SPECTRAL EMBEDDING OF REGULARIZED BLOCK MODELS
Nathan de Lara, Thomas Bonald

To cite this version:
Nathan de Lara, Thomas Bonald. SPECTRAL EMBEDDING OF REGULARIZED BLOCK MODELS. ICLR, 2020, Addis Abeba, Ethiopia. hal-02420827

HAL Id: hal-02420827
https://hal.telecom-paris.fr/hal-02420827
Submitted on 20 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
SPECTRAL EMBEDDING OF REGULARIZED BLOCK MODELS

Nathan De Lara & Thomas Bonald
Institut Polytechnique de Paris
Paris, France
{nathan.delara, thomas.bonald}@telecom-paris.fr

ABSTRACT

Spectral embedding is a popular technique for the representation of graph data. Several regularization techniques have been proposed to improve the quality of the embedding with respect to downstream tasks like clustering. In this paper, we explain on a simple block model the impact of the complete graph regularization, whereby a constant is added to all entries of the adjacency matrix. Specifically, we show that the regularization forces the spectral embedding to focus on the largest blocks, making the representation less sensitive to noise or outliers. We illustrate these results on both on both synthetic and real data, showing how regularization improves standard clustering scores.

1 INTRODUCTION

Spectral embedding is a standard technique for the representation of graph data. Given the adjacency matrix $A \in \mathbb{R}^{n \times n}$ of the graph, it is obtained by solving either the eigenvalue problem:

$$LX = X\Lambda, \quad \text{with} \quad X^T X = I, \quad (1)$$

or the generalized eigenvalue problem:

$$LX = DX\Lambda, \quad \text{with} \quad X^T DX = I, \quad (2)$$

where $D = \text{diag}(A1_n)$ is the degree matrix, with $1_n$ the all-ones vector of dimension $n$, $L = D - A$ is the Laplacian matrix of the graph, $\Lambda \in \mathbb{R}^{k \times k}$ is the diagonal matrix of the $k$ smallest (generalized) eigenvalues of $L$ and $X \in \mathbb{R}^{n \times k}$ is the corresponding matrix of (generalized) eigenvectors. In this paper, we only consider the generalized eigenvalue problem, whose solution is given by the spectral decomposition of the normalized Laplacian matrix $L_{\text{norm}} = I - D^{-1/2} AD^{-1/2}$ (Luxburg, 2007).

The spectral embedding can be interpreted as equilibrium states of some physical systems (Snell & Doyle, 2000; Spielman, 2007; Bonald et al., 2018), a desirable property in modern machine learning. However, it tends to produce poor results on real datasets if applied directly on the graph (Amini et al., 2013). One reason is that real graphs are most often disconnected due to noise or outliers in the dataset.

In order to improve the quality of the embedding, two main types of regularization have been proposed. The first artificially increases the degree of each node by a constant factor (Chaudhuri et al., 2012; Qin & Rohe, 2013), while the second adds a constant to all entries of the original adjacency matrix (Amini et al., 2013; Joseph et al., 2016; Zhang & Rohe, 2018). In the practically interesting case where the original adjacency matrix $A$ is sparse, the regularized adjacency matrix is dense but has a so-called sparse + low rank structure, enabling the computation of the spectral embedding on very large graphs (Lara, 2019).

While (Zhang & Rohe, 2018) explains the effects of regularization through graph conductance and (Joseph et al., 2016) through eigenvector perturbation on the Stochastic Block Model, there is no simple interpretation of the benefits of graph regularization. In this paper, we show on a simple block model that the complete graph regularization forces the spectral embedding to separate the blocks in decreasing order of size, making the embedding less sensitive to noise or outliers in the data.
Indeed, (Zhang & Rohe 2018) identified that, without regularization, the cuts corresponding to the first dimensions of the spectral embedding tend to separate small sets of nodes, so-called dangling sets, loosely connected to the rest of the graph. Our work shows more explicitly that regularization forces the spectral embedding to focus on the largest clusters. Moreover, our analysis involves some explicit characterization of the eigenvalues, allowing us to quantify the impact of the regularization parameter.

The rest of this paper is organized as follows. Section 2 presents block models and an important preliminary result about their aggregation. Section 3 presents the main result of the paper, about the regularization of block models, while Section 4 extends this result to bipartite graphs. Section 5 presents the experiments and Section 6 concludes the paper.

2 Aggregation of Block Models

Let \( A \in \mathbb{R}^{n \times n} \) be the adjacency matrix of an undirected, weight graph, that is a symmetric matrix such that \( A_{ij} > 0 \) if and only if there is an edge between nodes \( i \) and \( j \), with weight \( A_{ij} \). Assume that the \( n \) nodes of the graph can be partitioned into \( K \) blocks of respective sizes \( n_1, \ldots, n_K \) so that any two nodes of the same block have the same neighborhood, i.e., the corresponding rows (or columns) of \( A \) are the same. Without any loss of generality, we assume that the matrix \( A \) has rank \( K \). We refer to such a graph as a block model.

Let \( Z \in \mathbb{R}^{n \times K} \) be the associated membership matrix, with \( Z_{ij} = 1 \) if index \( i \) belongs to block \( j \) and 0 otherwise. We denote by \( W = Z^T Z \in \mathbb{R}^{K \times K} \) the diagonal matrix of block sizes.

Now define \( \bar{A} = Z^T A Z \in \mathbb{R}^{K \times K} \). This is the adjacency matrix of the aggregate graph, where each block of the initial graph is replaced by a single node; two nodes in this graph are connected by an edge of weight equal to the total weight of edges between the corresponding blocks in the original graph. We denote by \( \bar{D} = \text{diag}(\bar{A}1_K) \) the degree matrix and by \( \bar{L} = \bar{D} - \bar{A} \) the Laplacian matrix of the aggregate graph.

The following result shows that the solution to the generalized eigenvalue problem (2) follows from that of the aggregate graph:

**Proposition 1.** Let \( x \) be a solution to the generalized eigenvalue problem:

\[
Lx = \lambda Dx.
\]  

Then either \( Z^T x = 0 \) and \( \lambda = 1 \) or \( x = Z y \) where \( y \) is a solution to the generalized eigenvalue problem:

\[
\bar{L} y = \lambda \bar{D} y.
\]

**Proof.** Consider the following reformulation of the generalized eigenvalue problem (3):

\[
Ax = D x (1 - \lambda).
\]

Since the rank of \( A \) is equal to \( K \), there are \( n - K \) eigenvectors \( x \) associated with the eigenvalue \( \lambda = 1 \), each satisfying \( Z^T x = 0 \). By orthogonality, the other eigenvectors satisfy \( x = Z y \) for some vector \( y \in \mathbb{R}^K \). We get:

\[
A Z y = D Z y (1 - \lambda),
\]

so that

\[
\bar{A} y = \bar{D} y (1 - \lambda).
\]

Thus \( y \) is a solution to the generalized eigenvalue problem (4).

3 Regularization of Block Models

Let \( A \) be the adjacency matrix of some undirected graph. We consider a regularized version of the graph where an edge of weight \( \alpha \) is added between all pairs of nodes, for some constant \( \alpha > 0 \). The corresponding adjacency matrix is given by:

\[
A_{\alpha} = A + \alpha J,
\]
where \( J = 1_n 1_n^T \) is the all-ones matrix of same dimension as \( A \). We denote by \( D_\alpha = \text{diag}(A_\alpha 1_n) \) the corresponding degree matrix and by \( L_\alpha = D_\alpha - A_\alpha \) the Laplacian matrix.

We first consider a simple block model where the graph consists of \( K \) disjoint cliques of respective sizes \( n_1 > n_2 > \cdots > n_K \) nodes, with \( n_K \geq 1 \). In this case, we have \( A = ZZ^T \), where \( Z \) is the membership matrix.

The objective of this section is to demonstrate that, in this setting, the generalized eigenvalue problem of the aggregate graph, with adjacency matrix:

\[
\bar{A}_\alpha = Z^T A_\alpha Z = Z^T (A + \alpha J) Z.
\]

Since \( A = ZZ^T \) and \( W = ZZ^T \), we have \( \bar{A}_\alpha = W^2 + \alpha ZZ^T J Z \). Using the fact that \( Z 1_K = 1_n \), we get \( J = 1_n 1_n^T = Z J K Z^T \) with \( J K = 1_K 1_K^T \) the all-ones matrix of dimension \( K \times K \), so that:

\[
\bar{A}_\alpha = W (I_K + \alpha J_K) W,
\]

where \( I_K \) is the identity matrix of dimension \( K \times K \). We deduce the degree matrix:

\[
\bar{D}_\alpha = W (W + \alpha I_K),
\]

and the Laplacian matrix:

\[
\bar{L}_\alpha = \bar{D}_\alpha - \bar{A}_\alpha = \alpha W (n I_K - J_K W).
\]

The generalized eigenvalue problem associated with the aggregate graph is:

\[
\bar{L}_\alpha y = \lambda \bar{D}_\alpha y.
\]

After multiplication by \( W^{-1} \), we get:

\[
\alpha (n I_K - J_K W) y = \lambda (W + \alpha n I_K) y.
\]

Observing that \( J_K W y = 1_K 1_K^T W y = (1_K^T W y) 1_K \propto 1_K \), we conclude that:

\[
(\alpha (1 - \lambda) - \lambda W) y \propto 1_K,
\]

and since \( W = \text{diag}(n_1, \ldots, n_K) \),

\[
\forall j = 1, \ldots, K, \quad y_j \propto \frac{1}{\lambda n_j - \alpha (1 - \lambda) n}.
\]

The result then follows from the fact that \( x = Z y \). \( \square \)
Lemma 3. The $K$ smallest eigenvalues satisfy:

$$0 = \lambda_1 < \mu_1 < \lambda_2 < \mu_2 < \cdots < \lambda_K < \mu_K,$$

where for all $j = 1, \ldots, K$,

$$\mu_j = \frac{\alpha n}{\alpha n + n_j}.$$

Proof. We know from Lemma 1 that the $K$ smallest eigenvalues are in $[0, 1)$. Let $x$ be a solution to the generalized eigenvalue problem (6) with $\lambda \in (0, 1)$. We know that $x = Zy$ where $y$ is an eigenvector associated with the same eigenvalue $\lambda$ for the aggregate graph. Since $1_K$ is an eigenvector for the eigenvalue 0, we have $y^T \bar{D}_\alpha 1_K = 0$. Using the fact that $\bar{D}_\alpha = W(W + \alpha n I_K)$, we get:

$$\sum_{j=1}^{K} n_j (n_j + \alpha n) y_j = 0.$$

We then deduce from (7) and (8) that $\lambda \notin \{\mu_1, \ldots, \mu_K\}$ and

$$\sum_{j=1}^{K} \frac{n_j (n_j + \alpha n)}{\lambda / \mu_j - 1} = 0.$$

This condition cannot be satisfied if $\lambda < \mu_1$ or $\lambda > \mu_K$ as the terms of the sum would be either all positive or all negative.

Now let $y'$ be another eigenvector for the aggregate graph, with $y^T \bar{D}_\alpha y' = 0$, for the eigenvalue $\lambda' \in (0, 1)$. By the same argument, we get:

$$\sum_{j=1}^{K} n_j (n_j + \alpha n) y_j y'_j = 0,$$

and

$$\sum_{j=1}^{K} \frac{n_j (n_j + \alpha n)}{\lambda / \mu_j - 1} \frac{1}{\lambda' / \mu_j - 1} = 0.$$

with $\lambda' \notin \{\mu_1, \ldots, \mu_K\}$. This condition cannot be satisfied if $\lambda$ and $\lambda'$ are in the same interval $(\mu_j, \mu_{j+1})$ for some $j$ as the terms in the sum would be all positive. There are $K - 1$ eigenvalues in $(0, 1)$ for $K - 1$ such intervals, that is one eigenvalue per interval.

The main result of the paper is the following, showing that the $k - 1$ largest cliques of the original graph can be recovered from the spectral embedding of the regularized graph in dimension $k$.

Theorem 1. Let $X$ be the spectral embedding of dimension $k$, as defined by (2), for some $k$ in the set $\{2, \ldots, K\}$. Then $\text{sign}(X)$ gives the $k - 1$ largest blocks of the graph.

Proof. Let $x$ be the $j$-th column of the matrix $X$, for some $j \in \{2, \ldots, k\}$. In view of Lemma 3, this is the eigenvector associated with eigenvalue $\lambda_j \in (\mu_{j-1}, \mu_j)$, so that

$$\alpha \frac{1 - \lambda_j}{\lambda_j} n \in (n_{j-1}, n_j).$$

In view of Lemma 2, all entries of $x$ corresponding to blocks of size $n_1, n_2, \ldots, n_{j-1}$ have the same sign, the other having the opposite sign.

Theorem 1 can be extended in several ways. First, the assumption of distinct block sizes can easily be relaxed. If there are $L$ distinct values of block sizes, say $m_1, \ldots, m_L$ blocks of sizes $n_1 > \cdots > n_L$, there are $L$ distinct values for the thresholds $\mu_j$ and thus $L$ distinct values for the eigenvalues $\lambda_j$ in $[0, 1)$, the multiplicity of the $j$-th smallest eigenvalue being equal to $m_j$. The spectral embedding in dimension $k$ still gives $k - 1$ cliques of the largest sizes.
Second, the graph may have edges between blocks. Taking \( A = ZZ^T + \varepsilon J \) for instance, for some parameter \( \varepsilon \geq 0 \), the results are exactly the same, with \( \alpha \) replaced by \( \varepsilon + \alpha \). A key observation is that regularization really matters when \( \varepsilon \to 0 \), in which case the initial graph becomes disconnected and, in the absence of regularization, the spectral embedding may isolate small connected components of the graph. In particular, the regularization makes the spectral embedding much less sensitive to noise, as will be demonstrated in the experiments.

Finally, degree correction can be added by varying the node degrees within blocks. Taking \( A = \theta ZZ^T \theta \), for some arbitrary diagonal matrix \( \theta \) with positive entries, similar results can be obtained under the regularization \( A_\alpha = A + \alpha \theta J \theta \). Interestingly, the spectral embedding in dimension \( k \) then recovers the \( k - 1 \) largest blocks in terms of normalized weight, the ratio of the total weight of the block to the number of nodes in the block.

4 Regularization of Bipartite Graphs

Let \( B = \mathbb{R}^{n \times m} \) be the biadjacency matrix of some bipartite graph with respectively \( n, m \) nodes in each part, i.e., \( B_{ij} > 0 \) if and only if there is an edge between node \( i \) in the first part of the graph and node \( j \) in the second part of the graph, with weight \( B_{ij} \). This is an undirected graph of \( n + m \) nodes with adjacency matrix:

\[
A = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}
\]

The spectral embedding of the graph \( \mathcal{G} \) can be written in terms of the biadjacency matrix as follows:

\[
\begin{align*}
BX_2 &= D_1 X_1 (I - \Lambda) \\
B^T X_1 &= D_2 X_2 (I - \Lambda)
\end{align*}
\]

where \( X_1, X_2 \) are the embeddings of each part of the graph, with respective dimensions \( n \times k \) and \( m \times k \), \( D_1 = \text{diag}(B_{1:m}) \) and \( D_2 = \text{diag}(B^T_{1:n}) \). In particular, the spectral embedding of the graph follows from the generalized SVD of the biadjacency matrix \( B \).

The complete regularization adds edges between all pairs of nodes, breaking the bipartite structure of the graph. Another approach consists in applying the regularization to the biadjacency matrix, i.e., in considering the regularized bipartite graph with biadjacency matrix:

\[ B_\alpha = B + \alpha J, \]

where \( J = 1_n 1_m^T \) is here the all-ones matrix of same dimension as \( B \). The spectral embedding of the regularized graph is that associated with the adjacency matrix:

\[
A_\alpha = \begin{bmatrix} 0 & B_\alpha \\ B_\alpha^T & 0 \end{bmatrix}
\]

As in Section 3, we consider a block model so that the biadjacency matrix \( B \) is block-diagonal with all-ones block matrices on the diagonal. Each part of the graph consists of \( K \) groups of nodes of respective sizes \( n_1 > \ldots > n_K \) and \( m_1 > \ldots > m_K \), with nodes of block \( j \) in the first part connected only to nodes of block \( j \) in the second part, for all \( j = 1, \ldots, K \).

We consider the generalized eigenvalue problem \( (9) \) associated with the above matrix \( A_\alpha \). In view of \( (9) \), this is equivalent to the generalized SVD of the regularized biadjacency matrix \( B_\alpha \). We have the following results, whose proofs are deferred to the appendix:

**Lemma 4.** Let \( \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \) be the eigenvalues associated with the generalized eigenvalue problem \( (9) \). We have \( \lambda_1 = 0 < \lambda_2 \leq \ldots \leq \lambda_K < \lambda_{K+1} = \ldots = \lambda_{n-2K} < \ldots < \lambda_n = 2 \).

**Lemma 5.** Let \( x \) be a solution to the generalized eigenvalue problem \( (9) \) with \( \lambda \in (0, 1) \). There exists \( s_1, s_2 \in \{+1, -1\} \) such that for each node \( i \) in block \( j \) of part \( p \) in \( \{1, 2\} \),

\[
\text{sign}(x_i) = s_p \iff \frac{n_j m_j}{(n_j + \alpha m)(m_j + \alpha m)} \geq 1 - \lambda.
\]

**Lemma 6.** The \( K \) smallest eigenvalues satisfy:

\[
0 = \lambda_1 < \mu_1 < \lambda_2 < \mu_2 < \ldots < \lambda_K < \mu_K,
\]

where for all \( j = 1, \ldots, K \),

\[
\mu_j = 1 - \frac{n_j m_j}{(n_j + \alpha m)(m_j + \alpha m)}.
\]
Theorem 2. Let $X$ be the spectral embedding of dimension $k$, as defined by (2), for some $k$ in the set $\{2, \ldots, K\}$. Then $\text{sign}(X)$ gives the $k-1$ largest blocks of each part of the graph.

Like Theorem 1 the assumption of decreasing block sizes can easily be relaxed. Assume that block pairs are indexed in decreasing order of $\mu_j$. Then the spectral embedding of dimension $k$ gives the $k-1$ first block pairs for that order. It is interesting to notice that the order now depends on $\alpha$: when $\alpha \to 0^+$, the block pairs $j$ of highest value $(\frac{n_j}{m_j} + \frac{m_j}{n_j})^{-1}$ (equivalently, highest harmonic mean of proportions of nodes in each part of the graph) are isolated first; when $\alpha \to +\infty$, the block pairs $j$ of highest value $\frac{n_j m_j}{nm}$ (equivalently, the highest geometric mean of proportions of nodes in each part of the graph) are isolated first.

The results also extend to non-block diagonal biadjacency matrices $B$ and degree-corrected models, as for Theorem 1.

5 EXPERIMENTS

We now illustrate the impact of regularization on the quality of spectral embedding. We focus on a clustering task, using both synthetic and real datasets where the ground-truth clusters are known. In all experiments, we skip the first dimension of the spectral embedding as it is not informative (the corresponding eigenvector is the all-ones vector, up to some multiplicative constant). The code to reproduce these experiments is available online.

5.1 TOY GRAPH

We first illustrate the theoretical results of the paper with a toy graph consisting of 3 cliques of respective sizes $5, 3, 2$. We compute the spectral embeddings in dimension 1, using the second smallest eigenvalue. Denoting by $Z$ the membership matrix, we get $X \approx Z(-0.08, 0.11, 0.05)^T$ for $\alpha = 1$, showing that the embedding isolates the largest cluster; this is not the case in the absence of regularization, where $X \approx Z(0.1, -0.1, 0.41)^T$.

5.2 DATASETS

This section describes the datasets used in our experiments. All graphs are considered as undirected. Table 1 presents the main features of the graphs.

Stochastic Block-Model (SBM)  We generate 100 instances of the same stochastic block model (Holland et al., 1983). There are 100 blocks of size 20, with intra-block edge probability set to 0.5 for the first 50 blocks and 0.05 for the other blocks. The inter-block edge probability is set to 0.001. Other sets of parameters can be tested using the code available online. The ground-truth cluster of each node corresponds to its block.

20newsgroup (NG)  This dataset consists of around 18000 newsgroups posts on 20 topics. This defines a weighted bipartite graph between documents and words. The label of each document corresponds to the topic.

Wikipedia for Schools (WS)  (Haruechaiyasak & Damrongrat, 2008). This is the graph of hyperlinks between a subset of Wikipedia pages. The label of each page is its category (e.g., countries, mammals, physics).

<table>
<thead>
<tr>
<th>dataset</th>
<th>SBM</th>
<th>NG</th>
<th>WS</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes ($n_i$)</td>
<td>2000</td>
<td>10723</td>
<td>4591</td>
</tr>
<tr>
<td># edges</td>
<td>$\approx 5.10^3$</td>
<td>$\approx 2.10^6$</td>
<td>$\approx 2.10^5$</td>
</tr>
<tr>
<td># clusters in ground truth</td>
<td>100</td>
<td>20</td>
<td>14</td>
</tr>
</tbody>
</table>

https://github.com/research-submissions/iclr20
5.3 Metrics

We consider a large set of metrics from the clustering literature. All metrics are upper-bounded by 1 and the higher the score the better.

**Homogeneity (H), Completeness (C) and V-measure score (V)** (Rosenberg & Hirschberg, 2007). Supervised metrics. A cluster is homogeneous if all its data points are members of a single class in the ground truth. A clustering is complete if all the members of a class in the ground truth belong to the same cluster in the prediction. Harmonic mean of homogeneity and completeness.

**Adjusted Rand Index (ARI)** (Hubert & Arabie, 1985). Supervised metric. This is the corrected for chance version of the Rand Index which is itself an accuracy on pairs of samples.

**Adjusted Mutual Information (AMI)** (Vinh et al., 2010) Supervised metric. Adjusted for chance version of the mutual information.

**Fowlkes-Mallows Index (FMI)** (Fowlkes & Mallows, 1983). Supervised metric. Geometric mean between precision and recall on the edge classification task, as described for the ARI.

**Modularity (Q)** (Newman, 2006). Unsupervised metric. Fraction of edges within clusters compared to that is some null model where edges are shuffled at random.

**Normalized Standard Deviation (NSD)** Unsupervised metric. 1 minus normalized standard deviation in cluster size.

5.4 Experimental Setup

All graphs are embedded in dimension 20, with different regularization parameters. To compare the impact of this parameter across different datasets, we use a relative regularization parameter \((w/n^2)\alpha\), where \(w = \sum_{i=1}^{n} A_{ii}\) is the total weight of the graph.

We use the K-Means algorithm with to cluster the nodes in the embedding space. The parameter \(K\) is set to the ground-truth number of clusters (other experiments with different values of \(K\) are reported in the Appendix). We use the Scikit-learn (Pedregosa et al., 2011) implementation of K-Means and the metrics, when available. The spectral embedding and the modularity are computed with the Scikit-network package, see the documentation for more details.

5.5 Results

We report the results in Table 2 for relative regularization parameter \(\alpha = 0, 0.1, 1, 10\). We see that the regularization generally improves performance, the optimal value of \(\alpha\) depending on both the dataset and the score function. As suggested by Lemma 3, the optimal value of the regularization parameter should depend on the distribution of cluster sizes, on which we do not have any prior knowledge.

To test the impact of noise on the spectral embedding, we add isolated nodes with self loop to the graph and compare the clustering performance with and without regularization. The number of isolated nodes is given as a fraction of the initial number of nodes in the graph. Scores are computed only on the initial nodes. The results are reported in Table 3 for the Wikipedia for Schools dataset. We observe that, in the absence of regularization, the scores drop even with only 1% noise. The computed clustering is a trivial partition with all initial nodes in the same cluster. This means that the 20 first dimensions of the spectral embedding focus on the isolated nodes. On the other hand, the scores remain approximately constant in the regularized case, which suggests that regularization makes the embedding robust to this type of noise.

[https://scikit-network.readthedocs.io/](https://scikit-network.readthedocs.io/)
Table 2: Impact of regularization on clustering performance.

<table>
<thead>
<tr>
<th>α</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>NSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.19</td>
<td>0.27</td>
<td>0.22</td>
<td>0.0</td>
<td>0.01</td>
<td>0.03</td>
<td>0.45</td>
<td>0.76</td>
</tr>
<tr>
<td>0.1</td>
<td>0.33</td>
<td>0.35</td>
<td>0.34</td>
<td>0.0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.52</td>
<td>0.91</td>
</tr>
<tr>
<td>1</td>
<td><strong>0.36</strong></td>
<td><strong>0.37</strong></td>
<td><strong>0.36</strong></td>
<td>0.0</td>
<td>0.01</td>
<td>0.01</td>
<td>0.50</td>
<td><strong>0.92</strong></td>
</tr>
<tr>
<td>10</td>
<td>0.28</td>
<td>0.34</td>
<td>0.30</td>
<td>0.0</td>
<td>0.00</td>
<td>0.02</td>
<td>0.36</td>
<td>0.78</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>α</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>NSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.40</td>
<td><strong>0.70</strong></td>
<td>0.51</td>
<td>0.19</td>
<td>0.50</td>
<td>0.34</td>
<td><strong>0.21</strong></td>
<td>0.55</td>
</tr>
<tr>
<td>0.1</td>
<td>0.44</td>
<td><strong>0.70</strong></td>
<td><strong>0.54</strong></td>
<td>0.22</td>
<td>0.54</td>
<td><strong>0.35</strong></td>
<td><strong>0.21</strong></td>
<td>0.59</td>
</tr>
<tr>
<td>1</td>
<td><strong>0.46</strong></td>
<td>0.67</td>
<td><strong>0.54</strong></td>
<td>0.20</td>
<td><strong>0.54</strong></td>
<td>0.33</td>
<td>0.20</td>
<td><strong>0.60</strong></td>
</tr>
<tr>
<td>10</td>
<td>0.37</td>
<td>0.55</td>
<td>0.45</td>
<td>0.13</td>
<td>0.44</td>
<td>0.26</td>
<td>0.17</td>
<td>0.56</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>α</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>NSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.23</td>
<td>0.29</td>
<td>0.25</td>
<td>0.05</td>
<td>0.25</td>
<td>0.26</td>
<td>0.25</td>
<td>0.49</td>
</tr>
<tr>
<td>0.1</td>
<td><strong>0.26</strong></td>
<td><strong>0.29</strong></td>
<td><strong>0.28</strong></td>
<td><strong>0.10</strong></td>
<td><strong>0.27</strong></td>
<td><strong>0.26</strong></td>
<td><strong>0.29</strong></td>
<td>0.61</td>
</tr>
<tr>
<td>1</td>
<td>0.23</td>
<td>0.24</td>
<td>0.23</td>
<td>0.04</td>
<td>0.23</td>
<td>0.20</td>
<td><strong>0.30</strong></td>
<td><strong>0.65</strong></td>
</tr>
<tr>
<td>10</td>
<td>0.19</td>
<td>0.22</td>
<td>0.20</td>
<td>0.00</td>
<td>0.19</td>
<td>0.20</td>
<td>0.23</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Table 3: Impact of noise on clustering performance (WS dataset).

<table>
<thead>
<tr>
<th>noise</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 %</td>
<td>0.23</td>
<td>0.29</td>
<td>0.25</td>
<td>0.05</td>
<td>0.25</td>
<td>0.26</td>
<td>0.25</td>
<td>0.49</td>
</tr>
<tr>
<td>1 %</td>
<td>0.00</td>
<td>0.49</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.39</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>5 %</td>
<td>0.00</td>
<td>0.49</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.39</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>10 %</td>
<td>0.00</td>
<td>0.49</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.39</td>
<td>0.00</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>noise</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 %</td>
<td>0.23</td>
<td>0.24</td>
<td>0.23</td>
<td>0.04</td>
<td>0.23</td>
<td>0.2</td>
<td>0.3</td>
<td>0.65</td>
</tr>
<tr>
<td>1 %</td>
<td>0.24</td>
<td>0.24</td>
<td>0.24</td>
<td>0.04</td>
<td>0.23</td>
<td>0.2</td>
<td>0.3</td>
<td>0.66</td>
</tr>
<tr>
<td>5 %</td>
<td>0.23</td>
<td>0.23</td>
<td>0.23</td>
<td>0.05</td>
<td>0.22</td>
<td>0.2</td>
<td>0.3</td>
<td>0.67</td>
</tr>
<tr>
<td>10 %</td>
<td>0.24</td>
<td>0.23</td>
<td>0.23</td>
<td>0.05</td>
<td>0.23</td>
<td>0.2</td>
<td>0.3</td>
<td>0.67</td>
</tr>
</tbody>
</table>

6 Conclusion and Perspectives

In this paper, we have provided a simple explanation for the well-known benefits of regularization on spectral embedding. Specifically, regularization forces the embedding to focus on the largest clusters, making the embedding more robust to noise. This result was obtained through the explicit characterization of the embedding for a simple block model, and extended to bipartite graphs.

An interesting perspective of our work is the extension to stochastic block models, using for instance the concentration results proved in [Lei et al., 2015; Le et al., 2017]. Another problem of interest is the impact of regularization on other downstream tasks, like link prediction. Finally, we would like to further explore the impact of the regularization parameter, exploiting the theoretical results presented in this paper.
REFERENCES


## Appendix

We provide of proof of Theorem 2 as well as a complete set of experimental results.

### A Regularization of Bipartite Graphs

The proof of Theorem 2 follows the same workflow as that of Theorem 1. Let 
\[ Z_1 \in \mathbb{R}^{n \times K} \quad \text{and} \quad Z_2 \in \mathbb{R}^{m \times K} \]
be the left and right membership matrices for the block matrix \( B \in \mathbb{R}^{n \times m} \). The aggregated matrix is \( \overline{B} = Z_1^T B Z_2 \in \mathbb{R}^{K \times K} \). The diagonal matrices of block sizes are \( W_1 = Z_1^T Z_1 \) and \( W_2 = Z_2^T Z_2 \). We have the equivalent of Proposition 1:

**Proposition 2.** Let \( x_1, x_2 \) be a solution to the generalized singular value problem:

\[
\begin{align*}
B x_2 &= \sigma D_1 x_1 \\
B^T x_1 &= \sigma D_2 x_2
\end{align*}
\]

Then either \( Z_1^T x_1 = Z_2^T x_2 = 0 \) and \( \sigma = 0 \) or \( x_1 = Z_1 y_1 \) and \( x_2 = Z_2 y_2 \) where \( y_1, y_2 \) is a solution to the generalized singular value problem:

\[
\begin{align*}
\bar{B} y_2 &= \sigma \bar{D}_1 y_1 \\
\bar{B}^T y_1 &= \sigma \bar{D}_2 y_2.
\end{align*}
\]

**Proof.** Since the rank of \( B \) is equal to \( K \), there are \( n-K \) pairs of singular vectors \( (x_1, x_2) \) associated with the singular values 0, each satisfying \( Z_1^T x_1 = 0 \) and \( Z_2^T x_2 = 0 \). By orthogonality, the other pairs of singular vectors satisfy \( x_1 = Z_1 y_1 \) and \( x_2 = Z_2 y_2 \) for some vectors \( y_1, y_2 \in \mathbb{R}^K \). By replacing these in the original generalized singular value problem, we get that \( (y_1, y_2) \) is a solution to the generalized singular value problem for the aggregate graph. \( \square \)

In the following, we focus on the block model described in Section 4 where \( B = Z_1 Z_2^T \).

**Proof of Lemma 4.** The generalized eigenvalue problem (6) associated with the regularized matrix \( A_\alpha \) is equivalent to the generalized SVD of the regularized biadjacency matrix \( B_\alpha \):

\[
\begin{align*}
B_\alpha x_2 &= \sigma D_{\alpha,1} x_1 \\
B_\alpha^T x_1 &= \sigma D_{\alpha,2} x_2,
\end{align*}
\]

with \( \sigma = 1 - \lambda \).

In view of Proposition 2 the singular value \( \sigma = 0 \) has multiplicity \( n-K \), meaning that the eigenvalue \( \lambda = 1 \) has multiplicity \( n-K \). Since the graph is connected, the eigenvalue 0 has multiplicity 1. The proof then follows from the observation that if \( (x_1, x_2) \) is a pair of singular vectors for the singular value \( \sigma \), then the vectors \( x = (x_1, \pm x_2)^T \) are eigenvectors for the eigenvalues \( 1 - \sigma, 1 + \sigma \).

**Proof of Lemma 5.** By Proposition 2 we can focus on the generalized singular value problem for the aggregate graph:

\[
\begin{align*}
\bar{B}_\alpha y_2 &= \sigma \bar{D}_{\alpha,1} y_1 \\
\bar{B}_\alpha^T y_1 &= \sigma \bar{D}_{\alpha,2} y_2,
\end{align*}
\]

Since \( \bar{B}_\alpha = W_1 (I_K + \alpha J_K) W_2 \),
and
\[
\begin{align*}
\hat{D}_{\alpha,1} &= W_1(W_2 + \alpha nI), \\
\hat{D}_{\alpha,2} &= W_2(W_1 + \alpha mI),
\end{align*}
\]
we have:
\[
\begin{align*}
W_1(I_K + \alpha J_K)W_2y_2 &= W_1(W_2 + \alpha nI)y_1\sigma, \\
W_2(I_K + \alpha J_K)W_1y_1 &= W_2(W_1 + \alpha mI)y_2\sigma.
\end{align*}
\]
Observing that \(J_KW_1y_1 \propto 1_K\) and \(J_KW_2y_2 \propto 1_K\), we get:
\[
\begin{align*}
(W_2 + \alpha mI)\hat{y}_1\sigma - W_2y_2 &\propto 1_K, \\
(W_1 + \alpha nI)\hat{y}_2\sigma - W_1y_1 &\propto 1_K.
\end{align*}
\]
As two diagonal matrices commute, we obtain:
\[
\begin{align*}
(W_1 + \alpha nI_K)(W_2 + \alpha mI_K)\hat{y}_1\sigma - W_1W_2y_1 &= (\eta_1(W_1 + \alpha nI_K) + \eta_2W_2)1_K, \\
(W_1 + \alpha nI_K)(W_2 + \alpha mI_K)\hat{y}_2\sigma - W_1W_2y_1 &= (\eta_1W_1 + \eta_2(W_2 + \alpha mI_K))1_K,
\end{align*}
\]
for some constants \(\eta_1, \eta_2\), and
\[
\begin{align*}
y_{1,j} &= \frac{\eta_1(n_j + \alpha n) + \eta_2m_j}{(n_j + \alpha n)(m_j + \alpha m)\sigma - n_jm_j}, \\
y_{2,j} &= \frac{\eta_1n_j + \eta_2(m_j + \alpha m)}{(n_j + \alpha n)(m_j + \alpha m)\sigma - n_jm_j}.
\end{align*}
\]
Letting \(s_1 = -\text{sign}(\eta_1(n_j + \alpha n) + \eta_2m_j)\) and \(s_2 = -\text{sign}(\eta_1n_j + \eta_2(m_j + \alpha m))\), we get:
\[
\text{sign}(y_{1,j}) = s_1 \iff \text{sign}(y_{2,j}) = s_2 \iff \frac{n_jm_j}{(n_j + \alpha n)(m_j + \alpha m)} \geq \sigma = 1 - \lambda,
\]
and the result follows from the fact that \(x_1 = Z_1y_1\) and \(x_2 = Z_2y_2\).

**Proof of Lemma 6** The proof is the same as that of Lemma 5, where the threshold values follow from Lemma 5.

\[
\mu_j = 1 - \frac{n_jm_j}{(n_j + \alpha n)(m_j + \alpha m)}.
\]

**Proof of Theorem 2** Let \(x\) be the \(j\)-th column of the matrix \(X\), for some \(j \in \{2, \ldots, k\}\). In view of Lemma 6 this is the eigenvector associated with eigenvalue \(\lambda_j \in (\mu_{j-1}, \mu_j)\). In view of Lemma 4 all entries of \(x\) corresponding to blocks of size \(n_1, n_2, \ldots, n_{j-1}\) have the same sign, the other having the opposite sign.

**B Experimental Results**

In this section, we present more extensive experimental results.

Tables 4 and 5 present results for the same experiment as in Table 2 but for different values of \(K\), namely \(K = 2\) (bisection of the graph) and \(K = K_{\text{truth}}/2\) (half of the ground-truth value). As for \(K = K_{\text{true}}\), regularization generally improves clustering performance. However, the optimal value of \(\alpha\) remains both dataset dependent and metric dependent. Note that, for the NG and WS datasets, the clustering remains trivial in the case \(K = 2\), one cluster containing all the nodes, until a certain amount of regularization.

Table 6 presents the different scores for both types of regularization on the NG dataset. As we can see, preserving the bipartite structure of the graph leads to slightly better performance.

Finally, Table 7 shows the impact of regularization in the presence of noise for the NG dataset. The conclusions are similar as for the WS dataset: regularization makes the spectral embedding much more robust to noise.
Table 4: Impact of regularization on clustering performance. $K = 2$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>NSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00</td>
<td>0.43</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.10</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>0.1</td>
<td>0.00</td>
<td>0.47</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
<td>0.04</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>0.34</td>
<td>0.83</td>
</tr>
<tr>
<td>10</td>
<td>0.01</td>
<td>0.09</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.09</td>
<td>0.13</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table 5: Impact of regularization on clustering performance. $K = K_{true}/2$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>NSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.08</td>
<td>0.21</td>
<td>0.11</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.41</td>
<td>0.47</td>
</tr>
<tr>
<td>0.1</td>
<td>0.20</td>
<td>0.27</td>
<td>0.23</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.55</td>
<td>0.84</td>
</tr>
<tr>
<td>1</td>
<td>0.24</td>
<td>0.29</td>
<td>0.26</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.54</td>
<td>0.90</td>
</tr>
<tr>
<td>10</td>
<td>0.19</td>
<td>0.28</td>
<td>0.23</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.40</td>
<td>0.70</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>NSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.27</td>
<td>0.76</td>
<td>0.40</td>
<td>0.11</td>
<td>0.39</td>
<td>0.31</td>
<td>0.20</td>
<td>0.41</td>
</tr>
<tr>
<td>0.1</td>
<td>0.28</td>
<td>0.73</td>
<td>0.41</td>
<td>0.11</td>
<td>0.40</td>
<td>0.30</td>
<td>0.18</td>
<td>0.43</td>
</tr>
<tr>
<td>1</td>
<td>0.38</td>
<td>0.72</td>
<td>0.50</td>
<td>0.18</td>
<td>0.50</td>
<td>0.34</td>
<td>0.21</td>
<td>0.57</td>
</tr>
<tr>
<td>10</td>
<td>0.31</td>
<td>0.62</td>
<td>0.42</td>
<td>0.11</td>
<td>0.42</td>
<td>0.27</td>
<td>0.17</td>
<td>0.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>NSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.23</td>
<td>0.29</td>
<td>0.25</td>
<td>0.05</td>
<td>0.25</td>
<td>0.26</td>
<td>0.25</td>
<td>0.49</td>
</tr>
<tr>
<td>0.1</td>
<td>0.26</td>
<td>0.29</td>
<td>0.28</td>
<td>0.10</td>
<td>0.27</td>
<td>0.26</td>
<td>0.29</td>
<td>0.61</td>
</tr>
<tr>
<td>1</td>
<td>0.23</td>
<td>0.24</td>
<td>0.23</td>
<td>0.04</td>
<td>0.23</td>
<td>0.20</td>
<td>0.30</td>
<td>0.65</td>
</tr>
<tr>
<td>10</td>
<td>0.19</td>
<td>0.22</td>
<td>0.20</td>
<td>-0.00</td>
<td>0.19</td>
<td>0.20</td>
<td>0.23</td>
<td>0.53</td>
</tr>
</tbody>
</table>
Table 6: Regularization of the adjacency vs. biadjacency matrix on the NG dataset ($\alpha = 1$).

\[ K = K_{true}/2 \]

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj.</td>
<td>0.38</td>
<td>0.72</td>
<td>0.50</td>
<td>0.18</td>
<td>0.50</td>
<td>0.34</td>
<td>0.21</td>
<td>0.57</td>
</tr>
<tr>
<td>Biadj.</td>
<td>0.41</td>
<td>0.72</td>
<td><strong>0.52</strong></td>
<td><strong>0.19</strong></td>
<td><strong>0.52</strong></td>
<td><strong>0.35</strong></td>
<td>0.21</td>
<td><strong>0.61</strong></td>
</tr>
</tbody>
</table>

\[ K = K_{true} \]

<table>
<thead>
<tr>
<th></th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj.</td>
<td>0.46</td>
<td>0.67</td>
<td><strong>0.54</strong></td>
<td>0.20</td>
<td>0.54</td>
<td>0.33</td>
<td>0.2</td>
<td>0.60</td>
</tr>
<tr>
<td>Biadj.</td>
<td><strong>0.47</strong></td>
<td><strong>0.68</strong></td>
<td><strong>0.56</strong></td>
<td><strong>0.21</strong></td>
<td><strong>0.55</strong></td>
<td><strong>0.34</strong></td>
<td>0.2</td>
<td><strong>0.61</strong></td>
</tr>
</tbody>
</table>

Table 7: Impact of noise on clustering performance (NG dataset).

\[ \alpha = 0 \]

<table>
<thead>
<tr>
<th>noise</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 %</td>
<td>0.40</td>
<td>0.70</td>
<td>0.51</td>
<td>0.19</td>
<td>0.50</td>
<td>0.34</td>
<td>0.21</td>
<td>0.55</td>
</tr>
<tr>
<td>1 %</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>5 %</td>
<td>0.14</td>
<td>0.65</td>
<td>0.23</td>
<td>0.06</td>
<td>0.23</td>
<td>0.27</td>
<td>0.13</td>
<td>0.30</td>
</tr>
<tr>
<td>10 %</td>
<td>0.00</td>
<td>0.36</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.23</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\[ \alpha = 1 \]

<table>
<thead>
<tr>
<th>noise</th>
<th>H</th>
<th>C</th>
<th>V</th>
<th>ARI</th>
<th>AMI</th>
<th>FMI</th>
<th>Q</th>
<th>std</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 %</td>
<td>0.46</td>
<td>0.67</td>
<td>0.54</td>
<td>0.20</td>
<td>0.54</td>
<td>0.33</td>
<td>0.2</td>
<td>0.60</td>
</tr>
<tr>
<td>1 %</td>
<td>0.48</td>
<td>0.66</td>
<td>0.56</td>
<td>0.21</td>
<td>0.56</td>
<td>0.33</td>
<td>0.2</td>
<td>0.64</td>
</tr>
<tr>
<td>5 %</td>
<td>0.49</td>
<td>0.66</td>
<td>0.56</td>
<td>0.23</td>
<td>0.56</td>
<td>0.34</td>
<td>0.2</td>
<td>0.66</td>
</tr>
<tr>
<td>10 %</td>
<td>0.45</td>
<td>0.66</td>
<td>0.54</td>
<td>0.20</td>
<td>0.54</td>
<td>0.33</td>
<td>0.2</td>
<td>0.59</td>
</tr>
</tbody>
</table>