AMOS: AN AUTOMATED MODEL ORDER SELECTION ALGORITHM FOR SPECTRAL GRAPH CLUSTERING

Pin-Yu Chen  Thibaut Gensollen  Alfred O. Hero III, Fellow, IEEE

Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, USA
{pinyu, thibautg, hero}@umich.edu

ABSTRACT

One of the longstanding problems in spectral graph clustering (SGC) is the so-called model order selection problem: automated selection of the correct number of clusters. This is equivalent to the problem of finding the number of connected components or communities in an undirected graph. In this paper, we propose AMOS, an automated model order selection algorithm for SGC. Based on a recent analysis of clustering reliability for SGC under the random interconnection model, AMOS works by incrementally increasing the number of clusters, estimating the quality of identified clusters, and providing a series of clustering reliability tests. Consequently, AMOS outputs clusters of minimal model order with statistical clustering reliability guarantees. Comparing to three other automated graph clustering methods on real-world datasets, AMOS shows superior performance in terms of multiple external and internal clustering metrics.

1. INTRODUCTION

Undirected graphs are widely used for network data analysis, where nodes represent entities or data samples, and the existence and strength of edges represent relations or affinity between nodes. The goal of graph clustering is to group the nodes into clusters of high similarity. Applications of graph clustering, also known as community detection [1,2], include but are not limited to graph signal processing [3–11], multivariate data clustering [12–14], image segmentation [15,16], and network vulnerability assessment [17].

Spectral clustering [12–14] is a popular method for graph clustering, which we refer to as spectral graph clustering (SGC). It works by transforming the graph adjacency matrix into a graph Laplacian matrix [18], computing its eigendecomposition, and performing K-means clustering [19] on the eigenvectors to partition the nodes into clusters. Although heuristic methods have been proposed to automatically select the number of clusters [12,13,20], rigorous theoretical justifications on the selection of the number of eigenvectors for clustering are still lacking and little is known about the capabilities and limitations of spectral clustering on graphs.

Based on a recent development of clustering reliability analysis for SGC under the random interconnection model (RIM) [21], we propose a novel automated model order selection (AMOS) algorithm for SGC. AMOS works by incrementally increasing the number of clusters, estimating the quality of identified clusters, and providing a series of clustering reliability tests. Consequently, AMOS outputs clusters of minimal model order with statistical clustering reliability guarantees. Comparing the clustering performance on real-world datasets, AMOS outperforms three other automated graph clustering methods in terms of multiple external and internal clustering metrics.

2. RELATED WORK

Most existing model selection algorithms specify an upper bound $K_{\text{max}}$ on the number of clusters and then select $K$ based on optimizing some objective function, e.g., the goodness of fit of the $k$-cluster model for $k = 2, \ldots, K_{\text{max}}$. In [12], the objective is to minimize the sum of cluster-wise Euclidean distances between each data point and the centroid obtained from K-means clustering. In [20], the objective is to maximize the gap between the $K$-th largest and the $(K + 1)$-th largest eigenvalue. In [13], the authors propose to minimize an objective function that is associated with the cost of aligning the eigenvectors with a canonical coordinate system. In [22], the authors propose to iteratively divide a cluster based on the leading eigenvector of the modularity matrix until no significant improvement in the modularity measure can be achieved. The Louvain method in [23] uses a greedy algorithm for modularity maximization. In [24,25], the authors propose to use the eigenvectors of the nonbacktracking matrix for graph clustering, where the number of clusters is determined by the number of real eigenvalues with magnitude larger than the square root of the largest eigenvalue. The proposed AMOS algorithm not only automatically selects the number of clusters but also provides multi-stage statistical tests for evaluating clustering reliability of SGC.

3. THEORETICAL FRAMEWORK FOR AMOS

3.1. Random interconnection model (RIM)

Consider an undirected graph where its connectivity structure is represented by an $n \times n$ binary symmetric adjacency matrix $A$, where $n$ is the number of nodes in the graph. $[A]_{uv} = 1$ if there exists an edge between the node pair $(u, v)$, and otherwise $[A]_{uv} = 0$. An unweighted undirected graph is completely specified by its adjacency matrix $A$, while a weighted undirected graph is specified by a non-negative matrix $W$, where nonzero entries denote the edge weights.

Assume there are $K$ clusters in the graph and denote the size of cluster $k$ by $n_k$. The size of the largest and smallest cluster is denoted by $n_{\text{max}}$ and $n_{\text{min}}$, respectively. Let $A_k$ denote the $n_k \times n_k$ adjacency matrix representing the internal edge connections in cluster $k$ and let $C_{ij}$ ($i,j \in \{1, 2, \ldots, K\}$, $i \neq j$) be an $n_k \times n_j$ matrix representing the adjacency matrix of inter-cluster edge connections between the cluster pair $(i,j)$. The matrix $A_k$ is symmetric and $C_{ij} = C_{ji}^T$ for all $i \neq j$.

The random interconnection model (RIM) [21] assumes that: (1) the adjacency matrix $A_k$ is associated with a connected graph of $n_k$ nodes but is otherwise arbitrary; (2) the $K(K-1)/2$ matrices
\{C_{ij}\}_{i,j} are random mutually independent, and each $C_{ij}$ has i.i.d.
Bernoulli distributed entries with Bernoulli parameter $p_{ij} \in [0,1]$.
(3) For undirected weighted graphs the edge weight of each inter-cluster edge between clusters $i$ and $j$ is independently drawn from a common nonnegative distribution with mean $\bar{W}_{ij}$ and bounded fourth moment. In particular, we call this model a homogeneous RIM when all random interconnections have equal probability and mean edge weight, i.e., $p_{ij} = p$ and $\bar{W}_{ij} = \bar{W}$ for all $i \neq j$. Otherwise, the model is called an inhomogeneous RIM.

3.2. Spectral graph clustering (SGC)

The graph Laplacian matrix of the entire graph is defined as $L = S - W$, where $S = \text{diag}(W_1, \ldots, W_n)$ is a diagonal matrix and $W_i(0_{1n})$ is the $n \times 1$ column vector of ones (zeros). Similarly, the graph Laplacian matrix accounting for the within-cluster edges of cluster $k$ is denoted by $L_k$. We also denote the i-th smallest eigenvalue of $L$ by $\lambda_i(L)$ and define the partial eigenvalue sum $S_{2,k}(L) = \sum_{i=\lambda_k+1}^{K} \lambda_i(L)$. To partition the nodes in the graph into $K$ ($K \geq 2$) clusters, spectral clustering [14] uses the $K$ eigenvectors $\{u_k\}_{k=1}^{K}$ associated with the $K$ smallest eigenvalues of $L$. Each node can be viewed as a $K$-dimensional vector in the subspace spanned by these eigenvectors. $K$-means clustering [19] is then implemented on the $K$-dimensional vectors to group the nodes into $K$ clusters.

Throughout this paper we assume the graph is connected, otherwise the connected components can be easily found and the proposed algorithm can be applied to each connected component separately. If the graph is connected, by the definition of the graph Laplacian matrix $L$, the smallest eigenvector $u_1$ is a constant vector and $\lambda_1(L) > 0 \forall i \geq 2$. As a result, for connected undirected graphs, it suffices to use the $K-1$ eigenvectors $\{u_k\}_{k=2}^{K}$ of $L$ for SGC. In particular, these $K-1$ eigenvectors are represented by the columns of the eigenvector matrix $Y = [u_2, u_3, \ldots, u_K] \in \mathbb{R}^{n \times (K-1)}$.

3.3. Phase transitions under homogeneous RIM

Let $Y = [Y_1^T, Y_2^T, \ldots, Y_K^T]^T$ be the cluster partitioned eigenvector matrix associated with $L$ for SGC, where $Y_k \in \mathbb{R}^{n \times (K-1)}$ with its rows indexing the nodes in cluster $k$. Under the homogeneous RIM, let $t = p - \bar{W}$ be the inter-cluster edge connectivity parameter. Fixing the within-cluster edge connections and varying $t$, Theorem 1 below shows that there exists a critical value $t^*$ that separates the behavior of $Y$ for the cases of $t < t^*$ and $t > t^*$.

Theorem 1. Under the homogeneous RIM with parameter $t = p - \bar{W}$, there exists a critical value $t^*$ such that the following holds almost surely as $n_k \rightarrow \infty \forall k \in \{1, 2, \ldots, K\}$ and $\frac{n_{\min}}{\max} \rightarrow c > 0$:

(a) If $t < t^*$, $Y_k = [1_{n_k}1_{K-1}^T, v_k] = [v_1^T 1_{n_k}, v_2^T 1_{n_k}, \ldots, v_{K-1}^T 1_{n_k}]$, $\forall k$; $t > t^*$, $Y_k 1_{n_k} = 0_{n_k \times (K-1)}$, $\forall k$; $t = t^*$, $Y_k 1_{n_k} = 0_{n_k \times (K-1)}$, $\forall k$; where $v_k = \text{diag}(v_1, v_2, \ldots, v_{K-1}) \in \mathbb{R}^{(K-1) \times (K-1)}$.

In particular, when $t < t^*$, $Y$ has the following properties:

(a-1) The columns of $Y_k$ are constant vectors.

(a-2) Each column of $Y$ has at least two nonzero cluster-wise constant components, and these constant signs alternate such that their weighted sum equals 0 (i.e., $\sum_k n_k v_k^T = 0$, $\forall j \in \{1, 2, \ldots, K-1\}$).

(a-3) No two columns of $Y$ have the same sign on the cluster-wise nonzero components.

Furthermore, $t^*$ satisfies:

(b) $t_{LB} \leq t^* \leq t_{UB}$, where

$$t_{LB} = \frac{\min_{k \epsilon \{1, 2, \ldots, K\}} S_{2,K}(L_k)}{(K-1)n_{\min}} \quad t_{UB} = \frac{\min_{k \epsilon \{1, 2, \ldots, K\}} S_{2,K}(L_k)}{(K-1)n_{\max}}.$$ 

In particular, $t_{LB} = t_{UB}$ when $c = 1$.

Theorem 1 (a) shows that there exists a critical value $t^*$ that separates the behavior of the rows of $Y$ into two regimes: (1) when $t < t^*$, based on conditions (a-1) to (a-3), the rows of each $Y_k$ is identical and cluster-wise distinct such that SGC can be successful. (2) when $t > t^*$, the row sum of each $Y_k$ is zero, and the incoherence of the entries in $Y_k$ make it impossible for SGC to separate the clusters. Theorem 1 (b) provides closed-form upper and lower bounds on the critical value $t^*$, and these two bounds become tight when every cluster has identical size (i.e., $c = 1$).

3.4. Phase transitions under inhomogeneous RIM

We can extend the phase transition analysis of the homogeneous RIM to the inhomogeneous RIM. Let $Y \in \mathbb{R}^{n \times (K-1)}$ be the eigenvalue matrix of $L$ under the inhomogeneous RIM, and let $\widehat{Y} \in \mathbb{R}^{n \times (K-1)}$ be the eigenvalue matrix of the graph Laplacian $L$ of another random graph, independent of $L$, generated by a homogeneous RIM with cluster interconnectivity parameter $t$. We can specify the distance between the subspaces spanned by the columns of $Y$ and $\widehat{Y}$ by inspecting their principal angles [14]. Since $Y$ and $\widehat{Y}$ both have orthonormal columns, the angle $\vartheta$ of $K-1$ principal angles between their column spaces is $\vartheta = \max_{1 \leq k \leq K-1} (\sigma_1^2(\delta_k Y^T \widehat{Y}), \ldots, \sigma_{K-1}^2(\delta_k Y^T \widehat{Y}))$, where $\sigma_k(M)$ is the $k$-th largest singular value of a real rectangular matrix $M$. Let $\Theta(Y, \widehat{Y}) = \delta(t)$, and let $\sin \Theta(Y, \widehat{Y})$ be defined entrywise. When $t < t^*$, the following theorem provides an upper bound on the Frobenius norm of $\sin \Theta(Y, \widehat{Y})$, denoted by $\|\sin \Theta(Y, \widehat{Y})\|_{F}$. Theorem 2. Under the inhomogeneous RIM with interconnection parameters $\{t_{ij} = p_{ij} \cdot \bar{W}_{ij}\}$, let $t^*$ be the critical threshold value for the homogeneous RIM specified by Theorem 1, and define $\delta_{t,n} = \min\{t_i, |\lambda_{K+1}(L_k) - t_i|\}$. For a fixed $t$, if $t < t^*$ and $\delta_{t,n} \rightarrow 0$ as $n_k \rightarrow \infty \forall k \in \{1, 2, \ldots, K\}$, the following statement holds almost surely as $n_k \rightarrow \infty \forall k$ and $\frac{n_{\min}}{\max} \rightarrow c > 0$:

$$\|\sin \Theta(Y, \widehat{Y})\|_{F} \leq \frac{\|L - \widehat{L}\|_{F}}{n_{t,n}}.$$

Furthermore, let $t_{\max} = \max_{i \neq j} t_{ij}$. If $t_{\max} < t^*$, then

$$\|\sin \Theta(Y, \widehat{Y})\|_{F} \leq \min_{t \leq t_{\max}} \frac{\|L - \widehat{L}\|_{F}}{n_{t,n}}.$$

By Theorem 1, since under the homogeneous RIM the rows of $Y$ have cluster-wise separability when $t < t^*$, Theorem 2 shows that under the inhomogeneous RIM cluster-wise separability in $Y$ can still be expected provided that the subspace distance $\|\sin \Theta(Y, \widehat{Y})\|_{F}$ is small and $t < t^*$. Moreover, if $t_{\max} < t^*$, we can obtain a tighter upper bound on $\|\sin \Theta(Y, \widehat{Y})\|_{F}$. These two theorems serve as the cornerstone of the proposed AMOS algorithm, and the proofs are given in the extended version [21].

4. AUTOMATED MODEL ORDER SELECTION (AMOS) ALGORITHM FOR SPECTRAL GRAPH CLUSTERING

Based on the theoretical framework in Sec. 3, we propose an automated model order selection (AMOS) algorithm for automated cluster assignment for SGC. The flow diagram of AMOS is displayed in
Algorithm 1 p-value computation of V-test for the RIM test

**Input:** An \( n_1 \times n_2 \) interconnection matrix \( \mathbf{C}_{ij} \)

**Output:** p-value(i, j) 

\[
\begin{align*}
\text{x} &= \mathbf{C}_{ij}1_{n_1} \quad \text{(\# of nonzero entries of each row in } \mathbf{C}_{ij}) \\
y &= n_11_{n_2} \quad \text{(# of zero entries of each row in } \mathbf{C}_{ij}) \\
X &= x^T x - x^T 1_{n_1} \\
Y &= y^T y - y^T 1_{n_2} \\
N &= n_1 n_2 (n_j - 1) \\
V &= \left( \sqrt{X} + \sqrt{Y} \right)^2.
\end{align*}
\]

Compute test statistic \( Z = \frac{V - N}{\sqrt{2N}} \)

Compute p-value(i, j) = \( 2 \cdot \min \{ \Phi(Z), 1 - \Phi(Z) \} \)

---

4.1. RIM test via p-value for local homogeneity testing

Given clusters \( \{ \mathbf{G}_k \}_{k=1}^K \) obtained from SGC with model order \( K \), let \( \mathbf{C}_{ij} \) be the \( n_i \times n_j \) interconnection matrix of between-cluster edges connecting clusters \( i \) and \( j \). The goal of local homogeneity testing is to compute a p-value to test the hypothesis that the identified clusters satisfy the RIM. More specifically, we are testing the null hypothesis that \( \mathbf{C}_{ij} \) is a realization of a random matrix with i.i.d. Bernoulli entries (RIM) and the alternative hypothesis that \( \mathbf{C}_{ij} \) is not a realization of a random matrix with i.i.d. Bernoulli entries (not RIM), for all \( i \neq j, i > j \). To compute a p-value for the RIM test we use the V-test [26] for homogeneity testing of the row sums of each interconnection matrix \( \mathbf{C}_{ij} \). Specifically, the V-test tests that the rows of \( \mathbf{C}_{ij} \) are all identically distributed. For any \( \mathbf{C}_{ij} \) the test statistic \( Z \) of the V-test converges to a standard normal distribution as \( n_i, n_j \to \infty \), and the p-value for the hypothesis that the row sums of \( \mathbf{C}_{ij} \) are i.i.d. is \( \text{p-value}(i, j) = 2 \cdot \min \{ \Phi(Z), 1 - \Phi(Z) \} \), where \( \Phi(\cdot) \) is the cumulative distribution function (cdf) of the standard normal distribution. The proposed V-test procedure is summarized in Algorithm 1. The RIM test on \( \mathbf{C}_{ij} \) rejects the null hypothesis if \( \text{p-value}(i, j) \leq \eta \), where \( \eta \) is the desired single comparison significance level. The AMOS algorithm won’t proceed to the phase transition test stage (Sec. 4.2) unless every \( \mathbf{C}_{ij} \) passes the RIM test.

---

Algorithm 2 Automated model order selection (AMOS) algorithm for spectral graph clustering (SGC)

**Input:** a connected undirected weighted graph, p-value significance level \( \eta \), RIM confidence interval parameters \( \alpha, \alpha' \)

**Output:** number of clusters \( K \) and identified clusters \( \{ \hat{G}_k \}_{k=1}^K \)

Initialization: \( K = 2 \), Flag = 1. 

while Flag = 1 do

Obtain \( K \) clusters \( \{ \hat{G}_k \}_{k=1}^K \) via spectral clustering (+) # Local homogeneity testing #

for \( i = 1 \) to \( K \) do

for \( j = i + 1 \) to \( K \) do

Calculate p-value(i, j) from Algorithm 1.

if p-value(i, j) \( \leq \eta \) then Reject RIM

Go back to (+) with \( K = K + 1 \).

end if

end for

end for

end while

Estimate \( \hat{\rho}, \hat{W}, \{ \hat{\rho}_{ij} \}, \{ \hat{W}_{ij} \} \), and \( \hat{t}_{LB} \) specified in Sec. 4.2.

# Homogeneous RIM test #

if \( \hat{\rho} \) lies within the confidence interval in (1) then # Homogeneous RIM phase transition test #

if \( \hat{\rho} \cdot \hat{W} < \hat{t}_{LB} \) then Flag = 0.

else Go back to (+) with \( K = K + 1 \).

end if

else if \( \hat{\rho} \) does not lie within the confidence interval in (1) then # Inhomogeneous RIM phase transition test #

if \( \prod_{i=1}^{K} \prod_{j=i+1}^{K} F_{ij} \left( \frac{\hat{t}_{LB}}{\hat{W}_{ij}}, \hat{\rho}_{ij} \right) \geq 1 - \alpha' \) then

Flag = 0.

else Go back to (+) with \( K = K + 1 \).

end if

end if

end while

Output \( K \) clusters \( \{ \hat{G}_k \}_{k=1}^K \).
5. EXPERIMENTS ON REAL-WORLD DATASETS

We implement the proposed AMOS algorithm on the real-world network datasets in Table 1, and compare the clustering results with three other automated graph clustering methods in terms of multiple external and internal clustering metrics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Node</th>
<th>Edge</th>
<th>Ground truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE reliability test system (RTS) [27]</td>
<td>73 power stations</td>
<td>108 power lines</td>
<td>3 power subsystems</td>
</tr>
<tr>
<td>Hibernia Internet backbone map [28]</td>
<td>55 cities</td>
<td>162 connections</td>
<td>American &amp; Europe cities</td>
</tr>
<tr>
<td>Cogent Internet backbone map [28]</td>
<td>197 cities</td>
<td>243 connections</td>
<td>American &amp; Europe cities</td>
</tr>
<tr>
<td>Minnesota road map [29]</td>
<td>2640 intersections</td>
<td>3302 roads</td>
<td>None</td>
</tr>
<tr>
<td>Facebook [30]</td>
<td>4039 users</td>
<td>88234 friendships</td>
<td>None</td>
</tr>
</tbody>
</table>

**Table 1:** Summary of real-world datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>NMI</th>
<th>RI</th>
<th>F</th>
<th>C</th>
<th>NC</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEE RTS (3)</td>
<td>AMOS (3)</td>
<td>.89</td>
<td>.96</td>
<td>.94</td>
<td>.046</td>
<td>.068</td>
</tr>
<tr>
<td></td>
<td>Louvain (6)</td>
<td>.74</td>
<td>.84</td>
<td>.67</td>
<td>.144</td>
<td>.169</td>
</tr>
<tr>
<td></td>
<td>NB (3)</td>
<td>.75</td>
<td>.88</td>
<td>.61</td>
<td>.070</td>
<td>.100</td>
</tr>
<tr>
<td></td>
<td>ST (2)</td>
<td>.74</td>
<td>.78</td>
<td>.75</td>
<td>.021</td>
<td>.041</td>
</tr>
<tr>
<td>Hibernia (2)</td>
<td>AMOS (2)</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.030</td>
<td>.057</td>
</tr>
<tr>
<td></td>
<td>Louvain (6)</td>
<td>.27</td>
<td>.51</td>
<td>.33</td>
<td>.222</td>
<td>.263</td>
</tr>
<tr>
<td></td>
<td>NB (2)</td>
<td>.73</td>
<td>.89</td>
<td>.90</td>
<td>.027</td>
<td>.053</td>
</tr>
<tr>
<td></td>
<td>ST (2)</td>
<td>.88</td>
<td>.96</td>
<td>.97</td>
<td>.028</td>
<td>.050</td>
</tr>
<tr>
<td>Cogent (2)</td>
<td>AMOS (4)</td>
<td>.42</td>
<td>.63</td>
<td>.53</td>
<td>.036</td>
<td>.049</td>
</tr>
<tr>
<td></td>
<td>Louvain (11)</td>
<td>.25</td>
<td>.54</td>
<td>.26</td>
<td>.186</td>
<td>.204</td>
</tr>
<tr>
<td></td>
<td>NB (3)</td>
<td>.26</td>
<td>.54</td>
<td>.58</td>
<td>.073</td>
<td>.109</td>
</tr>
<tr>
<td></td>
<td>ST (14)</td>
<td>.34</td>
<td>.55</td>
<td>.29</td>
<td>.148</td>
<td>.164</td>
</tr>
<tr>
<td>Minnesota (-)</td>
<td>AMOS (46)</td>
<td>.074</td>
<td>.076</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Louvain (33)</td>
<td>.290</td>
<td>.299</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NB (35)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.140</td>
<td>.144</td>
</tr>
<tr>
<td></td>
<td>ST (100)</td>
<td>.119</td>
<td>.120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facebook (-)</td>
<td>AMOS (5)</td>
<td>.004</td>
<td>.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Louvain (17)</td>
<td>.076</td>
<td>.079</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>NB (55)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>.478</td>
<td>.486</td>
</tr>
<tr>
<td></td>
<td>ST (7)</td>
<td>.006</td>
<td>.007</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2:** Performance comparison of automated graph clustering algorithms. The number in the parenthesis of the Dataset (Method) column shows the number of ground-truth (identified) clusters. “-” means not available due to lack of ground-truth cluster information. For each metric, the best method is highlighted in bold face. For each dataset, AMOS has the most clustering metrics of best performance.

6. CONCLUSION

This paper presents an automated model order selection (AMOS) algorithm for spectral graph clustering (SGC). Stemming from the phase transition analysis on the clustering reliability of SGC under the random interconnection model, AMOS performs iterative SGC and multi-stage statistical tests such that it automatically finds the minimal number of clusters with statistical clustering reliability guarantees. Experiments on real-world datasets show that AMOS outperforms other three automated graph clustering methods in terms of multiple external and internal clustering metrics.
7. REFERENCES


