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.

This work was partially supported by Army Research Office grant W911NF-15-1-0479 and the Consortium for Verification Technology under Department of Energy National Nuclear Security Administration award number DE-NA0002534.

2. RELATED WORK

Most existing model selection algorithms specify an upper bound $K_{\max}$ on the number K of clusters and then select K based on optimizing some objective function, e.g., the goodness of fit of the k-cluster model for $k = 2, \ldots, K_{\max}$. 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 $\mathbf{A}$, where n is the number of nodes in the graph. $[\mathbf{A}]_{uv} = 1$ if there exists an edge between the node pair (u, v), and otherwise $[\mathbf{A}]_{uv} = 0$. An unweighted undirected graph is completely specified by its adjacency matrix $\mathbf{A}$, while a weighted undirected graph is specified by a nonnegative matrix $\mathbf{W}$, 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 $n_{\max}$ and $n_{\min}$, respectively. Let $\mathbf{A}_k$ denote the $n_k \times n_k$ adjacency matrix representing the internal edge connections in cluster k and let $\mathbf{C}_{ij}$ $(i, j \in \{1, 2, \ldots, K\}, i \neq j)$ be an $n_i \times n_j$ matrix representing the adjacency matrix of inter-cluster edge connections between the cluster pair (i, j). The matrix $\mathbf{A}_k$ is symmetric and $\mathbf{C}_{ij} = \mathbf{C}_{ji}^T$ for all $i \neq j$.

The random interconnection model (RIM) [21] assumes that: (1) the adjacency matrix $\mathbf{A}_k$ is associated with a connected graph of n_k nodes but is otherwise arbitrary; (2) the $K(K-1)/2$ matrices

$\{\mathbf{C}_{ij}\}_{i>j}$ are random mutually independent, and each $\mathbf{C}_{ij}$ 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 $\overline{W}_{ij}$ 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 $\overline{W}_{ij} = \overline{W}$ 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 $\mathbf{L} = \mathbf{S} - \mathbf{W}$, where $\mathbf{S} = \text{diag}(\mathbf{W}\mathbf{1}_n)$ is a diagonal matrix and $\mathbf{1}_n(\mathbf{0}_n)$ 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 $\mathbf{L}_k$. We also denote the i-th smallest eigenvalue of $\mathbf{L}$ by $\lambda_i(\mathbf{L})$ and define the partial eigenvalue sum $S_{2:K}(\mathbf{L}) = \sum_{i=2}^{K} \lambda_i(\mathbf{L})$. To partition the nodes in the graph into K ($K \geq 2$) clusters, spectral clustering [14] uses the K eigenvectors $\{\mathbf{u}_k\}_{k=1}^{K}$ associated with the K smallest eigenvalues of $\mathbf{L}$. 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 $\mathbf{L}$, the smallest eigenvector $\mathbf{u}_1$ is a constant vector and $\lambda_i(\mathbf{L}) > 0 \; \forall \; i \geq 2$. As a result, for connected undirected graphs, it suffices to use the $K-1$ eigenvectors $\{\mathbf{u}_k\}_{k=2}^{K}$ of $\mathbf{L}$ for SGC. In particular, these $K-1$ eigenvectors are represented by the columns of the eigenvector matrix $\mathbf{Y} = [\mathbf{u}_2, \mathbf{u}_3, \ldots, \mathbf{u}_K] \in \mathbb{R}^{n \times (K-1)}$.

3.3. Phase transitions under homogeneous RIM

Let $\mathbf{Y} = [\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_K^T]^T$ be the cluster partitioned eigenvector matrix associated with $\mathbf{L}$ for SGC, where $\mathbf{Y}_k \in \mathbb{R}^{n_k \times (K-1)}$ with its rows indexing the nodes in cluster k. Under the homogeneous RIM, let $t = p \cdot \overline{W}$ 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 $\mathbf{Y}$ for the cases of $t < t^*$ and $t > t^*$.

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

(a)
$$\begin{cases} \text{If } t < t^*, \; \mathbf{Y}_k = \mathbf{1}_{n_k}\mathbf{1}_{K-1}^T\mathbf{V}_k \\ \qquad\qquad = \left[v_1^k\mathbf{1}_{n_k}, v_2^k\mathbf{1}_{n_k}, \ldots, v_{K-1}^k\mathbf{1}_{n_k}\right], \; \forall \; k; \\ \text{If } t > t^*, \; \mathbf{Y}_k^T\mathbf{1}_{n_k} = \mathbf{0}_{K-1}, \; \forall \; k; \\ \text{If } t = t^*, \; \mathbf{Y}_k = \mathbf{1}_{n_k}\mathbf{1}_{K-1}^T\mathbf{V}_k \text{ or } \mathbf{Y}_k^T\mathbf{1}_{n_k} = \mathbf{0}_{K-1}, \; \forall \; k, \end{cases}$$

where $\mathbf{V}_k = \text{diag}(v_1^k, v_2^k, \ldots, v_{K-1}^k) \in \mathbb{R}^{(K-1)\times(K-1)}$.
In particular, when $t < t^*$, $\mathbf{Y}$ has the following properties:
(a-1) *The columns of $\mathbf{Y}_k$ are constant vectors.*
(a-2) *Each column of $\mathbf{Y}$ has at least two nonzero cluster-wise constant components, and these constants have alternating signs such that their weighted sum equals 0 (i.e., $\sum_k n_k v_j^k = 0, \; \forall \; j \in \{1, 2, \ldots, K-1\}$).*
(a-3) *No two columns of $\mathbf{Y}$ have the same sign on the cluster-wise nonzero components.*
Furthermore, t^* satisfies:
(b) $t_{LB} \leq t^* \leq t_{UB}$, *where*

$$t_{LB} = \frac{\min_{k\in\{1,2,\ldots,K\}} S_{2:K}(\mathbf{L}_k)}{(K-1)n_{\max}}; \;\; t_{UB} = \frac{\min_{k\in\{1,2,\ldots,K\}} S_{2:K}(\mathbf{L}_k)}{(K-1)n_{\min}}.$$

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 $\mathbf{Y}$ into two regimes: (1) when $t < t^*$, based on conditions (a-1) to (a-3), the rows of each $\mathbf{Y}_k$ is identical and cluster-wise distinct such that SGC can be successful. (2) when $t > t^*$, the row sum of each $\mathbf{Y}_k$ is zero, and the incoherence of the entries in $\mathbf{Y}_k$ 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 $\mathbf{Y} \in \mathbb{R}^{n\times(K-1)}$ be the eigenvector matrix of $\mathbf{L}$ under the inhomogeneous RIM, and let $\widetilde{\mathbf{Y}} \in \mathbb{R}^{n\times(K-1)}$ be the eigenvector matrix of the graph Laplacian $\widetilde{\mathbf{L}}$ of another random graph, independent of $\mathbf{L}$, generated by a homogeneous RIM with cluster interconnectivity parameter t. We can specify the distance between the subspaces spanned by the columns of $\mathbf{Y}$ and $\widetilde{\mathbf{Y}}$ by inspecting their principal angles [14]. Since $\mathbf{Y}$ and $\widetilde{\mathbf{Y}}$ both have orthonormal columns, the vector $\mathbf{v}$ of $K-1$ principal angles between their column spaces is $\mathbf{v} = [\cos^{-1}\sigma_1(\mathbf{Y}^T\widetilde{\mathbf{Y}}), \ldots, \cos^{-1}\sigma_{K-1}(\mathbf{Y}^T\widetilde{\mathbf{Y}})]^T$, where $\sigma_k(\mathbf{M})$ is the k-th largest singular value of a real rectangular matrix $\mathbf{M}$. Let $\mathbf{\Theta}(\mathbf{Y}, \widetilde{\mathbf{Y}}) = \text{diag}(\mathbf{v})$, and let $\sin\mathbf{\Theta}(\mathbf{Y}, \widetilde{\mathbf{Y}})$ be defined entrywise. When $t < t^*$, the following theorem provides an upper bound on the Frobenius norm of $\sin\mathbf{\Theta}(\mathbf{Y}, \widetilde{\mathbf{Y}})$, denoted by $\|\sin\mathbf{\Theta}(\mathbf{Y}, \widetilde{\mathbf{Y}})\|_F$.

Theorem 2. *Under the inhomogeneous RIM with interconnection parameters $\{t_{ij} = p_{ij} \cdot \overline{W}_{ij}\}$, let t^* be the critical threshold value for the homogeneous RIM specified by Theorem 1, and define $\delta_{t,n} = \min\{t, |\lambda_{K+1}(\frac{\mathbf{L}}{n}) - t|\}$. For a fixed t, if $t < t^*$ and $\delta_{t,n} \to \delta_t > 0$ as $n_k \to \infty \; \forall \; k \in \{1, 2, \ldots, K\}$, the following statement holds almost surely as $n_k \to \infty \; \forall \; k$ and $\frac{n_{\min}}{n_{\max}} \to c > 0$:*

$$\|\sin\mathbf{\Theta}(\mathbf{Y}, \widetilde{\mathbf{Y}})\|_F \leq \frac{\|\mathbf{L} - \widetilde{\mathbf{L}}\|_F}{n\delta_t}.$$

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

$$\|\sin\mathbf{\Theta}(\mathbf{Y}, \widetilde{\mathbf{Y}})\|_F \leq \min_{t \leq t_{\max}} \frac{\|\mathbf{L} - \widetilde{\mathbf{L}}\|_F}{n\delta_t}.$$

By Theorem 1, since under the homogeneous RIM the rows of $\widetilde{\mathbf{Y}}$ has cluster-wise separability when $t < t^*$, Theorem 2 shows that under the inhomogeneous RIM cluster-wise separability in $\mathbf{Y}$ can still be expected provided that the subspace distance $\|\sin\mathbf{\Theta}(\mathbf{Y}, \widetilde{\mathbf{Y}})\|_F$ is small and $t < t^*$. Moreover, if $t_{\max} < t^*$, we can obtain a tighter upper bound on $\|\sin\mathbf{\Theta}(\mathbf{Y}, \widetilde{\mathbf{Y}})\|_F$. 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

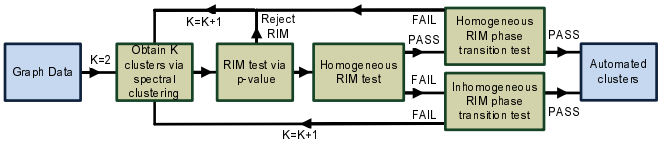

Fig. 1: Flow diagram of the proposed automated model order selection (AMOS) algorithm for spectral graph cluster (SGC).

Algorithm 1 p-value computation of V-test for the RIM test

Input: An $n_i \times n_j$ interconnection matrix $\widehat{\mathbf{C}}_{ij}$
Output: p-value(i,j)
$\mathbf{x} = \widehat{\mathbf{C}}_{ij}\mathbf{1}_{n_j}$ (# of nonzero entries of each row in $\widehat{\mathbf{C}}_{ij}$)
$\mathbf{y} = n_j\mathbf{1}_{n_i} - \mathbf{x}$ (# of zero entries of each row in $\widehat{\mathbf{C}}_{ij}$)
$X = \mathbf{x}^T\mathbf{x} - \mathbf{x}^T\mathbf{1}_{n_i}$ and $Y = \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{1}_{n_i}$.
$N = n_i n_j(n_j-1)$ and $V = \left(\sqrt{X} + \sqrt{Y}\right)^2$.
Compute test statistic $Z = \frac{V-N}{\sqrt{2N}}$
Compute p-value(i,j)= $2 \cdot \min\{\Phi(Z), 1-\Phi(Z)\}$

Fig. 1, and the algorithm is summarized in Algorithm 2. The AMOS codes can be downloaded from https://github.com/tgensol/AMOS.

AMOS works by iteratively increasing the number of clusters K and performing multi-stage statistical clustering reliability tests until the identified clusters are deemed reliable. The statistical tests in AMOS are implemented in two phases. The first phase is to test the RIM assumption based on the interconnectivity pattern of each cluster (Sec. 4.1), and the second phase is to test the homogeneity and variation of the interconnectivity parameter p_{ij} for every cluster pair i and j in addition to making comparisons to the critical phase transition threshold (Sec. 4.2). The proofs of the established statistical clustering reliability tests are given in the extended version [21].

The input graph data of AMOS is a matrix representing a connected undirected weighted graph. For each iteration in K, SGC is implemented to produce K clusters $\{\widehat{G}_k\}_{k=1}^K$, where $\widehat{G}_k$ is the k-th identified cluster with number of nodes $\widehat{n}_k$ and number of edges $\widehat{m}_k$.

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

Given clusters $\{\widehat{G}_k\}_{k=1}^K$ obtained from SGC with model order K, let $\widehat{\mathbf{C}}_{ij}$ be the $\widehat{n}_i \times \widehat{n}_j$ 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 $\widehat{\mathbf{C}}_{ij}$ *is a realization of a random matrix with i.i.d. Bernoulli entries* (RIM) and the alternative hypothesis that $\widehat{\mathbf{C}}_{ij}$ *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 $\widehat{\mathbf{C}}_{ij}$. Specifically, the V-test tests that the rows of $\widehat{\mathbf{C}}_{ij}$ are all identically distributed. For any $\widehat{\mathbf{C}}_{ij}$ 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 $\widehat{\mathbf{C}}_{ij}$ are i.i.d. is 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 $\widehat{\mathbf{C}}_{ij}$ rejects the null hypothesis if p-value$(i,j) \leq \eta$, where η is the desired single comparison significance level. The AMOS algorithm won't proceed to the phase transition test stage (Sec. 4.2) unless every $\widehat{\mathbf{C}}_{ij}$ 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 η, RIM confidence interval parameters α, α'
Output: number of clusters K and identified clusters $\{\widehat{G}_k\}_{k=1}^K$
Initialization: $K = 2$. Flag $= 1$.
while Flag$= 1$ **do**
  Obtain K clusters $\{\widehat{G}_k\}_{k=1}^K$ 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
  Estimate $\widehat{p}$, $\widehat{\overline{W}}$, $\{\widehat{p}_{ij}\}$, $\{\widehat{\overline{W}}_{ij}\}$, and $\widehat{t}_{\rm LB}$ specified in Sec. 4.2.
  # Homogeneous RIM test #
  if $\widehat{p}$ lies within the confidence interval in (1) **then**
    # Homogeneous RIM phase transition test #
    if $\widehat{p} \cdot \widehat{\overline{W}} < \widehat{t}_{\rm LB}$ **then** Flag$= 0$.
    else Go back to $(*)$ with $K = K+1$.
    end if
  else if $\widehat{p}$ does not lie within the confidence interval in (1) **then**
    # Inhomogeneous RIM phase transition test #
    if $\prod_{i=1}^K \prod_{j=i+1}^K F_{ij}\left(\frac{\widehat{t}_{\rm LB}}{\widehat{\overline{W}}_{ij}}, \widehat{p}_{ij}\right) \geq 1-\alpha'$ **then**
      Flag$= 0$.
    else Go back to $(*)$ with $K = K+1$.
    end if
  end if
end while
Output K clusters $\{\widehat{G}_k\}_{k=1}^K$.

4.2. Phase transition tests

Once the identified clusters $\{\widehat{G}_k\}_{k=1}^K$ pass the RIM test, one can empirically determine the reliability of the clustering results using the phase transition analysis in Sec. 3. AMOS first tests the assumption of homogeneous RIM, and performs the *homogeneous RIM phase transition test* by comparing the empirical estimate $\widehat{t}$ of the interconnectivity parameter t with the empirical estimate $\widehat{t}_{\rm LB}$ of the lower bound $t_{\rm LB}$ on t^* based on Theorem 1. If the test on the assumption of homogeneous RIM fails, AMOS then performs the *inhomogeneous RIM phase transition test* by comparing the empirical estimate $\widehat{t}_{\max}$ of $t_{\max}$ with $\widehat{t}_{\rm LB}$ based on Theorem 2.

• **Homogeneous RIM test**: The homogeneous RIM test is summarized as follows. Given clusters $\{\widehat{G}_k\}_{k=1}^K$, we estimate the interconnectivity parameters $\{\widehat{p}_{ij}\}$ by $\widehat{p}_{ij} = \frac{\widehat{m}_{ij}}{\widehat{n}_i\widehat{n}_j}$, where $\widehat{m}_{ij}$ is the number of inter-cluster edges between clusters i and j, and $\widehat{p}_{ij}$ is the maximum likelihood estimator (MLE) of p_{ij}. Under the homogeneous RIM, the estimate of the parameter p is $\widehat{p} = \frac{2(m - \sum_{k=1}^K \widehat{m}_k)}{n^2 - \sum_{k=1}^K \widehat{n}_k^2}$, where $\widehat{m}_k$ is the number of within-cluster edges of cluster k and m is the total number of edges in the graph. A generalized log-likelihood ratio test (GLRT) is used to test the validity of the homogeneous RIM. By the Wilk's theorem [31], an asymptotic $100(1-\alpha)\%$ confidence

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

Table 1: Summary of real-world datasets.

interval for p in an assumed homogeneous RIM is

$$\left\{ p : \xi_{\binom{K}{2}-1,1-\frac{\alpha}{2}} \leq 2\sum_{i=1}^{K}\sum_{j=i+1}^{K} \mathbb{I}_{\{\widehat{p}_{ij}\in(0,1)\}} \left[\widehat{m}_{ij}\ln\widehat{p}_{ij} \right. \right. \quad (1)$$

$$\left. +(\widehat{n}_i\widehat{n}_j - \widehat{m}_{ij})\ln(1-\widehat{p}_{ij})\right] - 2\left(m - \sum_{k=1}^{K}\widehat{m}_k\right)\ln p$$

$$\left. - \left[n^2 - \sum_{k=1}^{K}\widehat{n}_k^2 - 2\left(m-\sum_{k=1}^{K}\widehat{m}_k\right)\right]\ln(1-p) \leq \xi_{\binom{K}{2}-1,\frac{\alpha}{2}} \right\},$$

where $\xi_{q,\alpha}$ is the upper α-th quantile of the central chi-square distribution with degree of freedom q. The clusters pass the homogeneous RIM test if $\widehat{p}$ is within the confidence interval specified in (1).

• **Homogeneous RIM phase transition test**: By Theorem 1, if the identified clusters follow the homogeneous RIM, then they are deemed reliable when $\widehat{t} < \widehat{t}_{\text{LB}}$, where $\widehat{t} = \widehat{p}\cdot\widehat{\widehat{W}}$, $\widehat{\widehat{W}}$ is the average of all between-cluster edge weights, and $\widehat{t}_{\text{LB}} = \frac{\min_{k\in\{1,2,\ldots,K\}} S_{2:K}(\widehat{\mathbf{L}}_k)}{(K-1)\widehat{n}_{\max}}$.

• **Inhomogeneous RIM phase transition test**: If the clusters fail the homogeneous RIM test, we then use the maximum of MLEs of t_{ij}'s, denoted by $\widehat{t}_{\max} = \max_{i>j}\widehat{t}_{ij}$, as a test statistic for testing the null hypothesis H_0: $\widehat{t}_{\max} < t_{LB}$ against the alternative hypothesis H_1: $\widehat{t}_{\max} \geq t_{LB}$. The test accepts H_0 if $\widehat{t}_{\max} < t_{\text{LB}}$ and hence by Theorem 2 the identified clusters are deemed reliable. Using the Anscombe transformation on the $\widehat{p}_{ij}$'s for variance stabilization [32], let $A_{ij}(x) = \sin^{-1}\sqrt{\frac{x+\frac{c'}{\widehat{n}_i\widehat{n}_j}}{1+\frac{2c'}{\widehat{n}_i\widehat{n}_j}}}$, where $c' = \frac{3}{8}$. Under the null hypothesis that $\widehat{t}_{\max} < t_{\text{LB}}$, from [33, Theorem 2.1], an asymptotic $100(1-\alpha')\%$ confidence interval for $\widehat{t}_{\max}$ is $[0,\psi]$, where $\psi(\alpha',\{\widehat{t}_{ij}\})$ is a function of the precision parameter $\alpha' \in [0,1]$ and $\{\widehat{t}_{ij}\}$. Furthermore, it can be shown that verifying $\psi < \widehat{t}_{\text{LB}}$ is equivalent to checking the condition

$$\prod_{i=1}^{K}\prod_{j=i+1}^{K} F_{ij}\left(\frac{\widehat{t}_{\text{LB}}}{\widehat{\widehat{W}}_{ij}}, \widehat{p}_{ij}\right) \geq 1-\alpha', \quad (2)$$

where $F_{ij}(x,\widehat{p}_{ij}) = \Phi\left(\sqrt{4\widehat{n}_i\widehat{n}_j+2}\cdot(A_{ij}(x)-A_{ij}(\widehat{p}_{ij}))\right)\cdot \mathbb{I}_{\{\widehat{p}_{ij}\in(0,1)\}} + \mathbb{I}_{\{\widehat{p}_{ij}<x\}}\mathbb{I}_{\{\widehat{p}_{ij}\in\{0,1\}\}}$, and $\mathbb{I}$ is the indicator function.

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, including the self-tuning method (ST) [13], the nonbacktracking matrix method (NB) [24, 25], and the Louvain method [23]. For AMOS, we use the degree normalized adjacency matrix [14] as the input graph data, and set $\alpha = \alpha' = 0.05$ and $\eta = 10^{-5}$. For performance evaluation, multiple clustering metrics are computed for assessing the clustering quality. These metrics are normalized mutual information (NMI) [34], Rand index (RI) [34], F-measure (F) [34], conductance (C) [15], and normalized cut (NC) [15]. For NMI, RI, and F, higher value means better clustering performance, whereas for C and NC, lower value means better clustering performance.

Dataset	Method	NMI	RI	F	C	NC
IEEE RTS (3)	AMOS (3)	**.89**	**.96**	**.94**	.046	.068
	Louvain (6)	.74	.84	.67	.144	.169
	NB (3)	.75	.88	.81	.070	.100
	ST (2)	.74	.78	.75	**.021**	**.041**
Hibernia (2)	AMOS (2)	**1.0**	**1.0**	**1.0**	.030	.057
	Louvain (6)	.27	.51	.33	.222	.263
	NB (2)	.73	.89	.90	**.027**	.053
	ST (2)	.88	.96	.97	.028	**.050**
Cogent (2)	AMOS (4)	**.42**	**.63**	.53	**.036**	**.049**
	Louvain (11)	.25	.54	.26	.186	.204
	NB (3)	.26	.54	**.58**	.073	.109
	ST (14)	.34	.55	.29	.148	.164
Minnesota (-)	AMOS (46)	-	-	-	**.074**	**.076**
	Louvain (33)				.290	.299
	NB (35)				.140	.144
	ST (100)				.119	.120
Facebook (-)	AMOS (5)	-	-	-	**.004**	**.004**
	Louvain (17)				.076	.079
	NB (55)				.478	.486
	ST (7)				.006	.007

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.

Table 2 summarizes the clustering performance of the datasets in Table 1. For each dataset, AMOS has the most clustering metrics of best performance among these four methods, which demonstrates the robustness and reliability of AMOS. In particular, for the datasets with ground-truth cluster information such that the external clustering metrics NMI, RI, and F can be computed, AMOS shows significant improvement over other methods. In addition, for clustering metrics over which AMOS does not prevail, its performance is comparable to the best method.

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

[1] S. White and P. Smyth, "A spectral clustering approach to finding communities in graph." in *SIAM International Conference on Data Mining (SDM)*, vol. 5, 2005, pp. 76–84.

[2] P.-Y. Chen and A. Hero, "Phase transitions in spectral community detection," *IEEE Trans. Signal Process.*, vol. 63, no. 16, pp. 4339–4347, Aug 2015.

[3] A. Bertrand and M. Moonen, "Seeing the bigger picture: How nodes can learn their place within a complex ad hoc network topology," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 71–82, 2013.

[4] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[5] B. A. Miller, N. T. Bliss, P. J. Wolfe, and M. S. Beard, "Detection theory for graphs," *Lincoln Laboratory Journal*, vol. 20, no. 1, pp. 10–30, 2013.

[6] X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov, "Clustering with multi-layer graphs: A spectral perspective," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5820–5831, 2012.

[7] B. Oselio, A. Kulesza, and A. O. Hero, "Multi-layer graph analysis for dynamic social networks," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 514–523, Aug 2014.

[8] K. S. Xu and A. O. Hero, "Dynamic stochastic blockmodels for time-evolving social networks," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 4, pp. 552–562, 2014.

[9] S. Chen, A. Sandryhaila, J. Moura, and J. Kovacevic, "Signal recovery on graphs: Variation minimization," *IEEE Trans. Signal Process.*, vol. 63, no. 17, pp. 4609–4624, Sept. 2015.

[10] A. Sandryhaila and J. M. Moura, "Big data analysis with signal processing on graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Process. Mag.*, vol. 31, no. 5, pp. 80–90, 2014.

[11] X. Wang, P. Liu, and Y. Gu, "Local-set-based graph signal reconstruction," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2432–2444, May 2015.

[12] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems (NIPS)*, 2002, pp. 849–856.

[13] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in neural information processing systems (NIPS)*, 2004, pp. 1601–1608.

[14] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[15] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[16] S. Yu, R. Gross, and J. Shi, "Concurrent object segmentation and recognition with graph partitioning," in *Advances in neural information processing systems (NIPS)*, 2002, pp. 1383–1390.

[17] P.-Y. Chen and A. O. Hero, "Assessing and safeguarding network resilience to nodal attacks," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 138–143, Nov. 2014.

[18] R. Merris, "Laplacian matrices of graphs: a survey," *Linear Algebra and its Applications*, vol. 197-198, pp. 143–176, 1994.

[19] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clusterin algorithm," *Applied statistics*, pp. 100–108, 1979.

[20] M. Polito and P. Perona, "Grouping and dimensionality reduction by locally linear embedding," in *Advances in neural information processing systems (NIPS)*, 2001.

[21] P.-Y. Chen and A. O. Hero, "Phase transitions and a model order selection criterion for spectral graph clustering," *arXiv preprint arXiv:1604.03159*, 2016.

[22] M. E. J. Newman, "Modularity and community structure in networks," *Proc. National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[23] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, no. 10, 2008.

[24] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborova, and P. Zhang, "Spectral redemption in clustering sparse networks," *Proc. National Academy of Sciences*, vol. 110, pp. 20 935–20 940, 2013.

[25] A. Saade, F. Krzakala, M. Lelarge, and L. Zdeborova, "Spectral detection in the censored block model," *arXiv:1502.00163*, 2015.

[26] R. F. Potthoff and M. Whittinghill, "Testing for homogeneity: I. the binomial and multinomial distributions," *Biometrika*, vol. 53, no. 1-2, pp. 167–182, 1966.

[27] C. Grigg, P. Wong, P. Albrecht, R. Allan, M. Bhavaraju, R. Billinton, Q. Chen, C. Fong, S. Haddad, S. Kuruganty, W. Li, R. Mukerji, D. Patton, N. Rau, D. Reppen, A. Schneider, M. Shahidehpour, and C. Singh, "The IEEE reliability test system-1996. a report prepared by the reliability test system task force of the application of probability methods subcommittee," *IEEE Trans. Power Syst.*, vol. 14, no. 3, pp. 1010–1020, 1999.

[28] S. Knight, H. Nguyen, N. Falkner, R. Bowden, and M. Roughan, "The Internet topology zoo," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 9, pp. 1765–1775, Oct. 2011. [Online]. Available: http://www.topology-zoo.org/dataset.html

[29] D. Gleich, "MatlabBGL: A matlab graph library," *https://www.cs.purdue.edu/homes/dgleich*, 2008.

[30] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 2012, 2012, pp. 548–56.

[31] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[32] F. J. Anscombe, "The transformation of poisson, binomial and negative-binomial data," *Biometrika*, vol. 35, no. 3/4, pp. 246–254, 1948.

[33] Y.-P. Chang and W.-T. Huang, "Generalized confidence intervals for the largest value of some functions of parameters under normality," *Statistica Sinica*, pp. 1369–1383, 2000.

[34] M. J. Zaki and W. Meira Jr, *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, 2014.