Knowl Inf Syst (2010) 23:243–265
DOI 10.1007/s10115-009-0215-1

REGULAR PAPER

Enhancing the stability and efficiency of spectral ordering with partial supervision and feature selection

Dimitrios Mavroeidis · Ella Bingham

Received: 14 January 2009 / Revised: 28 March 2009 / Accepted: 11 April 2009 /
Published online: 3 June 2009

Abstract Several studies have demonstrated the prospects of spectral ordering for data mining. One successful application is seriation of paleontological findings, i.e. ordering the sites of excavation, using data on mammal co-occurrences only. However, spectral ordering ignores the background knowledge that is naturally present in the domain: paleontologists can derive the ages of the sites within some accuracy. On the other hand, the age information is uncertain, so the best approach would be to combine the background knowledge with the information on mammal co-occurrences. Motivated by this kind of partial supervision we propose a novel semi-supervised spectral ordering algorithm that modifies the Laplacian matrix such that domain knowledge is taken into account. Also, it performs feature selection by discarding features that contribute most to the unwanted variability of the data in bootstrap sampling. Moreover, we demonstrate the effectiveness of the proposed framework on the seriation of Usenet newsgroup messages, where the task is to find out the underlying flow of discussion. The theoretical properties of our algorithm are thoroughly analyzed and it is demonstrated that the proposed framework enhances the stability of the spectral ordering output and induces computational gains.

Keywords Spectral ordering · Semi-supervised learning · Laplacian · Matrix perturbation theory · Eigengap · Seriation · Paleontology

1 Introduction

In this paper, we consider the task of ordering the observations in the data, accompanied by partial supervision and feature selection, aiming at a more stable ordering. Although it may initially seem surprising the we employ partial supervision and feature selection in a common

D. Mavroeidis (✉)
Department of Informatics, Athens University of Economics and Business, Athens, Greece
e-mail: dmavr@aueb.gr

E. Bingham
Helsinki Institute for Information Technology, University of Helsinki, Helsinki, Finland
e-mail: ella@iki.fi

framework, it is analytically demonstrated in the paper, that each component addresses a different cause of instability of the results. In our context, stability refers to the variation of the end result with respect to small changes in the data; in practice we will measure this by bootstrap sampling.

In distance based ordering, the task is to find a permutation of objects such that similar objects become adjacent; in addition, the more dissimilar the objects are, the larger the order distance between them. The standard optimization problem formulation used for deriving the distance based ordering is known to be *NP*-hard [11], and spectral ordering [2,6] presents a popular, algorithmically feasible approach for deriving approximate solutions. Despite the name "ordering", the aim is not to rank the objects into any preference ranking, and the first and last object in the ordering are merely those that are maximally dissimilar to each other. Algorithmically, the order solution is derived by the eigenvector corresponding to the second eigenvalue of the data Laplacian matrix.

Our main application area is paleontology: our observations (objects, instances) are sites of excavation and our features (attributes, variables) are mammal genera whose remains are found at these sites. In addition, we have auxiliary information on the estimated ages of the sites; this information is uncertain to some degree. Spectral ordering of the sites of excavation can be based solely on the co-occurrences of the mammal genera, irrespective of the ages of the sites. It has been shown [9] that this kind of plain spectral ordering is a fast and standardized way of biochronological ordering of the sites. Albeit the favorable results in the biochronological ordering task, the spectral ordering does not take into account the background knowledge that naturally exists in the domain. The successful incorporation of domain knowledge is expected to increase the quality of the results.

In the current study, we take advantage of the domain knowledge of the ages of the sites and combine that with the spectral ordering, ending up with a semi-supervised spectral ordering.[1] In addition, we consider feature selection. Towards this target the features that contribute most to the unwanted variation of the data (measured by bootstrap sampling) will be removed. These features correspond to mammals whose observations are noisy. The paleontological data is noisy in many respects [10]: the preservation, recovery and identification of fossils are all random to some extent. These uncertainties are, however, hard to quantify, and a systematic way of characterizing the uncertainty would be most welcome—the behaviour of the features in bootstrap sampling is here chosen for this task.

Another domain of application is text document data in which the observations are Usenet newsgroup articles and the features are the most common terms. There is a natural underlying order in the data, as people respond to each others' postings. As "domain knowledge" we use the time stamps of the articles. This might lead to a slightly different ordering, as the users often respond to old postings instead of the newest ones, but it can however serve as initial domain knowledge.

The two components of the proposed framework, namely partial supervision and feature selection will make the resulting ordering more stable with respect to small variations in the data. As it is analyzed in detail in Sect. 6, each component of the framework addresses a different cause of instability of the spectral ordering results.

The theoretical analysis suggests and the experiments verify that the main advantages of the proposed framework as induced by the enhancement of stability are twofold:

– The results become more resilient to perturbations of the input, thus the reliability of the results is increased.

[1] In the context of this work we will use the terms "semi-supervised" and "partial supervision" to refer to the domain knowledge interchangeably.

– The power method [23] computes the ordering result more efficiently than in the original setting.

2 Spectral ordering

Given a set of n objects and a pairwise similarity measure between them, the task of distance based ordering is to derive the order indexes of the objects such that similar objects are placed in adjacent orders while dissimilar objects are placed far apart. More formally, distance sensitive ordering considers the following optimization problem:

$$\min_r \sum_{i,j} (r(i) - r(j))^2 w_{ij}$$

where w_{ij} is the similarity between objects i and j and vector r is the permutation of $\{1, 2, \ldots, n\}$ that optimizes the objective function. The values of the elements $r(i)$ of vector r reflect the ordering of the objects.

It is known that the general optimization problem related to distance based ordering is *NP*-hard [11], and thus approximate solutions should be considered. A popular approach is spectral ordering [2,6] that performs a continuous relaxation on the solution vector r, and reduces the optimization problem to a standard eigenvalue-eigenvector problem. Such relaxations are commonly used in data mining to effectively approximate several computationally hard problems with matrix-based algorithms [15]. In the context of this work we rely on a slight modification of the standard spectral ordering formulation as derived by [6], where the authors derive the ordering solution as the second eigenvector[2] of the normalized Laplacian matrix $L = D^{-1/2}WD^{-1/2}$. Here, W is the object-object similarity matrix $W = X^T X$, D is the diagonal degree matrix containing the row sums of W, and the data matrix X contains the objects as its columns and the features as its rows. Other choices of W are also possible: W can essentially be any positive semi-definite object-object similarity matrix. The use of the normalized Laplacian facilitates the theoretical analysis of the proposed semi-supervised spectral ordering framework and also presents theoretical advantages [22] over the unnormalized Laplacian that is commonly used for spectral ordering.

It should be noted that in the spectral graph theory literature the normalized Laplacian matrix is commonly referred to as $L = I - D^{-1/2}WD^{-1/2}$, however, in the context of this paper, we will employ the aforementioned notation and consider the normalized Laplacian as $L = D^{-1/2}WD^{-1/2}$. This matrix is well studied in the context of spectral graph theory (e.g. [21] and references therein) and it is known to have 1 as its largest eigenvalue. Moreover, by defining the object-similarity matrix $W = X^T X$, $L = D^{-1/2}WD^{-1/2}$ becomes positive semi-definite.

3 Two factors that determine the stability of spectral ordering

A common approach for measuring the stability of spectral algorithms requires the quantification, in the form of an error perturbation matrix E, of the uncertainty associated with the input matrix (e.g. [16]). Based on the matrix E, the stability of spectral ordering is determined by the similarity of the ordering solution as derived by the original Laplacian matrix L versus

[2] We consider the eigenvalues ordered in decreasing order, i.e. the first eigenvalue is the largest eigenvalue and so on. The first eigenvector is the eigenvector that corresponds to the largest eigenvalue and so on.

the perturbed Laplacian matrix $L + E$. Further details on the computation of E in the domain of interest will be provided in Sect. 6.3.

Based on this formulation, the stability of the ordering solution can be derived by Matrix Perturbation Theory, and more precisely Stewart's theorem on the perturbation of invariant subspaces [20]. Based on Stewart's theorem we can derive an upper bound on the difference between the ordering solution of L versus $L + E$. The upper bound applies when the smallest eigengap between the second eigenvalue of L and the rest is larger than four times the spectral norm of matrix E. In the case of spectral ordering the smallest eigengap is determined by the eigengap between the first and the second eigenvalue of the Laplacian matrix and the eigengap between the second and the third.

The upper bound gets smaller as the eigengap enlarges and the norm of the perturbation matrix E decreases. Thus, the stability depends on two factors: *the size of the eigengap and the norm of the perturbation.*

We can state the aforementioned result in a more formal manner using the *condition number of the eigenvector problem*. Recall that the condition number is a common tool in linear algebra for assessing the sensitivity of a solution with respect to small variations of the input. In the case of spectral ordering, we are interested in assessing the sensitivity of the eigenvector with respect to small perturbations of the Laplacian matrix. That is, we are interested in deriving an expression $||u - \tilde{u}|| \leq \kappa ||E||$, where u and $\tilde{u}$ are eigenvectors of L and $L + E$ respectively and κ is the condition number of eigenvector u. The general definition of the eigenvector condition number is rather complicated. However, it is largely simplified in the case of Hermitian matrices where it is defined as $\kappa = \frac{1}{\min\_\text{eigengap}}$, with min _eigengap being the minimum eigengap between the eigenvalue corresponding to eigenvector u and the rest. Thus, the condition number of the spectral ordering problem will depend on the eigengap between the first and the second eigenvalue, as well as the eigengap between the second and the third.

As we analyze further in the subsequent section, these eigengaps are not a mere theoretical artifact but are associated with the data-structure as well as computational issues related to the derivation of the spectral ordering solution.

4 Semantics of the eigengaps

4.1 Eigengap $\lambda_1 - \lambda_2$

The eigengap between the first and the second eigenvalue of the Laplacian matrix is associated with the level of data connectivity. More precisely, if we consider the Laplacian $D^{-1/2}WD^{-1/2}$ and the associated graph (i.e. a graph with edge weights $W(i, j)$), then the size of the second eigenvalue is associated with the cost of producing two separated clusters [7,21]. In fact when the eigengap is 0, i.e. the algebraic multiplicity of first eigenvalue is larger than 1, then the graph is disconnected and the clusters can be produced with zero cost. The following theorem illustrates this relation (note that we have appropriately changed the theorem statement from [21] to take into account that we consider the Laplacian $D^{-1/2}WD^{-1/2}$ instead of $I - D^{-1/2}WD^{-1/2}$):

Theorem 4.1 (can be found in [21]) *Let G be an indirected graph with non-negative weights W. Then the multiplicity k of the eigenvalue 1 of matrix $L = D^{-1/2}WD^{-1/2}$ equals the number of connected components in the graph. The eigenspace of 1 is spanned by the vectors $D^{1/2}e_{A_i}$ of those components, where e_{A_i} is such that $e(j)_{A_i} = 1$ for all vertices j that belong to the connected component A_i.*

Theorem 4.1 signifies that when the second eigenvalue is close to the first, a small amount of perturbation can make the graph disconnected, thus significantly affecting the second eigenvector. Thus, spectral graph theory provides us with the necessary tools for understanding the source of instability when the eigengap between the first and the second eigenvalue is small.

4.2 Eigengap $\lambda_2 - \lambda_3$

In order to study the eigengap between the second and the third eigenvalue of the Laplacian matrix L, we assume that the data is adequately connected (i.e. the algebraic multiplicity of the largest eigenvalue is 1) and consider the following transformation: $L' = L - vv^{\mathrm{T}}$, where v is the first eigenvector of Laplacian L (i.e. $v = \frac{D^{1/2}e}{||D^{1/2}e||}$ with D being the degree matrix of the Laplacian L and e a unit vector, $e(i) = 1$ for all i). With this definition the matrix L', apart from v, has exactly the same eigenvectors and eigenvalues as L. Thus the second eigenvalue of L is the largest eigenvalue of L'. This transformation is always possible and requires solely the computation of the degree matrix D.

The transformation of matrix L makes apparent the relevance of the power method [23] for computing the spectral ordering solution. Recall that the power method does not derive the full eigen-decomposition of a matrix and can compute solely the dominant eigenvalue and corresponding eigenvector. It starts with an initial vector b_0, and then computes iteratively $b_{k+1} = \frac{Ab_k}{||Ab_k||}$. If matrix A has an eigenvalue that is strictly larger than the rest and if the initial vector b_0 has a non-zero component in the direction of the dominant eigenvector, then the rate of convergence of b_k will be determined by $\frac{|\lambda_2|}{|\lambda_1|}$, where λ_1 is the dominant in magnitude eigenvalue of A and λ_2 is the second in magnitude eigenvalue. The larger the eigengap between $|\lambda_2|$ and $|\lambda_1|$, the faster the convergence.

Based on L', the power method can be used to derive the ordering solution. The power method will converge with rate $\frac{\lambda_3}{\lambda_2}$, where λ_2 is the second eigenvalue of L (and thus the dominant eigenvalue of L') and λ_3 is the third eigenvalue of L (and thus the second eigenvalue of L').

This analysis illustrates that the convergence of the power method for computing the ordering solution depends on the eigengap between the second and the third eigenvalue of the Laplacian matrix. A method that successfully enlarges this eigengap will increase the efficiency of the power method.

5 Elements of linear algebra

In order to study the behavior and the properties of the proposed spectral ordering framework, we need to recall certain elements of linear algebra. Firstly, we recall Weyl's theorem on the perturbation of eigenvalues.

Theorem 5.1 (Weyl, can be found in [20]) *Let A be a symmetric matrix with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and E a symmetric perturbation with eigenvalues $\epsilon_1 \geq \epsilon_2 \geq \cdots \geq \epsilon_n$. Then for $i = 1, \ldots, n$ the eigenvalues $\overline{\lambda}_i$ of $A + E$ will lie in the interval* $[\lambda_i + \epsilon_n, \lambda_i + \epsilon_1]$.

Another theorem we will employ is concerned with the affect of rank-k updates to matrix eigenvalues.

Theorem 5.2 (Wilkinson [23], can also be found in [18]) *Suppose $B = A + \tau \cdot uu^T$ where $A \in \mathbb{R}^{n \times n}$ is symmetric, $u \in \mathbb{R}^n$ has unit Euclidean norm and $\tau \in \mathbb{R}$. Then, there exist*

$m_1, \ldots, m_n \geq 0$, $\sum_{i=1}^{n} m_i = 1$, *such that*

$$\lambda_i(B) = \lambda_i(A) + m_i \tau, \quad i = 1, \ldots, n$$

Moreover, concerning rank-k updates $B = A + \sum_{i=1}^{k} \tau_i \cdot uu^T$, *there exist* $m_{ij} \geq 0$, $i = 1, \ldots, n$, $j = 1, \ldots, k$ *with* $\sum_{i=1}^{n} m_{ij} = 1$, *such that*

$$\lambda_i(B) = \lambda_i(A) + \sum_{j=1}^{k} m_{ij} \tau_j, \quad i = 1, \ldots, n.$$

6 Proposed spectral ordering framework

As we have mentioned in the introductory section, the proposed framework considers partial supervision and feature selection with the general aim of stabilizing the spectral ordering results. In this section we will present each component of the framework and demonstrate their contribution to the stability of the results. Recall that in Sect. 3 we have stated that stability essentially depends on two factors, namely the size of the relevant eigengaps as well as the uncertainty associated with the Laplacian matrix estimates. In the subsequent sections it is analytically demonstrated that the semi-supervised component is associated with the enlargement of the eigengaps, while the feature selection is concerned with the reduction of uncertainty.

6.1 Semi-supervised framework

The semi-supervised component assumes that an input ordering of the objects is provided and aims at adjusting the original object similarities such that the input ordering is taken into account. Recall that the original object similarities are used for computing the Laplacian matrix $L = D^{-1/2} W D^{-1/2}$ (here $W(i, j)$ is the similarity between object i and j) that derives the ordering solution. The proposed method essentially aims at adjusting the values of the W matrix based on the input ordering.

In order to achieve this goal, we initially construct a Laplacian matrix that produces the input ordering, i.e. whose second eigenvector gives the same result as the input order. If we consider the initial input ordering r (i.e. $r(i)$ is the order of object i) as a permutation of $\{1, 2, \ldots, n\}$, and a degree matrix D, we can define the initial input Laplacian as:

$$L_{\text{input}} = v_0 v_0^{\mathrm{T}} + \frac{1}{2} v_1 v_1^{\mathrm{T}}$$

where $v_0 = \frac{D^{1/2} e}{||(D^{1/2} e)||}$, with e being the unit vector (i.e. $e_i = 1$ for all i) and

$$v_1(i) = \frac{r(i) - \left(\sum_i r(i)\sqrt{d_i}\right) / \left(\sum_i \sqrt{d_i}\right)}{||r(i) - \left(\sum_i r(i)\sqrt{d_i}\right) / \left(\sum_i \sqrt{d_i}\right)||}, \tag{1}$$

with d_i being the ith diagonal element of the degree matrix D.

In order to understand the definition of the L_{input} matrix, one should initially observe that vector v_0 is essentially the largest eigenvector of any Laplacian matrix with degree matrix D (if there are no disconnected components). Moreover, vector v_1 is by construction orthogonal to v_0 and produces exactly the same ordering as r. Based on the above we can write L_{input} in the form of a Laplacian with degree matrix D, i.e. $L_{\text{input}} = D^{-1/2} W_{\text{input}} D^{-1/2}$, which has exactly two eigenvectors v_0 and v_1, with corresponding eigenvalues 1 and $\frac{1}{2}$. The W_{input}

matrix will contain the object similarities that generate the input ordering. Notice that this construction is possible for any degree matrix D.

It should also be noted that there exist different possible definitions of the v_1 eigenvector that are orthogonal to v_0 and also preserve the initial input order. However, the specific choice of v_1 imposes equal distances between the elements of the eigenvector v_1 and thus also on the "continuous" ordering solution between the objects. In the absence of further knowledge on the initial input ordering it would not be reasonable to impose the additional bias of unequal distances between the objects.

Based on the definition of L_{input} we derive the final Laplacian as a linear combination of the original data Laplacian (thereafter referred to as L_{data}) and L_{input} as:

$$L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$$

where $0 \leq c \leq 1$ is a confidence factor associated with each component of the summation. The behavior of L_{semi} can be understood if we write L_{semi} as:

$$L_{\text{semi}} = D^{-1/2}(cW_{\text{data}} + (1-c)W_{\text{input}})D^{-1/2}$$

which is possible since L_{input} is defined with the same degree matrix as L_{data}. This illustrates the main intuition of the semi-supervised framework that essentially adjusts the similarities of the original Laplacian such that the ordering is taken into account.

Intuitively one would expect that the use of supervision increases the reliability of the ordering results. This intuition is reflected in the eigengaps of L_{semi}. As demonstrated in the subsequent analysis, they can be enlarged with an appropriate choice of the c parameter, as compared to L_{data}.

6.2 Theoretical analysis of the semi-supervised framework

We will now analyze theoretically the behavior of the eigenvalues of L_{semi} with respect to the parameter c, the eigenstructure of L_{data} as well as the ordering solutions of L_{data} and L_{input}. In most theorems we derive the required amount of supervision (i.e. required value for $(1-c)$ or c) such that the desired eigenvalue bounds or eigengaps are achieved. We can summarize the theoretical results as follows:

- Theorem 6.1 demonstrates that the parameter c can fully control the eigenvalues of L_{semi}, almost independent of the structure of the Laplacians L_{data} and L_{input}.
- Theorem 6.2 demonstrates that if the eigenvalues of L_{data} are close to the bounds we wish to derive for the eigenvalues of L_{semi}, then these can be achieved with little supervision (i.e. small values for $(1-c)$).
- Theorems 6.3–6.5 demonstrate that the behavior of the eigenvalues depends also on the ordering solutions as derived by L_{data} and L_{input}. When the ordering solutions conform to a high degree, then the eigengaps are enlarged even with little supervision (i.e. small values for $(1-c)$).
- Theorem 6.6 demonstrates the dependency of the condition number of the spectral ordering problem with respect to the c parameter.

We will start with the dependence of the eigenvalues of L_{semi} with respect to the parameter c. The following theorem demonstrates that with an appropriate choice of the parameter c, large eigengaps can be achieved.

Theorem 6.1 *Let L_{data} be an $n \times n$ normalized Graph Laplacian, c a real number such that $0 \leq c \leq 1$ and L_{input} be the Laplacian as derived by an initial input ordering.*

Define the matrix $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$. *Its largest eigenvalue will be* $\lambda_1(L_{\text{semi}}) = 1$, *its second eigenvalue will reside in the interval* $\frac{1}{2} - \frac{c}{2} + c\lambda_n(L_{\text{data}}) \leq \lambda_2(L_{\text{semi}}) \leq \frac{1}{2} + \frac{c}{2}$, *where* $\lambda_n(L_{\text{data}})$ *is the smallest eigenvalue of matrix* L_{data}, *and its third eigenvalue will be smaller than c,* $\lambda_3(L_{\text{semi}}) \leq c$.

Proof In order to compute the appropriate bounds for the eigenvalues of L_{semi} we can employ Weyl's theorem on the matrices cL_{data}, $(1-c)L_{\text{input}}$ and $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$ and derive for the largest eigenvalue of L_{semi}, $\lambda_1(L_{\text{semi}})$:

$$\lambda_1(L_{\text{semi}}) \leq \lambda_1(cL_{\text{data}}) + \lambda_1((1-c)L_{\text{input}})$$

Based on the fact that $\lambda_1(cL_{\text{data}}) = c \cdot 1 = c$ (since the largest eigenvalue of L_{data} is 1) and $\lambda_1((1-c)L_{\text{input}}) = (1-c) \cdot 1$ (since the largest eigenvalue of L_{input} is 1) we can derive:

$$\lambda_1(L_{\text{semi}}) \leq 1.$$

Moreover for the first Laplacian eigenvector v_0, we have that $L_{\text{semi}} v_0 = [cL_{\text{data}} + (1-c)L_{\text{input}}]v_0 = cL_{\text{data}} v_0 + (1-c)L_{\text{semi}} v_0 = c \cdot v_0 + (1-c) \cdot v_0 = v_0$. Thus v_0 is an eigenvector of L_{semi} with corresponding eigenvalue 1. Thus

$$\lambda_1(L_{\text{semi}}) = 1.$$

Concerning the second eigenvalue of L_{semi} we can employ Weyl's theorem and state:

$$\lambda_2((1-c)L_{\text{input}}) + \lambda_n(cL_{\text{data}}) \leq \lambda_2(L_{\text{semi}}) \leq \lambda_2((1-c)L_{\text{input}}) + \lambda_1(cL_{\text{data}}).$$

It holds $\lambda_2((1-c)L_{\text{input}}) = (1-c)\frac{1}{2}$, $\lambda_n(cL_{\text{data}}) = c\lambda_n(L_{\text{data}})$ and $\lambda_1(cL_{\text{data}}) = c$. Thus,

$$\begin{aligned} &(1-c)\frac{1}{2} + c\lambda_n(L_{\text{data}}) \leq \lambda_2(L_{\text{semi}}) \leq (1-c)\frac{1}{2} + c \\ \Leftrightarrow\ &\frac{1}{2} - \frac{c}{2} + c\lambda_n(L_{\text{data}}) \leq \lambda_2(L_{\text{semi}}) \leq \frac{1}{2} + \frac{c}{2}. \end{aligned}$$

Concerning the third eigenvalue of L_{semi} we can employ Weyl's theorem and state:

$$\lambda_3(L_{\text{semi}}) \leq \lambda_3((1-c)L_{\text{input}}) + \lambda_1(cL_{\text{data}}).$$

We have $\lambda_3((1-c)L_{\text{input}}) = (1-c) \cdot 0 = 0$ (since L_{input} has only two non-zero eigenvalues) and $\lambda_1(cL_{\text{data}}) = c$. Thus

$$\lambda_3(L_{\text{semi}}) \leq c.$$ □

The bounds derived in the theorem above depend solely on the parameter c and illustrate that with an appropriate choice of c, large eigengaps can be achieved. However, if the eigengaps of matrix L_{data} are already large, then little supervision (i.e. smaller values of $(1-c)$) is required. The subsequent theorem illustrates this connection.

Theorem 6.2 *Let* L_{data} *be an* $n \times n$ *normalized Graph Laplacian, and* L_{input} *be the Laplacian as derived by an initial input ordering. Define the matrix* $L_{\text{semi}} = [cL_{\text{data}} + (1-c)L_{\text{input}}]$. *In order to derive an upper bound* $\overline{\lambda_2} \geq \frac{1}{2}$ *on the second eigenvalue of* L_{semi}, $\lambda_2(L_{\text{semi}}) \leq \overline{\lambda_2}$, *we must set* $c = \frac{\overline{\lambda_2} - \frac{1}{2}}{\lambda_2(L_{\text{data}}) - \frac{1}{2}}$. *In order to derive an upper bound on the third eigenvalue of* L_{semi}, $\lambda_3(L_{\text{semi}}) \leq \overline{\lambda_3}$, *we must set* $c \leq \frac{\overline{\lambda_3} + \overline{\lambda_2} - \frac{1}{2}}{\lambda_3(L_{\text{data}}) + \lambda_2(L_{\text{data}}) - \frac{1}{2}}$.

Proof In order to apply Wilkinson's theorem, we consider that matrix L_{semi} is composed by a rank-2 update on matrix cL_{data}. We can write for the three largest eigenvalues of L_{semi}:

$$\lambda_1(L_{\text{semi}}) = c\lambda_1(L_{\text{data}}) + m_{11}(1-c) + m_{12}\frac{1-c}{2}$$
$$\lambda_2(L_{\text{semi}}) = c\lambda_2(L_{\text{data}}) + m_{21}(1-c) + m_{22}\frac{1-c}{2}$$
$$\lambda_3(L_{\text{semi}}) = c\lambda_3(L_{\text{data}}) + m_{31}(1-c) + m_{32}\frac{1-c}{2}$$

Since the largest eigenvalue of L_{semi} is equal to 1, we have: $\lambda_1(L_{\text{semi}}) = 1 \Rightarrow c\lambda_1(L_{\text{data}}) + m_{11}(1-c) + m_{12}\frac{1-c}{2} = 1 \Rightarrow c + (m_{11} + \frac{m_{12}}{2})(1-c) = 1 \Rightarrow m_{11} + \frac{m_{12}}{2} = 1$.

Moreover, we have $\sum_{i=1}^{n}(m_{i1} + \frac{m_{i2}}{2}) = 1 + \frac{1}{2} \Rightarrow m_{11} + \frac{m_{12}}{2} + \sum_{i=2}^{n}(m_{i1} + \frac{m_{i2}}{2}) = 1 + \frac{1}{2} \Rightarrow \sum_{i=2}^{n}(m_{i1} + \frac{m_{i2}}{2}) = \frac{1}{2}$.

Thus $m_{21} + \frac{m_{22}}{2} \leq \frac{1}{2}$.

Now for the second eigenvalue we can write:

$$\lambda_2(L_{\text{semi}}) = c\lambda_2(L_{\text{data}}) + m_{21}(1-c) + m_{22}\frac{1-c}{2} \leq c\lambda_2(L_{\text{data}}) + \frac{1-c}{2}.$$

Recall that we aim at determining the appropriate c such that the upper bound $\overline{\lambda_2}$ is achieved. Thus we have:

$$c\lambda_2(L_{\text{data}}) + \frac{1-c}{2} = \overline{\lambda_2} \Rightarrow c = \frac{\overline{\lambda_2} - \frac{1}{2}}{\lambda_2(L_{\text{data}}) - \frac{1}{2}}.$$

In order to derive the appropriate bound for the third eigenvalue we should initially observe that $m_{21} + \frac{m_{22}}{2} = \frac{\lambda_2(L_{\text{semi}}) - c\lambda_2(L_{\text{data}})}{1-c}$.

Thus, $\sum_{i=3}^{n}(m_{i1} + \frac{m_{i2}}{2}) = \frac{1}{2} - \frac{\lambda_2(L_{\text{semi}}) - c\lambda_2(L_{\text{data}})}{1-c} \Rightarrow m_{31} + \frac{m_{32}}{2} \leq \frac{1}{2} - \frac{\lambda_2(L_{\text{semi}}) - c\lambda_2(L_{\text{data}})}{1-c}$.

Now for the third eigenvalue we can write:

$$\lambda_3(L_{\text{semi}}) = c\lambda_3(L_{\text{data}}) + m_{31}(1-c) + m_{32}\frac{1-c}{2} \leq c\lambda_3(L_{\text{data}}) + \frac{1-c}{2} - \lambda_2(L_{\text{semi}}) + c\lambda_2(L_{\text{data}}).$$

Recall that we aim at determining the appropriate c such that the upper bound $\overline{\lambda_3}$ is achieved. Thus we have:

$$c\lambda_3(L_{\text{data}}) + \frac{1-c}{2} - \lambda_2(L_{\text{semi}}) + c\lambda_2(L_{\text{data}}) = \overline{\lambda_3} \Rightarrow$$
$$c = \frac{\overline{\lambda_3} + \lambda_2(L_{\text{semi}}) - \frac{1}{2}}{\lambda_2(L_{\text{data}}) + \lambda_3(L_{\text{data}}) - \frac{1}{2}} \leq \frac{\overline{\lambda_3} + \overline{\lambda_2} - \frac{1}{2}}{\lambda_2(L_{\text{data}}) + \lambda_3(L_{\text{data}}) - \frac{1}{2}}$$

□

The derived c for the second eigenvalue is meaningful when the desired upper bound $\overline{\lambda_2(L_{\text{semi}})}$ is smaller than $\lambda_2(L_{\text{data}})$, and when both are larger that $\frac{1}{2}$, as this ensures that $c \in [0, 1]$. This is a natural setup because in order to achieve stability one should lower the second eigenvalue, as this will enlarge the eigengap between the first eigenvalue (which is always equal to 1) and the second. Concerning the derived c for the third eigenvalue, it is meaningful (i.e. $c \in [0, 1]$), when $\overline{\lambda_3(L_{\text{semi}})}$ is smaller than $\lambda_3(L_{\text{data}})$.

One would generally expect the behavior of the $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$ matrix to also depend on the eigenvectors of L_{data} and L_{input} and not solely on the eigenvalues.

It would be intuitive to consider that when the ordering solutions as derived by L_{data} and L_{input} conform to a high degree, then even with little supervision (i.e. small values of $(1-c)$), the reliability of the ordering results is rapidly increased. This is demonstrated in the following theorems.

Theorem 6.3 (Best case scenario) *Let $L_{\text{data}} = v_0 v_0^T + \lambda_2 v_2 v_2^T + \cdots + \lambda_n v_n v_n^T$ be the data Laplacian matrix and $L_{\text{input}} = v_0 v_0^T + \frac{1}{2} v_1 v_1^T$. If the ordering solution as derived by the second eigenvector of L_{data} is equal to the provided supervision $v_2 = v_1$, then the eigenvalues of matrix $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$ will be $\lambda_1(L_{\text{semi}}) = 1$, $\lambda_2(L_{\text{semi}}) = c\lambda_2(L_{\text{data}}) + \frac{1-c}{2}$, and $\lambda_i(L_{\text{semi}}) = c\lambda_i(L)$, for $i = 3, \ldots, n$. Moreover, the required supervision for achieving the eigengap $\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}}) = gap$, is $c = \frac{1/2-gap}{\lambda_2(L_{\text{data}})-1/2}$, and the required supervision for achieving the eigengap $\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}}) = gap$, is $c = \frac{1/2-gap}{1/2-(\lambda_2(L_{\text{data}})-\lambda_3(L_{\text{data}}))}$.*

Proof We have that the original data Laplacian is decomposed as $L_{\text{data}} = v_0 v_0^T + \lambda_2 v_2 v_2^T + \cdots + \lambda_n v_n v_n^T$ and $L_{\text{input}} = v_0 v_0^T + \frac{1}{2} v_2 v_2^T$ (since the two matrices induce the same order solution, i.e. $v_2 = v_1$). Thus:

$$\begin{aligned} L_{\text{semi}} &= cL_{\text{data}} + (1-c)L_{\text{input}} \\ &= v_0 v_0^T + \left(c\lambda_2(L_{\text{data}}) + \frac{1-c}{2}\right) v_2 v_2^T + c\lambda_3(L_{\text{data}}) v_3 v_3^T + \cdots + c\lambda_n(L_{\text{data}}) v_n v_n^T. \end{aligned}$$

Based on the above, we can derive the required c value as:

$$\begin{aligned} &\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}}) = gap \Rightarrow 1 - c\lambda_2(L) - \frac{1-c}{2} = gap \Rightarrow \\ &c = \frac{1/2 - gap}{\lambda_2(L_{\text{data}}) - 1/2} \end{aligned}$$

and

$$\begin{aligned} &\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}}) = gap \Rightarrow c\lambda_2(L_{\text{data}}) + \frac{1-c}{2} - c\lambda_3(L_{\text{data}}) = gap \Rightarrow \\ &c = \frac{1/2 - gap}{1/2 - (\lambda_2(L_{\text{data}}) - \lambda_3(L_{\text{data}}))}. \end{aligned}$$

□

On the other hand, when the initial input ordering solution corresponds to the eigenvector of L_{data} that is associated with the smallest eigenvector, then more supervision (i.e. larger values of $(1-c)$) is required.

Theorem 6.4 (Worst case scenario) *Let $L_{\text{data}} = v_0 v_0^T + \lambda_2 v_2 v_2^T + \cdots + \lambda_n v_n v_n^T$ be the data Laplacian matrix and $L_{\text{input}} = v_0 v_0^T + \frac{1}{2} v_1 v_1^T$. If the provided supervision is equal to the last eigenvector of the Laplacian matrix $v_1 = v_n$, then the eigenvalues of matrix $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$, will be $\lambda_1(L_{\text{semi}}) = 1$, $\lambda_n(L_{\text{semi}}) = c\lambda_n(L_{\text{data}}) + \frac{1-c}{2}$ and the rest will be $\lambda_i(L_{\text{semi}}) = c\lambda_i(L_{\text{data}})$, for $i = 2, \ldots, n-1$. Moreover, the required supervision for achieving the eigengap $\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}}) = gap$, is $c = \frac{1/2-gap}{1/2+(\lambda_2(L_{\text{data}})-\lambda_n(L_{\text{data}}))}$.*

Proof We have that $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}} = v_0 v_0^T + c\lambda_2(L_{\text{data}}) v_2 v_2^T + c\lambda_3(L_{\text{data}}) v_3 v_3^T + \cdots + (c\lambda_n(L_{\text{data}}) + \frac{1-c}{2}) v_n v_n^T$. Thus the eigengap between the second and the third eigenvalue will steadily become smaller as supervision increases (i.e. $(1-c)$ increases), until the eigenvalue corresponding to the eigenvector v_n gets larger than $c\lambda_2(L_{\text{data}})$. Based on the above, we can derive the required c value as:

$$\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}}) = gap \Rightarrow c\lambda_n(L_{\text{data}}) + \frac{1-c}{2} - c\lambda_2(L_{\text{data}}) = gap \Rightarrow$$
$$c = \frac{1/2 - gap}{1/2 - (\lambda_n(L_{\text{data}}) - \lambda_2(L_{\text{data}}))}.$$

□

In general, we can express the initial input ordering solution (i.e. the second eigenvector of L_{input}) as a linear combination of the eigenvectors of L_{data}. Based on this decomposition, it would be intuitive to expect that the eigenvectors that do not participate in the input ranking solutions are downgraded in importance. This is demonstrated in the subsequent theorem.

Theorem 6.5 *Let* $L_{\text{data}} = v_0v_0^T + \lambda_2 v_2 v_2^T + \cdots + \lambda_n v_n v_n^T$ *be the data Laplacian matrix and* $L_{\text{input}} = v_0v_0^T + \frac{1}{2}v_1v_1^T$. *Write* v_1 *as a linear combination of the eigenvectors of the data Laplacian matrix,*[3] $v_1 = w_2v_2 + w_3v_3 + \cdots + w_nv_n$. *Then the eigenvalues of* $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$ *are* $\lambda_1(L_{\text{semi}}) = 1$, *and* $\lambda_i(L_{\text{semi}}) = c\lambda_i(L_{\text{data}})$ *for all* i *such that* $w_i = 0$.

Proof We have that $L_{\text{input}} = v_0v_0^T + \frac{1}{2}v_1v_1^T = v_0v_0^T + \frac{1}{2}(w_2v_2 + w_3v_3 + \cdots + w_nv_n)(w_2v_2^T + w_3v_3^T + \cdots + w_nv_n^T)$. It is evident that for those v_i such that $w_i = 0$ we will have $L_{\text{input}}v_i = 0$. Thus, $L_{\text{semi}}v_i = c\lambda_i(L_{\text{data}})v_i$. □

This theorem signifies that the eigenvectors that do not participate in the input ranking solution will be quickly downgraded in importance (through the shrinkage of their eigenvalues), while the rest will finally converge to v_1. The same effect will take place concerning the eigenvectors that have small significance in the solution (i.e. $w_i \approx 0$).

The following theorem is concerned with the condition number of the spectral ordering problem versus the eigengaps and the level of supervision.

Theorem 6.6 *Let* L_{data} *be an* $n \times n$ *normalized Graph Laplacian,* c *a real number such that* $0 \le c \le 1$ *and* L_{input} *be the Laplacian as derived by an initial input ordering. Define the matrix* $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$. *The condition number of the spectral ordering problem (i.e. the second eigenvector) of* L_{semi} *is* $\kappa = \max\{\frac{1}{\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}})}, \frac{1}{\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}})}\}$. *Moreover,*

$if \kappa = \frac{1}{\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}})}$ *then* $\frac{2}{1+c-2c\lambda_n(L_{\text{data}})} \le \kappa \le \frac{2}{1-c}$;
$if \kappa = \frac{1}{\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}})}$ *then* $\kappa \ge \frac{2}{1+c}$;
$if \kappa = \frac{1}{\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}})}$ *and* $c < \frac{1}{3-2\lambda_n(L_{\text{data}})}$ *then* $\kappa \le \frac{2}{1-3c+2c\lambda_n(L_{\text{data}})}$.

Proof Recall that in the case of a Hermitian matrix the condition number of an eigenvector with corresponding eigenvalue λ is defined as $\kappa = \frac{1}{\min\_\text{eigengap}}$, where min _eigengap is the minimum eigengap between λ and the rest of the eigenvalues. Thus, in the case of the second eigenvector of L_{semi}, the minimum eigengap is determined by $\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}})$ and $\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}})$, thus $\kappa = \max\{\frac{1}{\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}})}, \frac{1}{\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}})}\}$.

[3] This is always possible since $\{u_2, \ldots, u_n\}$ are orthogonal to v_0 and to each other, thus forming a basis for every vector that is orthogonal to v_0.

In order to derive the range of values that κ will assume, recall that in Theorem 6.1 we have shown that $\lambda_1(L_{\text{semi}}) = 1$ and $\frac{1}{2} - \frac{c}{2} + c\lambda_n(L_{\text{data}}) \leq \lambda_2(L_{\text{semi}}) \leq \frac{1}{2} + \frac{c}{2}$. Based on these inequalities we can derive that

$$\frac{2}{1+c-2c\lambda_n(L_{\text{data}})} \leq \frac{1}{\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}})} \leq \frac{2}{1-c}.$$

Also, in Theorem 6.1 we have shown that $\lambda_3(L_{\text{semi}}) \leq c$. Thus we can derive that when $c < \frac{1}{3-2\lambda_n(L_{\text{data}})}$ (which implies that $1 - 3c + 2c\lambda_n(L_{\text{data}}) > 0$) it holds that

$$\frac{2}{1+c} \leq \frac{1}{\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}})} \leq \frac{2}{1-3c+2c\lambda_n(L_{\text{data}})}.$$

In the case $c \geq \frac{1}{3-2\lambda_n(L_{\text{data}})}$, we can solely prove that

$$\frac{2}{1+c} \leq \frac{1}{\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}})}.$$

□

Recall that the condition number κ of the spectral ordering problem essentially provides us with an upper bound on the error of the ordering solution when an E perturbation is applied on the Laplacian matrix, i.e. $||u - \tilde{u}|| \leq \kappa||E||$. Here u and $\tilde{u}$ are eigenvectors of L and $L + E$, respectively. It can be observed that when $c \to 1$, no upper bound on the error can be induced by the inequalities derived in Theorem 6.6. This is an expected behavior as when $c \to 1$, κ depends mostly on the eigengaps of the data-Laplacian matrix. However, as c assumes smaller values the solution becomes more heavily biased by the input Laplacian L_{input} and the upper bounds rapidly decrease. This is also observed in the experiments.

6.3 Quantification of uncertainty

As we have analyzed in Sect. 3, an integral component of stability assessment is the quantification of uncertainty in the form of an error-perturbation matrix E. Since we have already defined three matrices in the previous section (L_{data}, L_{input} and L_{semi}), we will need to define an appropriate perturbation matrix E for each. We will begin with L_{input} that is associated with the initial input ordering. Suppose we can characterize the degree of reliability in the supervision by comparing two rankings produced by the domain knowledge: if these are close to each other, then the domain knowledge is reliable. For both rankings we generate a corresponding eigenvector as was described in Sect. 6.1, and the difference between these vectors will be denoted as $v = u_1 - u_2$ where u_1 and u_2 are the two ranking eigenvectors. The element $v(i)$ gives the uncertainty related to object i. The input perturbation matrix E_{input} is a rank-1 matrix

$$E_{\text{input}} = 1/2vv^{\mathrm{T}}. \tag{2}$$

We will define the error-perturbation matrix for the L_{data} matrix in a way that will enable feature selection for uncertainty reduction. We initially observe that the order solution of $L_{\text{data}} = D^{-1/2}WD^{-1/2} = D^{-1/2}X^{\mathrm{T}}XD^{-1/2}$ can be derived by $D^{-1/2}X^{\mathrm{T}}u_2$, where u_2 is the second eigenvector of the "feature Laplacian" $L_{\text{feat}} = XD^{-1}X^{\mathrm{T}}$. Notice that L_{data} and L_{feat} have the same eigenvalues and if u is an eigenvector of L_{feat}, then $D^{-1/2}X^{\mathrm{T}}u$ is an eigenvector of L_{data}. Thus the stability of the ordering solution can be derived by the stability of the L_{feat} matrix. In order to quantify the uncertainty associated with the elements of L_{feat}, we bootstrap the observations and produce bootstrap confidence intervals for the elements of the L_{feat} matrix (pair-wise feature similarities). Consequently, we define matrix E_{data} such that $E_{\text{data}}(i, j)$ is the maximum difference between $L_{\text{feat}}(i, j)$ and the endpoints of the respective confidence interval.

The error-perturbation matrix of L_{semi} is derived by the norms of the matrices that take part in the summation. More precisely, we define

$$||E_{\text{semi}}||_2 = c||E_{\text{data}}||_2 + (1 - c)||E_{\text{input}}||_2. \tag{3}$$

Having defined all the appropriate error-perturbation matrices, we can move on to evaluate the stability of the spectral ordering framework and explore possible approaches for uncertainty reduction.

6.4 Feature selection for uncertainty reduction

Based on the definition of E_{data} as the perturbation of a *feature* × *feature* matrix, we can consider feature selection for uncertainty reduction. The proposed framework is similar in spirit to [16], where the features that contribute maximally to the norm of E_{data} matrix are sequentially removed. More precisely, at each step of the algorithm, the feature that corresponds to the column (or row) of matrix E_{data} that has the highest norm is removed. Although we employ feature selection in the same manner as in [16], we should stress that there are some important differences. The main difference is concerned with the fact that the new perturbation matrix E'_{data}, as induced by the removal of a feature, will not be a principal submatrix of E_{data}. This is because the removal of a feature will influence the values of the degree matrix D, thus affecting the confidence intervals of all the feature-pairs. In order to address this issue, we recompute the confidence intervals and E_{data} matrix after each feature is removed. However, it should be noted that when there is a large number of features, we can expect that the degree matrix is not severally affected and thus we can consider the principal submatrix of E_{data} (after removing the row and column i that corresponds to the removed feature) as an accurate approximation of the new perturbation matrix E'_{data}. When this is the case, it is guaranteed that the uncertainty as expressed by the norm $||E_{\text{data}}||_2$ will be reduced.

7 Related work

Concerning the semi-supervised component, our work is conceptually related to Pagerank [4]. Pagerank is considered as one of the top algorithms is Data Mining [24] and aims at deriving the stationary probability of the random walk based on a weighted linear combination of the transition matrix and a random jump or prior knowledge matrix, in the form of $A = [cP + (1 - c)S]^{\mathrm{T}}$, where P is the row-stochastic transition matrix and $S = eu^{\mathrm{T}}$, where u contains the random jump component or the prior distribution. Apart from the intuitive probabilistic interpretation of the A matrix, it has been shown that parameter c can control the eigengap between the largest and the second eigenvalue.

Theorem 7.1 (Haveliwala and Kamvar [12]) *Let P be an $n \times n$ row-stochastic matrix. Let c be a real number such that $0 \leq c \leq 1$. Let S be the $n \times n$ rank-one row-stochastic matrix $S = eu^{\mathrm{T}}$, where e is the n-vector whose elements are all $e_i = 1$ and u is an n-vector that represents a probability distribution. Define the matrix $A = [cP + (1 - c)S]^{\mathrm{T}}$. Its second eigenvalue is $|\lambda_2| \leq c$.*

As it illustrated in Sect. 6.1, the proposed semi-supervised approach essentially biases the original object-similarities such that the input ordering is taken into account. Under this view, we can consider our work as related to other frameworks that aim at learning the object similarity matrix for spectral learning (such as [3] and references therein). Conceptually closer to our approach is the work of Meila et al. [17], where the eigengap is explicitly used for constructing the appropriate objective function. More precisely, the authors consider the task of learning the object similarity matrix and define an optimization problem that maximizes the appropriate eigengap and minimizes a modified *MNCut* criterion. The need for a large eigengap is justified both by theoretical and empirical findings. Interestingly, although we do not explicitly require that the appropriate eigengap is maximized in our semi-supervised framework, this is achieved as a consequence of our L_{semi} construction process.

Concerning the feature selection component our work is conceptually related to Stability based Sparse PCA [16]. In this work the authors consider the use of feature selection for uncertainty reduction in the context of PCA, and demonstrate empirically that feature selection can stabilize the PCA results in several real-world UCI datasets.

We use results from matrix perturbation theory [20], stating that the rank-k approximation of a matrix A is close to a rank-k approximation of $A + E$, if E has weak spectral properties compared to those of A. Somewhat similar properties have been used in a different setting, namely speeding up SVD and kernel PCA: Achlioptas [1] shows how to choose the perturbation E based on the elements of the A matrix, such that the matrix $A + E$ is either a quantized or sampled version of A, making eigenvalue algorithms work faster.

The prospects of spectral ordering in the paleontological domain have been demonstrated by Fortelius et al. [9]. In this work, plain spectral ordering of the sites, based on mammal co-occurrences and discarding the age information of the sites, was considered. In addition, Puolamäki et al. [19] present a full probabilistic model that again only considers the co-occurrences in the data.

We have presented a semi-supervised approach for spectral ordering; the semi-supervision is realized by feeding an initial ordering into the process. The initial input ordering suggests which objects should stay close together and which objects should be placed far away in the final ordering. This is similar in spirit to constraint-based clustering in which the user provides pairwise constraints on some data objects, specifying whether they must or cannot be clustered to the same cluster.[4] Chen et al. [5] have presented a semi-supervised non-negative matrix factorization framework for clustering. In their approach, the user can provide “must-link” or “cannot-link” constraints on a few data objects.

Kalousis et al. [13] have studied the stability of feature selection algorithms. Similarly to our approach, they measure the sensitivity of the end result with respect to variations in the training set. Their problem setting is classification in high dimensional spaces, and the task is to select a small number of features that accurately classify the learning examples.

8 Empirical results

In the experiments we aim at verifying that the proposed framework enhances the stability of spectral ordering and increases the relevant eigengaps. Recall that this will increase the reliability of the ordering results and improve on the convergence rate of the power method. The experiments indeed verify the anticipated behavior.

[4] Remember that spectral ordering can be seen as continuous clustering.

8.1 Data sets

8.1.1 Paleontological data

The paleontological data we are considering consists of findings of land mammal genera in Europe and Asia 25 to 2 million years ago. The data set is stored and coordinated in the University of Helsinki, Department of Geology [8]. Our version of the data set was downloaded on June 26, 2007.

The observations in our data are the sites of excavation, and our features are mammal genera whose remains are found at these sites. In total we have 1,887 observations and 823 features. The data matrix is 0-1 valued: an entry $x_{ij} = 1$ means that mammal i was found at site j, and 0 otherwise. The data is very sparse: about 1 per cent of the entries are nonzero. We will also work with a small subset of data containing 1,123 observations and 18 features (the most common ones); this subset is more dense, having 12 per cent of its entries nonzero. Thereafter we will refer to the sparse dataset as $paleo_{sp}$ and the dense dataset as $paleo_d$.

In addition, we have auxiliary information on the estimated ages of the sites: an approximate age for each site, and also a more precise age for some sites; the methods available for estimating the ages vary from site to site, and thus at some sites the information is more certain than at others. The approximate ages will be used to construct an initial ranking r_{input} of the sites, and this will be used as an input in the semi-supervised setting, results of which will be presented in Sect. 8.2. Both the precise and approximate ages will be needed when quantifying our belief in the initial ranking, that is, defining the perturbation matrix E_{input} of L_{input} as discussed in Sect. 6.3; empirical results on this will be shown in Sect. 8.5.

We will assume that the data is fully connected in that the algebraic multiplicity of the first eigenvalue of the data Laplacian is 1. If this is not the case, the removal of disconnected observations will be a preprocessing step. In addition, we will preprocess the data such that almost-disconnected components are removed too: these correspond to objects that are very weakly connected to the rest of the objects. For such objects j, the value $r(j)$ in the order vector r (obtained by sorting the second eigenvector) is very large compared to other $r(j')$.

8.1.2 Newsgroup data

The other data set we will consider is a subset of the 20 Newsgroup corpus,[5] consisting of Usenet messages from four newsgroups ‘sci.crypt’, ‘sci.med’, ‘sci.space’ and ‘soc.religion.christian’. We have converted the documents into a binary term by document matrix using the Bow toolkit.[6] The 2,000 features of the data set consist of the most common terms in the documents, except for the stop words. The number of observations (documents, articles) is 3,987; there are 1,000 documents from each newsgroup, except for some empty documents that contain none of the 2,000 terms.

In the semi-supervised setting we will again need an input ordering that is either given by a domain expert or otherwise known. For newsgroup data, the input ordering is simply the time ordering of the documents in each newsgroup. The aim of the spectral ordering would be to reveal the flow of the discussion: who responds to whom. Spectral ordering is based on the co-occurrences of the terms, and documents belonging to the same discussion tend to share more

[5] http://www.cs.cmu.edu/~textlearning/

[6] http://www.cs.cmu.edu/~mccallum/bow/

terms than any two randomly chosen documents. The flow of the discussion is to some degree given by the time stamps of the documents, but not completely: people might respond to old documents, and there might be several discussions going on simultaneously. Thus our confidence in the input ordering is limited, similarly to what was the case in paleontological data.

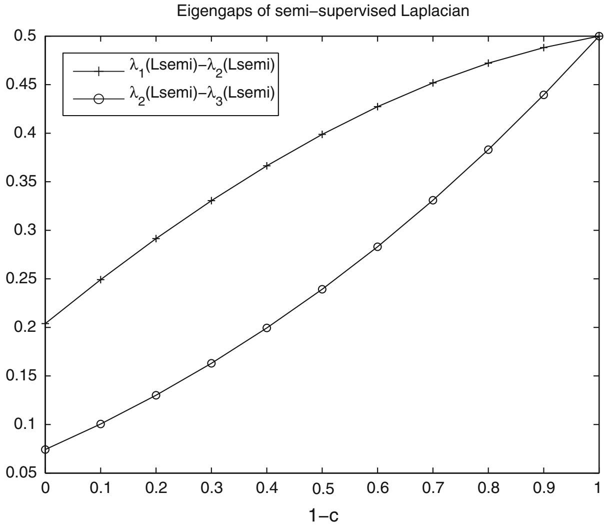
The newsgroup data is quite dense, as opposed to the sparse paleontological data, and there is no need to preprocess the data by removing disconnected components. In addition, we will see that the eigengaps are quite large in the original newsgroup data and the supervision cannot significantly increase them. However, there are other benefits in the semi-supervision and feature selection that we will demonstrate in the sequel.

8.2 Effect of supervision on the stability

Let us first demonstrate that the eigengaps of the data Laplacian increase when domain knowledge is taken into account. These experiments are performed on the sparse and large $paleo_{sp}$ dataset, where the initial eigengaps are small. Recall that the stability of spectral ordering essentially depends on two factors, one of which are the eigengaps between the first and second eigenvalue and the second and third eigenvalue of the Laplacian. Figure 1 shows the behavior of the eigengaps of the semi-supervised Laplacian $L_{\text{semi}} = cL_{\text{data}} + (1-c)L_{\text{input}}$ at a varying level of supervision. Choosing $1-c=0$ corresponds to no supervision, in which domain knowledge is not taken into account and the spectral ordering is done based on feature co-occurrences only; the eigengaps at $c=1$ thus show the eigengaps of the data Laplacian. In contrast, $1-c=1$ corresponds to the trivial case of full supervision of the ranking, in which co-occurrences in the data are not taken into account but only the domain knowledge ranking is used. We observe that both eigengaps increase rapidly when the level of supervision increases. Thus the spectral ordering becomes more stable as more emphasis is put on the domain knowledge.

For newsgroup data, the eigengaps are large in the original data, and we do not need to apply supervision to increase them. The data are dense, and the co-occurrences in the data give a quite stable spectral ordering.

Fig. 1 Eigengaps $\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}})$ (+) and $\lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}})$ (∘) versus the level of supervision. Horizontal axis: $1-c$, confidence in domain knowledge. $1-c=0$: no supervision; $1-c=1$: full supervision. Paleontological data



8.3 Computational gains of the supervision

As analyzed in Sect. 4.2, increasing the eigengap between the second and the third eigenvalue of the Laplacian matrix will enhance the convergence of the power method.

For any square matrix A whose dominant eigenvalues are $\lambda_1 > \lambda_2$, the rate of convergence of the power method is determined by $\frac{\lambda_2}{\lambda_1}$. The power method will output the first eigenvector of A. In spectral ordering we do not need the first eigenvector but instead the second eigenvector of the Laplacian matrix L. To take advantage of the power method, we apply a trivial transformation $L' = L - vv^{\mathrm{T}}$ where v is the first eigenvector of L, obtained easily from the degree matrix of the Laplacian, as demonstrated in Sect. 4.2. The first eigenvalue of L' will equal the second eigenvalue of L, and similarly for the eigenvectors. Thus applying the power method on L' will give us the spectral ordering solution. The rate of convergence of the power method on L' is dependent on $\frac{\lambda_3}{\lambda_2}$ where λ_2 and λ_3 are the second and third eigenvalues of L, and respectively equal to the first and second eigenvalue of L'. Thus increasing the eigengap $\lambda_2 - \lambda_3$ will speed up the power method.

Theorems 6.1 and 6.2 demonstrated that the eigengaps will depend on the amount of supervision, that is, value of $1 - c$. In this section we will show that the choice of c will indeed affect the number of iterations needed in the power method on two data sets: the large and sparse paleontological data set having 1,887 observations and 823 features, and the newsgroup data set having 3,987 observations and 2,000 features.

Figure 2 shows the number of iterations needed until convergence in the power method on the paleontological and newsgroup data sets. The power method was iterated until the Euclidean norm of the difference between two consecutive solutions of the dominant eigenvector was smaller than 10^{-15}. Changing this limit did not seem to make a difference in the pattern observed in the figures. The error bars show the standard deviation over 20 random initializations of the power method. The horizontal axis shows $1 - c$, the amount of supervision: the larger $1 - c$ is, the more we emphasize the input ordering given by a domain expert. We can see that as the supervision increases, the power method converges more easily. The pattern is quite strong even though the data sets are modest in size, and on larger data sets the computational gains are expected to be significant.

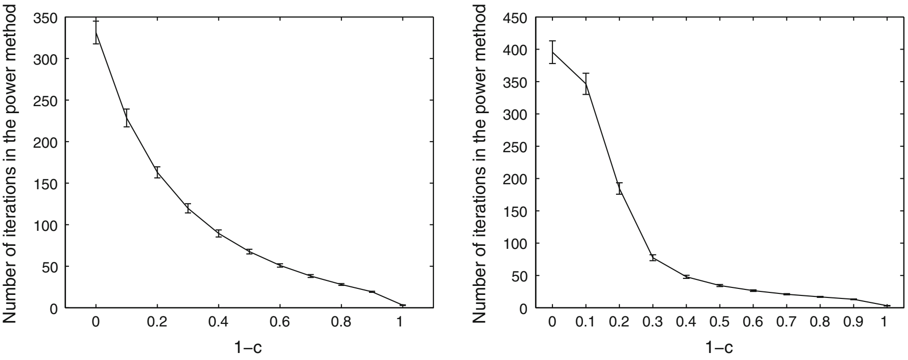


Fig. 2 Number of iterations needed in the power method versus the level of supervision. *Left*: paleontological data, *right*: newsgroup data. Horizontal axis: $1-c$, confidence in domain knowledge. $1-c = 0$: no supervision; $1 - c = 1$: full supervision. The error bars show the standard deviation over 20 random initializations

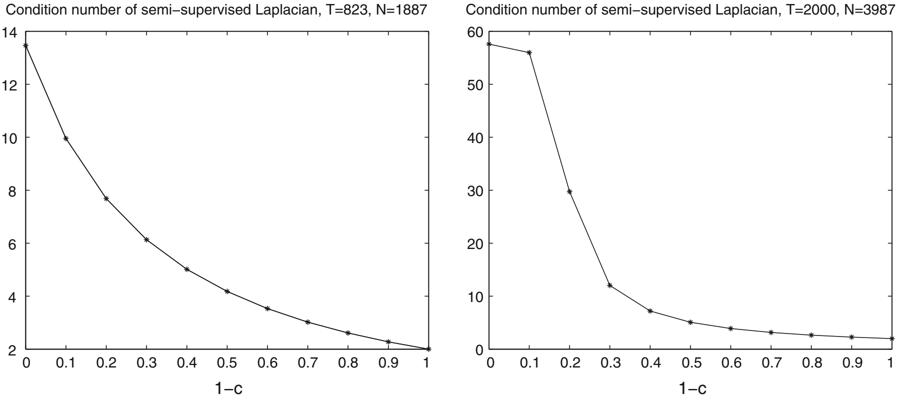


Fig. 3 Condition number of the spectral ordering problem versus the level of supervision. *Left*: paleontological data, *right*: newsgroup data. Horizontal axis: $1 - c$, confidence in domain knowledge. $1 - c = 0$: no supervision; $1 - c = 1$: full supervision

8.4 Condition number versus supervision

The condition number, measuring the sensitivity of the eigenvector with respect to small variations in the Laplacian, is a well defined tool to assess the stability of the spectral ordering problem. The smaller the condition number, the better. In Theorem 6.6 we have shown how the condition number κ is dependent on the eigengaps and the level of supervision. Figure 3 depicts the behaviour of the condition number as a function of $1 - c$: at $1 - c = 0$, no supervision is applied, and the figure shows the condition number of the data Laplacian. At $1 - c = 1$, the spectral ordering solution is dictated by the input ordering only and not by the data. We can see that the condition number decreases as the level of supervision increases. The data sets employed here are the paleontological data having 1,887 observations and 823 features, and the newsgroup data having 3,987 observations and 2,000 features.

8.5 Effect of feature selection on the stability

We will then demonstrate that the stability of the spectral ordering increases as features are removed step by step. The removed features will be chosen based on their contribution on the variability of the feature-feature similarity matrix, measured as matrix E_{data} discussed in Sect. 6.4. It should be noted that after each feature is removed, L_{semi} is reevaluated based on L_{input} and L_{data} which are appropriately recomputed.

We will measure the stability of the spectral ordering by a "stability factor" sf that depends on the eigengaps and the norm of the perturbation matrix:

$$sf_{\text{semi}} = \frac{\min(\lambda_1(L_{\text{semi}}) - \lambda_2(L_{\text{semi}}), \lambda_2(L_{\text{semi}}) - \lambda_3(L_{\text{semi}}))}{||E_{\text{semi}}||_2} \tag{4}$$

For the stability factor we will need appropriate values for L_{semi} and $||E_{\text{semi}}||_2$.

Let us first construct the semi-supervised Laplacian $L_{\text{semi}} = cL_{\text{data}} + (1 - c)L_{\text{input}}$. For this we need to carefully choose the confidence factor c reflecting our belief in the observed data versus the input ranking. We can either rely on a domain expert or better still, derive c from the body of domain knowledge: for the paleontological data we will choose to define

$$c = ||E_{\text{input}}||_2/(||E_{\text{data}}||_2 + ||E_{\text{input}}||_2) \tag{5}$$

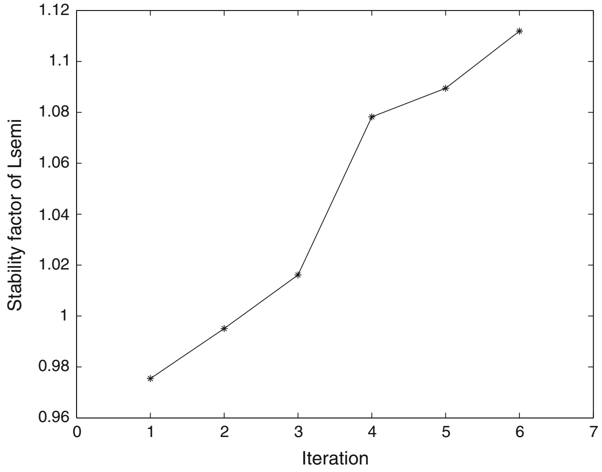


Fig. 4 Stability factor during feature selection. Horizontal axis: iteration. One feature is removed at each iteration. Paleontological data

which naturally characterizes the confidence such that a large perturbation in the initial input ranking leads to a high confidence in the observed data, and vice versa. In the definition (5), the data perturbation E_{data} will be obtained by bootstrap sampling as discussed in Sect. 6.3. The input perturbation E_{input} for paleontological data will be derived based on the availability of approximate or precise ages for each site: in addition to the initial ranking r_{input} based on approximate ages of the sites, we construct another initial ranking r_s using the precise ages available for some of the sites. (The sites for which a precise age is not available will get an average ranking in r_s.) For both rankings r_{input} and r_s we generate a corresponding eigenvector, v_{input} and v_s, using Eq. (1). We then take the difference between these eigenvectors as $v = v_{\text{input}} - v_s$ and use that in place of v in Eq. (2), to measure the difference between the two orderings. This gives us the perturbation E_{input} associated with the domain knowledge.

In the case of newsgroup data, we cannot objectively assess the quality of the input ranking, because we do not have several alternative rankings available as we did in the case of paleontological data above. The confidence factor c must thus be selected manually via experimentation.

Having now defined the value for c, we can construct the matrix $L_{\text{semi}} = cL_{\text{data}} + (1 - c)L_{\text{input}}$.

For the stability factor in Eq. (4) we also need the value for $||E_{\text{semi}}||_2$. Based on the definition for c in Eq. (5), Eq. (3) now simplifies

$$||E_{\text{semi}}||_2 = c||E_{\text{data}}||_2 + (1-c)||E_{\text{input}}||_2 = \frac{2||E_{\text{data}}||_2||E_{\text{input}}||_2}{||E_{\text{data}}||_2 + ||E_{\text{input}}||_2} \tag{6}$$

Having now defined all components of the stability factor (4) let us see how it behaves when features are iteratively removed. Figure 4 shows the results on paleontological data. We have employed the subset $paleo_d$ of 1,123 observations and 18 features in this experiment. At each iteration, one feature is removed based on its contribution to the data perturbation. Simultaneously, a few observations typically get removed, as they have become disconnected with the other observations due to the removal of the feature in question. The stability factor of the semi-supervised Laplacian increases during feature selection, showing that feature selection enhances the stability of spectral ordering.

In the results shown in Fig. 4, the value of the confidence parameter c was computed anew at each iteration. The value of c increased slightly but monotonically during the iterations:

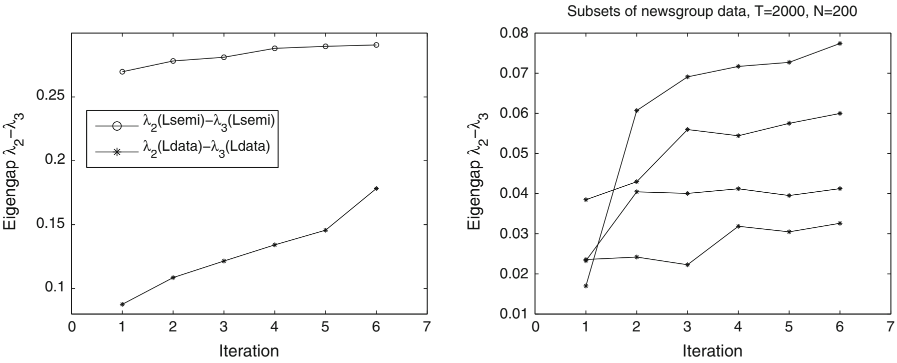


Fig. 5 Eigengap between the 2nd and 3rd eigenvalue during feature selection. One feature is removed at each iteration (horizontal axis). *Left*: Paleontological data. Semi-supervised spectral ordering (○), original spectral ordering (∗). *Right*: Newsgroup data, original spectral ordering. Each curve is a random subset having 200 observations

in the beginning, $c \approx 0.42$, and after the six iterations shown in the figure, $c \approx 0.46$. Recall that $c = 0$ would correspond to a perfect confidence in the domain expert opinion, and $c = 1$ a perfect confidence in the observed data, so the change in c corresponds to an increased confidence in the observed data. In feature selection, the feature that most contributes to $||E_{\text{data}}||$ is removed; this decreases the value of $||E_{\text{data}}||$. In the light of the definition of c in (5) the increase of c is now no surprise.

Again, for the newsgroup data we cannot in practice measure the value of the stability factor as the input perturbation $||E_{\text{input}}||$ is not available due to reasons discussed above: we do not have two or more alternative input rankings.

8.6 Computational gains of feature selection

Finally, let us describe a nice side-effect of feature selection. Recall that the eigengap between the second and third eigenvalue affects the convergence of the power method, as discussed in Sect. 4.2. Figure 5(left) shows that this eigengap increases during feature selection, both in the original paleontological data and in the semi-supervised setting. Thus the removal of "noisy" features can enhance the behaviour of the power method. The data set employed here is the smaller paleontological data having 1,123 observations and 18 features.

We also demonstrate the behaviour of the eigengap in the newsgroup data, without supervision. We have taken random subsets containing 200 observations. In each subset, the eigengap between the second and third eigenvalue increases during feature selection, as seen in Fig. 5(right).

This behaviour is not directly predicted by our theoretical analysis presented in the previous sections—but not prevented either. Further research is needed to find the theoretical justification of the findings in Fig. 5.

9 Discussion

In this paper, we have shown how to increase the stability of spectral ordering using two separate tools: partial supervision in the form of a (possibly uncertain) domain knowledge ordering, and feature selection.

We have presented a detailed theoretical analysis showing how the eigengaps of the Laplacian affect the stability, and how partial supervision will increase the eigengaps. The eigengaps are those between the first and second eigenvalue of the Laplacian, and similarly between the second and third eigenvalue. Feature selection in turn will decrease the norm of the perturbation matrix E that quantifies the uncertainty associated with the observed data.

Our main application area is paleontology: we have considered the ordering of the sites of excavation in paleontological data, by complementing spectral ordering with domain knowledge of the approximate ages of the sites. The paleontological data is noisy in that many observations are missing, and prone to small changes when the findings are more carefully examined. Also, we never have access to the exact ages of the sites. Thus when ordering the sites, the best we can aim at is an ordering that is as stable as possible with respect to small variations in the data. This motivates our task of optimizing the stability of spectral ordering. We have shown that in the paleontological data, the eigengaps quickly increase as semi-supervision is used. Also, feature selection, by removing the mammals that contribute most to the variation of the results in bootstrap sampling, is demonstrated to increase the stability of spectral ordering.

Another data set we have employed is newsgroup data in which the observations are newsgroup documents and the features are the most common terms. Although very different in nature, this data set shares the problem of noisy observations in some respect: In natural language documents, many terms are omitted although they would fit in the topic, as the documents are short and synonymous terms might have been used instead.

In order to illustrate the effect of the eigengaps on the stability of the ordering, we have also reported the condition number of the spectral ordering problem. This measure illustrates in a more explicit manner the dependence of stability on the size of the eigengap: if the eigengap tends to zero, the spectral ordering problem becomes ill-conditioned. Empirical results demonstrate that the condition number decreases as the amount of supervision increases in both paleontological and newsgroup data.

We have also shown that the supervision enhances the efficiency of spectral ordering when the power method is employed. This is demonstrated empirically in both application areas. The observed pattern is a direct consequence of the enlargement of the eigengap between the second and the third eigenvalue.

An interesting avenue for future research is to consider extrapolation methods for accelerating spectral ordering computations. Kamvar et al. [14] have shown how to accelerate PageRank computations and it might be possible to follow their approach.

In future work we also aim at exploring the potentials of our framework in different application domains, where partial supervision is naturally present. Moreover, we aim at extending the proposed framework to spectral clustering.

Acknowledgments The authors wish to thank professor Mikael Fortelius for fruitful discussions regarding paleontological data and professor Heikki Mannila for his insights in spectral ordering. E. Bingham was supported by Academy of Finland grant 118653 (ALGODAN).

References

1. Achlioptas D (2004) Random matrices in data analysis. In: Boulicaut J-F, Esposito F, Giannotti F, Pedreschi D (eds) Proceedings of the 15th European conference on machine learning (ECML), number 3201 in Lecture notes in computer science. Springer, Heidelberg, pp 1–7
2. Atkins JE, Boman EG, Hendrickson B (1998) A spectral algorithm for seriation and the consecutive ones problem. SIAM J Comput 28(1):297–310

3. Bach FR, Jordan MI (2006) Learning spectral clustering, with application to speech separation. J Mach Learn Res 7:1963–2001
4. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw 30(1–7):107–117
5. Chen Y, Rege M, Dong M, Hua J (2008) Non-negative matrix factorization for semi-supervised data clustering. Knowl Inf Syst 17(3):355–379
6. Ding CHQ, He X (2004) Linearized cluster assignment via spectral ordering. In: Brodley CE (ed) Proceedings of the 21st international conference on machine learning (ICML)', vol. 69 of ACM International Conference Proceeding Series. ACM, pp 233–240
7. Ding CHQ, He X, Zha H (2001) A spectral method to separate disconnected and nearly-disconnected web graph components. In: Proceedings of the 7th international conference on knowledge discovery and data mining (KDD), pp 275–280
8. Fortelius (coordinator) M (2007) Neogene of the Old World database of fossil mammals (NOW), University of Helsinki. http://www.helsinki.fi/science/now/
9. Fortelius M, Gionis A, Jernvall J, Mannila H (2006) Spectral ordering and biochronology of European fossil mammals. Paleobiology 32(2):206–214
10. Fortelius M, Werdelin L, Andrews P, Bernor RL, Gentry A, Humphrey L, Mittmann W, Viranta S (1996) Provinciality, diversity, turnover and paleoecology in land mammal faunas of the later Miocene of western Eurasia. In: Bernor R, Fahlbusch V, Mittmann W (eds) The Evolution of Western Eurasian Neogene Mammal Faunas. Columbia University Press, New York, pp 414–448
11. George A, Pothen A (1997) An analysis of spectral envelope reduction via quadratic assignment problems. SIAM J Matrix Anal Appl 18(3):706–732
12. Haveliwala T, Kamvar S (2003) The second eigenvalue of the Google matrix. Technical report, Stanford University. http://dbpubs.stanford.edu:8090/pub/2003-35
13. Kalousis A, Prados J, Hilario M (2007) Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inf Syst 12(1):95–116
14. Kamvar SD, Haveliwala TH, Manning CD, Golub GH (2003) Extrapolation methods for accelerating PageRank computations. In: Proceedings of the 12th international world wide web conference, pp 261–270
15. Li T (2008) Clustering based on matrix approximation: a unifying view. Knowl Inf Syst 17(1):1–15
16. Mavroeidis D, Vazirgiannis M (2007) Stability based sparse LSI/PCA: Incorporating feature selection in LSI and PCA. In: Kok JN, Koronacki J, de Mántaras RL, Matwin S, Mladenic D, Skowron A (eds) Proceedings of the 18th European conference on machine learning (ECML). Lecture notes in computer science, vol 4701. Springer, Heidelberg, pp 226–237
17. Meilă M, Shortreed S, Xu L (2005) Regularized spectral learning. In: Cowell RG, Ghahramani Z (eds) Proceedings of the Tenth international workshop on artificial intelligence and statistics (AISTATS). Society for Artificial Intelligence and Statistics, pp 230–237
18. Mika S (2002) Kernel Fisher discriminants. Ph.D. thesis, University of Technology, Berlin
19. Puolamäki K, Fortelius M, Mannila H (2006) Seriation in paleontological data using Markov Chain Monte Carlo methods. PLoS Comput Biol 2(2):e6
20. Stewart GW, Sun G-J (1990) Matrix perturbation theory. Academic Press, London
21. von Luxburg U (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416
22. von Luxburg U, Belkin M, Bousquet O (2008) Consistency of spectral clustering. Ann Stat 36(2): 555–586
23. Wilkinson JH (2004) The algebraic eigenvalue problem. Oxford University Press, New York
24. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng AFM, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37

Author Biographies

Dimitrios Mavroeidis received his B.Sc. degree in Mathematics from University of Athens, Greece, in 2001, his M.Sc. degree in Advanced computing from University of Bristol, UK, in 2002, and his Ph.D. in Computer Science from Athens University of Economics and Business, Greece, in 2009. His research interests include machine learning, data mining and spectral methods in data mining.

Ella Bingham received her M.Sc. degree in Engineering Physics and Mathematics at Helsinki University of Technology in 1998, and her Dr.Sc. degree in Computer Science at Helsinki University of Technology in 2003. She is currently at Helsinki Institute for Information Technology, located at the University of Helsinki. Her research interests include statistical data analysis and machine learning.