louvain (G)</td>
<td>(Modularity Optimization)</td>
</tr>
<tr>
<td>Resolution-Limit (γ)</td>
<td>Tunable Parameter</td>
</tr>
<tr>
<td>Leiden (G)</td>
<td>(Modularity Optimization)</td>
</tr>
<tr>
<td>Resolution-Limit (γ)</td>
<td>Tunable Parameter</td>
</tr>
<tr>
<td>Fluid-Communities.</td>
<td>(k)</td>
</tr>
<tr>
<td></td>
<td>Number of Clusters</td>
</tr>
<tr>
<td>Spectral Methods</td>
<td>ε-neighborhood, k-nearest</td>
</tr>
<tr>
<td></td>
<td>Connectivity parameter</td>
</tr>
<tr>
<td>k-means</td>
<td>(k)</td>
</tr>
<tr>
<td></td>
<td>Number of Clusters</td>
</tr>
<tr>
<td>Expectation Maximization</td>
<td>(k)</td>
</tr>
<tr>
<td></td>
<td>Number of Clusters</td>
</tr>
</tbody>
</table>
Graph Clustering Quality Metrics Categorization

**Internal Connectivity parameter**

- **Coverage:**
  \[ \frac{\sum_{i=1}^{k} \text{int}_{deg}(C_i)}{e} \]

- **Internal-Edge Density:**
  \[ \delta_{int}(C_i) = \frac{\text{int}_{deg}(C_i)}{C_n(C_n-1)/2} \]

- **Average Embeddedness:**
  \[ \text{emb}(G\mid C_1, C_2, \ldots, C_k) = \frac{1}{k} \sum_{i=1}^{k} \text{emb}(v, C_i) \]

- **Triangle Participation Ratio:**
  \[ \text{TPR}(G\mid C_1, C_2, \ldots, C_k) = \frac{1}{k} \sum_{i=1}^{k} \text{TPR}(C_i) \]

- **FOMD (Fraction Over Median Degree):**
  \[ \text{FOMD}(G\mid C_1, C_2, \ldots, C_k) = \frac{1}{k} \sum_{i=1}^{k} \text{FOMD}(C_i) \]

**External Connectivity parameter**

- **Conductance:**
  \[ \phi(G\mid C_1, C_2, \ldots, C_k) = \frac{1}{k} \sum_{i=1}^{k} \phi(C_i) \phi(C_i) = \frac{\text{deg}_{ext}(C_i)}{\min(\text{deg}_{total}(C_i), 2m - \text{deg}_{total}(C_i))} \]

- **Flake-Odf (Out Degree Fraction):**
  \[ \text{Flake} \cdot \text{ODF}(C_i) = \frac{1}{k} \sum_{i=1}^{k} \text{Flake - ODF}(C_i) \]

- **Nor-Cut (Normalized Cut):**
  \[ \frac{\text{ext}_{deg}(C_i)}{2 \cdot \text{int}_{deg}(C_i) + \text{ext}_{deg}(C_i)} + \frac{\text{ext}_{deg}(C_i)}{2 \cdot (m - \text{int}_{deg}(C_i)) + \text{ext}_{deg}(C_i)} \]

- **Cut-Ratio:**
  \[ \text{CR}(C_i) = \frac{\text{ext}_{deg}(C_i)}{C_n(n-C_n)} \]

**Graph Clustering Quality Evaluation Metrics**

- **Graph Clustering Quality Metrics Categorization**

3
Graph Clustering Algorithm Based on Clique Strategy

Relative Density Metric/ User-Specified Relative Density $U(\delta_r)$

Equation: $C_i(\delta_r) = \frac{\sum_{v \in C_i} int_{deg}(v)}{\sum_{v \in C_i} int_{deg}(v) + ext_{deg}(C_i)}$ (Value ranges from 0 to 1)

Phase 1: Algorithm-1 $\rightarrow$ Cliques Determination in Graph $(G)$

Input: Connected Un-weighted and Un-directed Graph

- Spanning Tree
- Fundamental Cycles

Outcome $\rightarrow$ Cliques (Triangles) and their associated degrees detection

Phase 2: Algorithm-2 $\rightarrow$ Cluster Determination

- User-Specified Relative Density $U(\delta_r)$
- Connected Un-weighted and Un-directed Graph, Clique List

Outcome $\rightarrow$ Clusters with equal or Closer to the user desired Density $U(\delta_r)$
Algorithm Illustration with a Simple Random Graph Example

Input: Random Graph

\[ V = 12 \]
\[ E = 19 \]
\[ \text{Triangles} = 7 \]
\[ U(\delta_r) = 0.45 \]
• 8 Real-world networks have been experimented with in this study

• Diverse properties and connectivity structures.

1. Zachary's Karate Club
2. American College Football
3. US-Grid Power Network
Depicts the relative frequency distribution of 8 real-world networks which tell us about the characteristic's comprehension. \( p_k \) is relative frequency and can also be thought of as a probability, the likelihood that a node has a degree of exactly \( k \). \( n_k \) represents the number of nodes with degree \( (k) \), and \( n \) is the count of nodes in a graph.

Connectivity structures and Degree Distribution of the networks

- **Zachary's Karate Network**
  - Average Degree = 4.5

- **Aves-Weaver Network**
  - Average Degree = 7.2

- **Dolphins Network**
  - Average Degree = 5

- **Les Misérables Network**
  - Average Degree = 6.6

- **Political Books Network**
  - Average Degree = 8.4

- **American Football Network**
  - Average Degree = 10.8

- **FB-Pages Food Network**
  - Average Degree = 6.8

- **Power Grid Network**
  - Average Degree = 2.7
Results (Real-world Networks)

- Real structures of three networks: (a) Zachary's Karate Club, (b) American College Football, and (c) US Grid Power.
- The average degree of connectivity is 4.5, 10.8, and 2.7 respectively.
Results (Real-world Networks):

Clustering results with a $U(\delta_r)$ value of 0.4

(a) Zachary's Karate Network
$U(\delta_r) = 0.4$
Identified 4 Clusters

(b) American college Football Network
$U(\delta_r) = 0.4$
Identified 12 Clusters

(c) US Grid Power Network
$U(\delta_r) = 0.4$
Identified 180 Clusters
Results (Real-world Networks):

Clustering results with a $U(\delta_r)$ value of 0.6

Zachary's Karate Network
$U(\delta_r) = 0.6$
Identified 3 Clusters

American college Football Network
$U(\delta_r) = 0.6$
Identified 4 Clusters

US-Grid Power Network
$U(\delta_r) = 0.6$
Identified 84 Clusters
### Results (Real-world Networks)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No of Clusters ($C_k$)</th>
<th>Average Relative Density $Avg C(\delta_r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zachary's Karate</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>American College Football</td>
<td>12</td>
<td>0.4</td>
</tr>
<tr>
<td>US-Grid Power</td>
<td>180</td>
<td>0.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>No of Clusters ($C_k$)</th>
<th>Average Relative Density $Avg C(\delta_r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zachary's Karate</td>
<td>3</td>
<td>0.6</td>
</tr>
<tr>
<td>American College Football</td>
<td>4</td>
<td>0.6</td>
</tr>
<tr>
<td>US-Grid Power</td>
<td>84</td>
<td>0.6</td>
</tr>
</tbody>
</table>
• $U(\delta_r)$ value of 0.4

• Graph structural and clustering type impact on cluster size with respect to number of nodes.
• $U(\delta_r)$ value of 0.6
• Graph Structural and clustering type impact on cluster size with respect to number of nodes.
Algorithm Effectiveness and Process Evaluation with the Existing Quality Metrics:

- Clustering Results Evaluation based on Internal Connectivity

<table>
<thead>
<tr>
<th>Network</th>
<th>Conductance</th>
<th>Cut-Ratio</th>
<th>Nor-Cut</th>
<th>Flake-Odf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zachary's Karate</td>
<td>0.41</td>
<td>0.25</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td>Aves-weaver</td>
<td>0.42</td>
<td>0.34</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>Dolphins-interaction</td>
<td>0.4</td>
<td>0.27</td>
<td>0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>Les Misérables</td>
<td>0.42</td>
<td>0.22</td>
<td>0.06</td>
<td>0.03</td>
</tr>
<tr>
<td>Political-Books</td>
<td>0.47</td>
<td>0.33</td>
<td>0.03</td>
<td>0.02</td>
</tr>
<tr>
<td>American college Football</td>
<td>0.42</td>
<td>0.25</td>
<td>0.04</td>
<td>0.02</td>
</tr>
<tr>
<td>FB-Pages-Food</td>
<td>0.43</td>
<td>0.26</td>
<td>0.006</td>
<td>0.005</td>
</tr>
<tr>
<td>US-Grid power</td>
<td>0.41</td>
<td>0.37</td>
<td>0.003</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

- Clustering Results Evaluation based on External Connectivity

<table>
<thead>
<tr>
<th>Network</th>
<th>Coverage</th>
<th>Average Embeddedness</th>
<th>FOMD</th>
<th>Average TPR</th>
<th>Internal Edge Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zachary's Karate</td>
<td>0.46</td>
<td>0.69</td>
<td>0.7</td>
<td>0.8</td>
<td>0.3</td>
</tr>
<tr>
<td>Aves-weaver</td>
<td>0.51</td>
<td>0.7</td>
<td>0.6</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>Dolphins-interaction</td>
<td>0.52</td>
<td>0.63</td>
<td>0.6</td>
<td>0.7</td>
<td>0.3</td>
</tr>
<tr>
<td>Les Misérables</td>
<td>0.52</td>
<td>0.69</td>
<td>0.6</td>
<td>0.8</td>
<td>0.29</td>
</tr>
<tr>
<td>Political-Books</td>
<td>0.54</td>
<td>0.65</td>
<td>0.58</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>American college Football</td>
<td>0.53</td>
<td>0.8</td>
<td>0.6</td>
<td>0.74</td>
<td>0.17</td>
</tr>
<tr>
<td>FB-Pages-Food</td>
<td>0.51</td>
<td>0.64</td>
<td>0.7</td>
<td>0.78</td>
<td>0.35</td>
</tr>
<tr>
<td>US-Grid power</td>
<td>0.22</td>
<td>0.49</td>
<td>0.65</td>
<td>0.79</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Algorithm Effectiveness and Process Evaluation:

- Clustering Results Evaluation based on Internal Connectivity
Algorithm Effectiveness and Process Evaluation:

- Clustering Results Evaluation based on External Connectivity
Proposed Quality Metric Mean Relative Density Deviation Coefficient (MRDDC):

- The notion of "Mean Relative Density Deviation Coefficient" is defined as a metric (MRDDC) for assessing the proximity between the relative density specified by the user and the cluster calculated density discover by the algorithm.
- The equation below represents the metric's formal mathematical definition.

$$MRDDC = 1 - \frac{1}{k} \sum_{k=1}^{C_k} \left| \frac{U(\delta_r) - C_i(\delta_r)}{U(\delta_r)} \right|$$

- The optimal deviation value for each cluster is zero, while the ideal value for this mean coefficient is 1.
The algorithm is not sensitive to structural characteristics variations in the inputs and can still produce meaningful results under a wide range of conditions.
Continuation of Prior Work

- Overlapping Clustering
- Hierarchal Clustering

2: Edge-based Clustering
- Partitioning Clustering
- Overlapping Clustering
- Hierarchal Clustering

3: Algorithm Sensitivity
Thank You

Q/A