

---

# Multiway Cuts and Spectral Clustering

---

Marina Meilă and Liang Xu

University of Washington

mmp@stat.washington.edu, lxu@math.washington.edu

## Abstract

We look at spectral clustering as optimization. We show that near some special points called *perfect*, spectral clustering optimizes simultaneously two criteria: a dissimilarity measure that we call the *multiway normalized cut* (*MNCut*) and a cluster coherence measure that we call the *gap*. The immediate implication from the user's p.o.v is that spectral clustering will optimize any tradeoff between *MNCut* and *gap* which may explain its success in practice. Finally, we propose new methods for selecting  $K$  based on the gap and show their superior performance in experiments.

## 1 Introduction

Spectral clustering methods, i.e methods that use eigenvectors of a suitably chosen matrix to partition the data, have recently become popular for similarity-based tasks. Several new algorithms [1, 4, 10, 14] and practical applications [13, 3] have been published. The general belief that “spectral clustering works” is based on two kinds of results: On one side, proofs that if clusters are well separated (i.e very dissimilar), spectral clustering will be able to find the clusters [4, 10]. On the other side, accumulated evidence that spectral methods find good or acceptable clusterings as judged by human experts on a variety of real data sets (e.g image segmentation, information retrieval) and on artificial cases.

The latter results are encouraging, but they do not offer an explanation of what is a “good” clustering from the point of view of the spectral algorithm. The former results are highly predictive, but of restricted applicability. Many situations where spectral algorithms work well empirically do not have well separated clusters. We assume a more general situation in which spectral clustering is expected to work and for this case we set out to define what criteria are optimized by spectral clustering. First, we prove that spectral clustering optimizes a criterion we call the multiway normalized cut (*MNCut*). Second, we show that from the user's p.o.v spectral clustering simultaneously optimizes two different criteria (and therefore an infinity of combinations thereof) and discuss the implications for selecting the number of clusters  $K$ . Third, we use this result to derive criteria for selecting the number of clusters  $K$  and validate them by experiments.

## 2 Spectral clustering – notation and previous results

In spectral clustering, the data is a set of *similarities*  $S_{ij}$ , satisfying  $S_{ij} = S_{ji} \geq 0$ , between pairs of points  $i, j$  in a set  $V$ ,  $|V| = n$ . The matrix  $S = [S_{ij}]_{i,j \in V}$  is called the *similarity*

*matrix*. We denote by

$$D_i \equiv \text{Vol} \{i\} = \sum_{j \in V} S_{ij} \quad (1)$$

the *volume* of node  $i \in V$  and by  $D$  a diagonal matrix formed with  $D_i, i \in V$ . The volume of a set  $A \subseteq V$  is  $\text{Vol} A = \sum_{i \in A} D_i$ . W.l.o.g we assume that no node has volume 0.

**The random walks view** Many properties of spectral clustering are elegantly expressed in terms of the stochastic *transition matrix*  $P$  obtained by normalizing the rows of  $S$  to sum to 1.

$$P = D^{-1}S \quad \text{or} \quad P_{ij} = S_{ij}/D_i \quad (2)$$

This matrix can be viewed as defining a Markov random walk over  $V$ ,  $P_{ij}$  being the *transition probability*  $\text{Pr}[i \rightarrow j|i]$ . The eigenvalues of  $P$  are  $1 = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n \geq -1$  and the corresponding eigenvectors are  $v^1, \dots, v^n$ . Note that because  $S = DP$  is symmetric, the eigenvalues of  $P$  are real and the eigenvectors linearly independent. Define  $[\pi_i]_{i \in V}$  by

$$\pi_i = D_i/\text{Vol} V$$

It is easy to verify that  $P^T \pi = \pi$  and thus that  $\pi$  is a *stationary distribution* of the Markov chain. For a set  $A \subseteq V$ , we denote by  $\pi_A = \text{Vol} A/\text{Vol} V$  the probability of  $A$  under the stationary distribution.

**The NCut criterion** A clustering  $\Delta = \{C_1, \dots, C_K\}$  is defined as a partition of the set  $V$  into the disjoint nonempty sets  $C_1, \dots, C_K$ . The *normalized cut (NCut)* clustering criterion was introduced by [13] for clusterings with  $K = 2$  clusters.

$$\text{NCut}(\Delta) = \text{Cut}(C_1, C_2) \left( \frac{1}{\text{Vol} C_1} + \frac{1}{\text{Vol} C_2} \right) \quad (3)$$

where

$$\text{Cut}(A, B) = \sum_{i \in A} \sum_{j \in B} S_{ij} \quad (4)$$

As a criterion for clustering quality, the *NCut* balances the size of the cut with the sizes of the resulting clusters. It is virtually equivalent to the *Lovasz conductance* and to the *isoperimetric number* in a discrete setting (see e.g [6])

$$\psi(A, B) = \frac{\text{Cut}(A, B)}{\text{Vol} A \text{Vol} B}$$

as

$$\text{NCut}(A, B) = \psi(A, B) \text{Vol} V$$

In [13, 4] it is shown with a number of examples that minimizing the *NCut* agrees with the intuition of a good clustering much better than other criteria in wide use. It is also known [13] that finding the two-way clustering that minimizes the normalized cut is NP hard, but that in a special case, the optimum can be found by the following algorithm.

**A spectral clustering algorithm** The Shi-Malik (SM) algorithm [13] uses  $v^2$ , the second eigenvector of  $P$ , to partition the data set into two. Data point  $i$  is mapped to  $v_i^2 \in R$ ; the resulting set of points on the real axis is partitioned by thresholding. It has been shown [13] that when the elements of  $v^2$  satisfy

$$v_i^2 = \begin{cases} \alpha, & i \in C_1 \\ \beta, & i \in C_2 = V \setminus C_1 \end{cases}$$

the clustering  $\Delta = \{C_1, C_2\}$  minimizes the normalized cut. Moreover, for such a  $v^2$ , each cluster projects into a single point on the real axis and the SM algorithm will find the optimal clustering.

We call an  $n$ -dimensional vector  $v$  *piecewise constant* w.r.t a clustering  $\Delta$  if  $v_i = v_j$  whenever  $i, j$  are in the same cluster in  $\Delta$ . The work of [13] establishes an intriguing

connection between the optimal normalized cuts and piecewise constant eigenvectors (PCE for short) for the case  $K = 2$ . It is then natural to ask: Could one define a “ $K$ -way normalized cut” similar to the *NCut* defined in (3)? Can this be optimized by using one or more of the leading eigenvectors of  $P$ ? The next section will answer yes to both questions. In the meanwhile, we introduce another useful result about PCE.

**Stochastic matrices with piecewise constant eigenvectors** If a set of vectors  $v^1, \dots, v^K$  are piecewise constant w.r.t a clustering  $\Delta$  then the *spectral mapping*  $i \rightarrow (v_i^1, v_i^2, \dots, v_i^K)$  maps each cluster  $C_k \in \Delta$  into a unique point in  $R^K$ . Many spectral algorithms [16, 10, 4], SM included, follow this pattern. The vectors  $v^1, v^2, \dots, v^K$  are obtained from  $S$  by a process involving an eigen-decomposition. We call a similarity matrix  $S$  from which piecewise constant vectors result *perfect*. Note that being perfect is a function of the spectral mapping. If an  $S$  is perfect, then clustering the mapped data in  $R^K$  is extremely easy; moreover, by the continuity of eigenvectors the data will be easy to cluster in  $R^K$  in a neighborhood of a perfect  $S$ . In other words, if a matrix  $S$  is (nearly) perfect for a spectral algorithm and a clustering  $\Delta$ , the algorithm will be guaranteed to return  $\Delta$ . The following theorem establishes the necessary and sufficient conditions when this happens.

**Theorem 1 (Lumpability)**[9] *Let  $P$  be a matrix with rows and columns indexed by  $V$  that has independent eigenvectors. Let  $\Delta = (C_1, C_2, \dots, C_k)$  be a partition of  $V$ . Then,  $P$  has  $K$  eigenvectors that are piecewise constant w.r.t.  $\Delta$  and correspond to non-zero eigenvalues if and only if:*

1. for all  $k, l = 1, \dots, k$  the sums  $P_{ik} = \sum_{j \in C_k} P_{ij}$  are equal for all  $i \in C_l$ , and
2. the matrix  $\hat{P} = [\hat{P}_{kl}]_{k,l=1,\dots,K}$  (with  $\hat{P}_{kl} = \sum_{j \in C_k} P_{ij}$ ,  $i \in C_l$ ) is non-singular.

A stochastic matrix  $P$  satisfying the conditions of Theorem 1 is called *block stochastic*. Because of the lumpability theorem, block stochastic matrices will play a central role in this paper. Therefore, in the following, unless it is otherwise specified, a perfect  $S$  will denote an  $S$  that engenders a block stochastic  $P$  and  $v^1, v^2, \dots, v^K$  will represent the leading eigenvectors of  $P$ .

## 2.1 Some useful lemmas

This section groups other results that help one get more insights into the properties of matrices with PCE and that are used in the proofs. They can be skipped at first reading. The first two lemmas were proved in [9], while lemma 4 is proved in the Appendix.

The Laplacian [2] of  $S$  is defined as

$$L = I - D^{-1/2} S D^{-1/2} \tag{5}$$

where  $I$  is the unit matrix. There are strong ties between the Laplacian and  $P$ , as the following lemma shows.

**Lemma 2 (Relationship between the Laplacian and the Markov random walk transition matrix)** *Denote by  $1 = \lambda_1 \geq \lambda_2 \geq \dots \lambda_n \geq -1$  the eigenvalues of  $P$  and by  $v^1, \dots, v^n$  the corresponding eigenvectors. Denote by  $0 = \mu_1 \leq \mu_2 \leq \dots \mu_n$  the eigenvalues of  $L$  and by  $u^1, \dots, u^n$  the corresponding eigenvectors. Then,*

$$\mu_i = 1 - \lambda_i \tag{6}$$

$$u^i = D^{1/2} v^i \tag{7}$$

for all  $i = 1, \dots, n$ .

Note that this lemma ensures that the eigenvalues of  $P$  are always real and the eigenvectors linearly independent.

**Lemma 3 (Relationship between  $P$  and  $\hat{P}$ )** Assume that the conditions of Lemma 1 hold. Let  $v^1, \dots, v^K$  and  $1 = \lambda_1 \geq \lambda_2 \geq \dots \lambda_K$  be the piecewise constant eigenvectors of  $P$  and their eigenvalues. Denote by  $1 = \hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \hat{\lambda}_K$  and  $\hat{v}^1, \dots, \hat{v}^K$  the eigenvalues and eigenvectors of  $\hat{P}$ . Then for all  $k = 1, \dots, K$  we have

$$\hat{\lambda}_k = \lambda_k \quad (8)$$

$$\hat{v}_i^k = v_i^k \text{ for } l = 1, \dots, K \text{ and } i \in C_l \quad (9)$$

**Lemma 4 (A generalized Rayleigh quotient)** Let  $L$  be a symmetric  $n \times n$  positive definite matrix with eigenvalues  $0 = \mu_1 \leq \mu_2 \leq \dots \mu_n$  and corresponding eigenvectors  $u^1, \dots, u^n$ . Then

$$\min \sum_{k=1}^K (y^k)^T L y^k \text{ s.t. } (y^l)^T y^k = \delta_{kl}$$

equals  $\sum_{k=1}^K \mu_k$  and the minimizing  $y^1, \dots, y^K$  lie in the subspace spanned by  $u^1, \dots, u^K$ .

### 3 Spectral clustering optimizes the Multiway Normalized Cut

We define the *multiway normalized cut* ( $MNCut$ ) of a clustering  $\Delta = \{C_1, C_2, \dots, C_K\}$  by

$$MNCut(\Delta) = \sum_{k=1}^K \left( 1 - \frac{Cut(C_k, C_k)}{Vol C_k} \right) \quad (10)$$

Noting that  $Vol C_k = \sum_{k'=1}^K Cut(C_k, C_{k'})$ , we obtain the alternate expression for  $MNCut$

$$MNCut(\Delta) = \sum_{k=1}^K \sum_{k'=k+1}^K NCut(C_k, C_{k'}) \quad (11)$$

The definition of  $MNCut$  is best motivated by the Markov random walk view. Define  $P_{AB} = Pr[A \rightarrow B|A]$  as the probability of the random walk going from set  $A \subset V$  to set  $B \subset V$  in one step if the current state is in  $A$  and the random walk is in its stationary distribution  $\pi$ .

$$P_{AB} = \frac{\sum_{i \in A, j \in B} \pi_i P_{ij}}{\pi_A} = \frac{\sum_{i \in A, j \in B} S_{ij}}{Vol A} = \frac{Cut(A, B)}{Vol A} \quad (12)$$

It follows that the multiway normalized cut represents the sum of the “out-of-cluster” transition probabilities at the cluster level.

$$MNCut(\Delta) = K - \sum_{k=1}^K P_{C_k C_k} = \sum_{k=1}^K \sum_{k \neq k'} P_{C_k C_{k'}} \quad (13)$$

If  $MNCut(\Delta)$  is small for a certain partition  $\Delta$ , then the probabilities of evading  $C_k$ , once the walk is in it, is small. For  $K = 2$ ,  $MNCut$  is equivalent to  $NCut$ . Just like its two way counterpart, the  $MNCut$  balances sizes of two way cuts with volumes of clusters. How appropriate is  $MNCut$  as a clustering criterion? We can use as indirect evidence the successes of spectral clustering so far; we can also look at motivating examples designed for  $NCut$  by [13] on which  $MNCut$  works as well as  $NCut$ . These suggest that the new criterion is reasonable and potentially useful in practice. Unfortunately, its optimization is in general intractable.

**Theorem 5** For any given  $K \geq 2$ , the  $K$ -way  $MNCut$  is NP hard to optimize.

Now we turn to the block stochastic  $P$  case and show that, in this situation, the  $MNCut$  can be minimized by a spectral algorithm. The following lemma paves the way and theorem 7 states the main result.

**Lemma 6** *If  $P$  is block stochastic w.r.t  $\Delta$  and  $v^1, \dots, v^K$  are its PCE, then*

$$MNCut(\Delta) = K - \sum_{k=1}^K \lambda_k \quad (14)$$

**Theorem 7 (Multicut Lemma)** *Let  $S$  be an  $n \times n$  symmetric matrix with nonnegative elements, and let  $P$  be as in (2). Assume that  $P$  has  $K$  PCE  $v^1, \dots, v^K$  w.r.t a partition  $\Delta^*$ ,  $|\Delta^*| = K$ . Denote the corresponding eigenvalues by  $\lambda_1, \dots, \lambda_K$  and assume that  $\lambda_1, \dots, \lambda_K$  are the  $K$  largest eigenvalues of  $P$ , are all non-zero, and  $\lambda_K > \lambda_{K+1}$ . Then the minimum  $K$ -way normalized cut for  $S$  is given by the partition  $\Delta^*$ .*

Hence, for a block stochastic  $P$ , the spectral mapping perfectly reflects the clustering  $\Delta^*$  that minimizes the multiway normalized cut. Any algorithm that uses this spectral mapping, as for example the Meila-Shi algorithm [15], will optimize  $MNCut$ .

Let us now examine other algorithms published in the literature. One of the most popular ones is the algorithm of Ng & al. [10]. This algorithm uses a different spectral mapping but, as it was shown in [15], the Ng & al.'s spectral mapping is perfect iff  $P$  is block stochastic (for the same  $K$  and  $\Delta$ ). Therefore, the Ng & al. algorithm implicitly minimizes the  $MNCut$ .

The following algorithms have been published in the image segmentation literature: the Perona-Freeman [11] algorithm's spectral mapping uses the first eigenvector of  $S$  and splits by finding the largest difference in the sorted values. For this algorithm,  $S$  is perfect when it is block stochastic. The Scott-Longuet-Higgins algorithm [12] uses the leading  $K$  eigenvectors of  $S$ , then constructs an  $n \times n$  matrix  $Q$  in a manner similar to the Ng & al algorithm and clusters based on the values of its elements.  $S$  is perfect for this algorithm if  $Q$  has only 1's or 0's. In this case too,  $P$  will have PCE [15]. In [16], a combination of the SM algorithm and Scott-Longuet-Higgins was published. For this algorithm as well, it can be shown that the perfect case implies a block stochastic  $P$ .

In general, noticing that if  $S$  satisfies the conditions of Theorem 1, then  $P$  is block stochastic, any algorithm that uses a spectral mapping based on the eigenvectors of  $S$  will be minimizing the  $MNCut$  at its perfect point.

For the recursive algorithms the case is less clear: it has been shown in [15] that a variant of the SM algorithm is equivalent to the Meila-Shi algorithm when  $P$  is block stochastic. There is no similar result for the standard version of the SM algorithm, nor for the "spectral algorithm II" in [4]. We do not know what are the conditions for a perfect  $S$  for these two algorithms. Experiments in [15] indicate that the SM algorithm behaves very similarly to the Meila-Shi and Ng & al algorithms near the perfect point, while "spectral algorithm II" does not (and in fact seems not to be optimizing the same function as the other algorithms). "Spectral algorithm I", and "spectral algorithm III" in the same paper are not directly comparable with the framework presented here.

Hence, a significant number of the spectral algorithms in the literature implicitly minimizes the  $MNCut$  at their perfect point. This entitles us from now on to simplify the exposition by talking about "spectral clustering", with the understanding that the term applies to those algorithms that in the perfect case are minimizing  $MNCut$ .

**Corollary 8** *For a clustering  $\Delta$  with  $K$  clusters  $MNCut(\Delta) \geq l(K) = K - \sum_{k=1}^K \lambda_k$ .*

The corollary follows from the proof of theorem 7. It provides a computable lower bound

$l(K)$  on the best *MNCut* obtainable in  $S$  for a given  $K$ . The bound holds for any  $S$  and  $K$  and is attained if  $S$  is perfect.

For  $K = 2$  the bound becomes  $l(2) = 1 - \lambda_2$  and is known (see e.g [2]); another, intractable, lower bound for  $K = 2$  is  $\phi(V)$  the *conductance* defined in section 5.

## 4 Spectral clustering is bicriterial

**The gap as a measure of cluster coherence** We define the *gap* to be the difference between a normalized cut and the lower bound

$$gap(\Delta) = MNCut(\Delta) - \left( K - \sum_{k=1}^K \lambda_k \right) \quad \text{if } |\Delta| = K \quad (15)$$

It is obvious and it may even seem redundant to state that, if  $S$  is perfect, then spectral clustering optimizes the gap. However, this section will argue that the gap behaves like a measure of distortion, showing spectral clustering in a new light.

**Lemma 9** Define  $x_i = [v_i^1 \dots v_i^K]^T$  the  $i$ 'th data point in the spectral mapping,  $c_k = \frac{1}{|C_k|} (\sum_{i \in C_k} x_i)$  the centroid of  $C_k \in \Delta$ , and  $dist(\Delta) = \sum_{C_k \in \Delta} \sum_{i \in C_k} \|x_i - c_k\|^2$  the distortion of  $\Delta$  in the spectral mapping. Then

$$dist(\Delta) < \epsilon \quad \Rightarrow \quad gap(\Delta) \leq \lambda_{K+1} \epsilon / \sigma$$

where  $\sigma$  is the smallest singular value of the  $K \times K$  matrix  $C = [c_1 \dots c_k]$ .

A converse result, bounding the distortion for small gap can also be proved using matrix perturbation theory. Hence, the gap is intimately related to the *coherence* of the clusters in the spectral mapping. We will denote coherence generically by  $coh(\Delta)$ ; in the next section we discuss and compare various definitions of coherence besides the one used in lemma 9. For now it suffices to assume that  $coh(\Delta) \geq 0$  with 0 (counterintuitively!) as optimal value attained whenever  $S$  is perfect.

Another interesting result reflecting on the ties between the gap and cluster coherence is proved in [7]: two clusterings  $\Delta, \Delta'$  with a sufficiently small gap, i.e  $gap(\Delta), gap(\Delta') \leq \epsilon < \lambda_K - \lambda_{K+1}$ , are close to each other. Intuitively, this result expresses the fact that two clusterings that are both coherent cannot "cut" each other.

**Spectral clustering is bicriterial** We see thus that it is equally correct to say that spectral clustering finds coherent clusters, when they exist, as it is to say that it finds a small multiway normalized cut, if one exists. But, from the point of view of a user, small *MNCut* and good coherence are two conceptually different goodness criteria. It is a special feature of spectral clustering near a perfect  $S$  to optimize both simultaneously. We can bank on this feature by noting that there are an infinite number of criteria  $f(MNCut, coh)$  that are minimized at a perfect  $S$ . In the vicinity of a perfect  $S$ , each will represent a different tradeoff between *MNCut* and *coh* and a user can choose the one that best suits her problem. This property of spectral clustering of being bicriterial may well account for its success in so many different situations.

Looking at a combined criterion  $f(MNCut, coh)$  has even greater potential as it allows us to meaningfully compare clusterings with different  $K$ , something that cannot be done by *MNCut* alone. This is the more significant because typically spectral clustering algorithms take the number of clusters  $K$  as given. In the remainder of the paper we focus on this topic. We leave the problem of defining a suitable, application-dependent  $f(MNCut, coh)$ , for future research and study only how to formulate practical, but if possible algorithm and application independent measures of coherence.

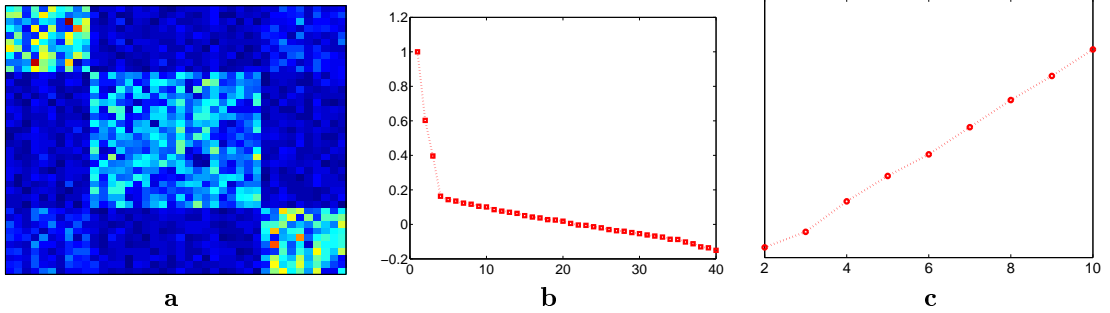


Figure 1: The criteria *eig* and *MNCut* can fail to select the correct  $K$ : (a) A block stochastic  $P$  with 3 clusters and small additive noise. The block structure is still clearly visible. (b) The eigenvalues of  $P$  in decreasing order, showing that the eigengap as a function of  $K$  chooses  $K_{eig} = 1$  (no clusters). (c) The *MNCut* as a function of  $K$  chooses  $K_{MNCut} = 2$ .

## 5 Coherence and the choice of $K$

In the following, we shall demonstrate that (1) the *MNCut* alone is unfit for comparing clusterings with different  $K$ , and (2) measures of coherence based on the gap are capable of detecting the best clustering in controlled experiments where we know that one exists. To stress these points, we introduce several measures of coherence and examine their performance as criteria for selecting the number of clusters  $K$ . We compare these with some criteria that do not use coherence (which will be listed below).

The experimental methodology is standard: We fix a data set for which an optimal clustering is known. Then we run a clustering algorithm with values for  $K$  in the range  $2, \dots, K_{max}$  obtaining clusterings  $\Delta_2, \dots, \Delta_{K_{max}}$ . From among these, we choose the “best”  $\Delta_K$  according to each of the criteria and record the selected  $K$ .

Before we move on to the experiments, we shall discuss some other possibilities for selecting  $K$ , given a set of clusterings  $\Delta_2, \dots, \Delta_{K_{max}}$  with different numbers of clusters.

The *eigengap* defined as  $eig(K) = \lambda_K - \lambda_{K+1}$  (or alternatively  $\lambda_K/\lambda_{K+1}$ ) was often suggested as a way to guess the number of clusters by  $K_{eig} = \operatorname{argmax} eig(K)$ . Note however that the eigengap, is unrelated to the eigenvectors and gives no indication about the coherence of the clustering. Figure 1 shows an very simple example where the eigengap fails to indicate the true number of clusters.

The *MNCut* is undoubtedly a measure of clustering quality according to our previous standpoint. But *MNCut* by itself, however useful it may be in choosing between different clusterings with the same  $K$ , is inappropriate for comparisons between clusterings with different numbers of clusters. In particular, *MNCut* tends to grow with  $K$  (see lemma 10), favoring clusterings with minimal number of clusters. Figure 1 illustrates this.

In [4], the coherence is defined as the minimum intra-cluster conductance  $coh(\Delta) = \min_{C_k \in \Delta} \phi(C_k)$  with the *conductance*  $\phi(C)$  being

$$\phi(C) = \min_{A \in C} \frac{Cut(A, C \setminus A)}{\min[Vol A, Vol V \setminus A]}$$

The conductance is NP-hard to compute so this criterion is impractical for the evaluations we perform here.

Recalling that spectral clustering algorithms map  $V$  to  $R^K$  by the spectral mapping, one

can use a measure of distortion in the spectral domain for  $coh(\Delta)$ . For example,

$$\begin{aligned} coh_{dist}(\Delta) &= \sum_{k=1}^K \sum_{i \in C_k} \|x_i - c_k\|^2 & coh_{distK}(\Delta) &= coh_{dist}(\Delta)/K \\ coh_{avgdist}(\Delta) &= \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i \in C_k} \|x_i - c_k\|^2 & coh_{avgdistK}(\Delta) &= coh_{avgdist}(\Delta)/K \end{aligned} \quad (16)$$

where  $c_k$  is the  $K$ -dimensional centroid of cluster  $C_k$  defined in lemma 9. These measures have the disadvantage of being algorithm dependent, as different algorithms may use different mappings. Also, it is not clear how to define  $coh$  for recursive algorithms like SM.

The gap as a measure of clustering quality has some appealing properties: it is algorithm independent and easily computable; it is minimized for PCE, no matter how large the  $MNCut$ . In addition, the gap measures the normalized cut of a clustering  $\Delta$  relative to (a lower bound on) the best achievable value. Therefore, we suggest using coherence measures based on the gap

$$\begin{aligned} coh_{gap}(\Delta) &= gap(\Delta) & coh_{gap/l}(\Delta) &= gap(\Delta)/l(|\Delta|) \\ coh_{gap/K}(\Delta) &= gap(\Delta)/|\Delta| & coh_{gap/eig}(\Delta) &= gap(\Delta)/eig(|\Delta|) \end{aligned} \quad (17)$$

The last three coherence measures are motivated by the fact that, as noise increases, the difference between the lower bound  $l(K)$  and the best achievable  $MNCut$  tends to increase, putting a larger penalty on clusterings with larger  $K$ .

**Experiments.** The first experiment uses an artificially constructed  $S$  with  $n = 100$  and  $K^* = 5$  blocks of sizes 10, 20, 30, 20, 20. The matrix (figure 2) is not block diagonal, the node volumes are unequal ( $\max D_i / \min D_i = 15$ ) and unevenly distributed (smallest  $D_i$ 's in the smallest cluster); the average  $S_{ij}$  is 0.057. This matrix is perfect for the given  $\Delta$ . We add symmetric i.i.d noise to the elements of  $S_{ij}$ ,  $noise_{ij} \sim \epsilon/n \times \text{uniform}[0, 1]$ , then call a clustering algorithm with  $K = 2, \dots, 12$  on the perturbed  $S$  obtaining  $\Delta_2, \dots, \Delta_8$ . Each of the coherence measures described previously is used to select the best clustering from this set and its  $K$  is recorded.

In total, we used 11 criteria: the gap based measures from (17), the distortion measures from (16),  $MNCut$ ,  $MNCut/K$ , and  $eig(K)$ . We also find the closest  $\Delta_K$  to the optimal clustering (by the *variation of information* (VI) [8]) and record  $K_{VI}$ . We use  $K_{VI}$  as the ‘‘gold standard’’; VI is also used to check that the clustering algorithm finds a clustering close to the optimum. We repeat the experiment with two algorithms, Meila-Shi [9] and Ng & al [10], for different noise levels  $\epsilon$ , 10 or more times for each  $\epsilon$ .

Results (figure 3) show that the gap normalized by the lower bound  $l(K)$  dominates all other coherence measures in robustness, indicating the best clustering up to SNRs of order 1 ( $\epsilon = 4$ ). As predicted, the eigengap, the  $MNCut$  and even the scaled  $MNCut/K$  are poor measures of coherence, always choosing the lowest possible number of clusters. The other measures, including the gap, work well for low noise levels, gradually failing in higher noise. Of the distortion measures, the distortion normalized by  $K$  is more robust than the others. We have run the experiment on other artificial similarity matrices, with several variants of the two algorithms, and have obtained results consistent with the ones here. The present  $S$  was designed to be difficult; on easy block diagonal matrices the coherence measures work well to SNR's up to 10. It is also worth noting that for a block diagonal  $S$  and low noise the eigengap is also a reliable criterion for selecting  $K$ , albeit never as robust as the gap based measures.

We also ran experiments on handwritten digits data obtained from [5]. The similarities were computed as in [10] with a kernel width  $\sigma = 10$ . We show results for a data set `digit10` with  $n = 1000$  data, 100 for each digit, and for the 5 most separable digits from this data set (digits 0, 2, 4 6, 7) forming data set `digit5`. The results are shown in the

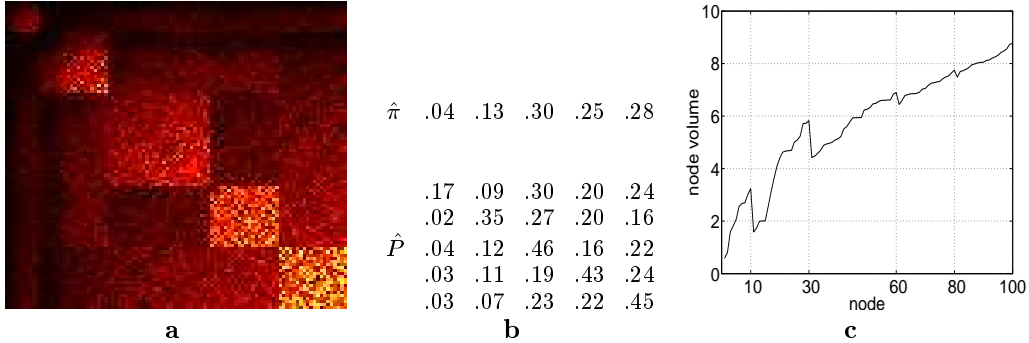


Figure 2: (a) The perfect  $S$  used in the experiments (white is higher, dark lower); (b) the aggregated transition matrix  $\hat{P}$  showing large crosstalk between clusters and the cluster probabilities  $\hat{\pi}$ ; (c) the values of  $D_i$  showing the concentration of small values in the smallest cluster.

table below<sup>1,2</sup>.

Data	VI	gap	gap/l	gap/K	gap/eig	MNCut/K	MNCut	eig	dist	dist/K	avgdist	avgdist/K
digit10	10	9	9.2	9	8	9	9	8	8	8	8	8
digit5	5	3	5	3	5	3	3	5	5	5	3	5

The `digit5` data set has an almost block diagonal  $S$ . Therefore, one sees that quality measures like the eigengap, and all distortion measures work well. The clustering obtained for  $K = 5$  is very close to the true clustering. By contrast, the `digit10` data represent a much harder task for the spectral clustering algorithms. Even the best clusterings leave a lot of confusion between the digits 1,3,5,8,9. Under these conditions, distortion measures fail by choosing the lowest possible  $K$ . The measures based on the gap and *MNCut* tend to choose  $K = 9$  (one lower than the best possible). By examining the detailed plots in figure 5 we saw that  $K = 10$  is often a close second for the normalized gap measures, but not for the other measures. For both data sets,  $gap/l(K)$  performs best.

These experiments are preliminary, but they strongly suggest that: (1) coherence is far superior to other criteria (eigengap, *MNCut*) in comparing clusterings with different  $K$ , and (2) of the variety of coherence measures tried here, the gap divided by the lower bound stands out as the most robust over algorithms, data sets and noise levels.

## 6 Discussion

This paper has provided a comprehensive analysis of spectral clustering near the *perfect point*, when the transition matrix  $P$  has PCE. How significant is this case for general spectral clustering algorithms on general data sets? We cannot answer what a spectral algorithm will do far away from the perfect point, but, based on our experience with spectral clustering, we conjecture that in practice, performance near this point accounts for a large proportion of successes of spectral clustering. If this is so, our analysis is highly significant. Second, as it was shown in [9, 15] there is a large family of spectral algorithms, including some of the most popular ones, whose perfect points are subsumed by the case when  $P$  has PCE and for which all the results proved here will hold. This is visible in the experiments in the last section, where the coherence measures behave similarly

<sup>1</sup>No noise was added, but we repeated each experiment 5 times with different random initializations for the k-means (or analog) part of the clustering algorithm. All standard deviations are 0, except for the 9.2 value which is the average of four 9's and a 10.

<sup>2</sup>By using another measure of comparison between clusterings, the clustering error [15], we also obtained  $K^* = 10$  and 5 respectively for these data sets.

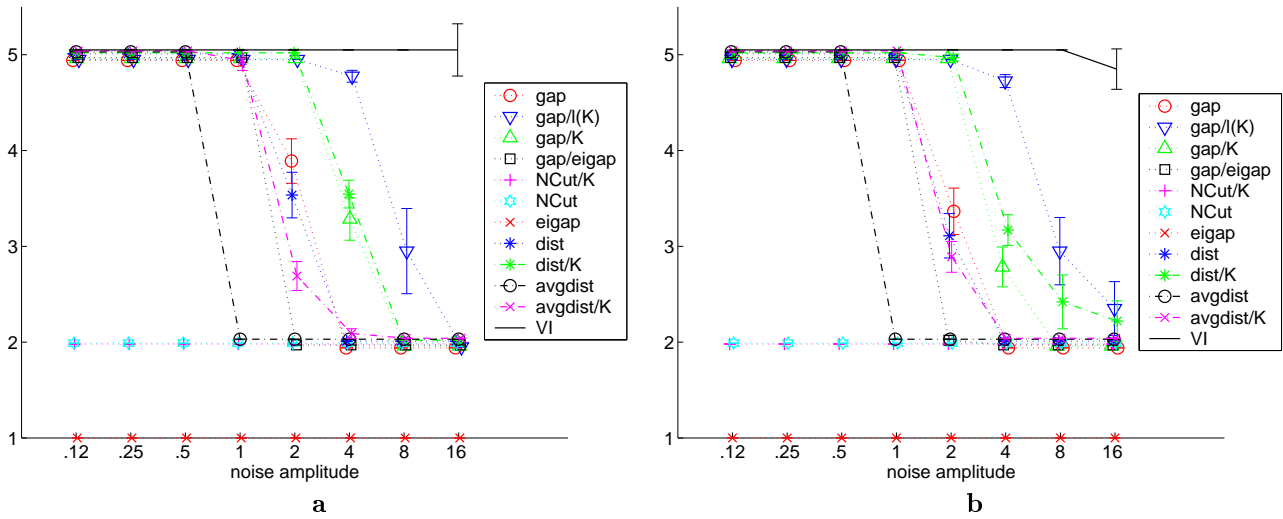


Figure 3: Selection of  $K$  on an artificial  $S$  as a function of the noise level  $\epsilon$ , for 11 criteria: (a) average values of  $K$  for clustering with the Ng & al algorithm; (b) average values of  $K$  for clusterings obtained with the Meila-Shi algorithm. The distortion based values are shown with dashed lines, the others are shown with dotted lines; a continuous line shows the “gold standard”: the  $K$  that would be chosen if we could compare with the true clustering. Note that the two algorithms have different spectral mappings; the distortion measures were computed appropriately.

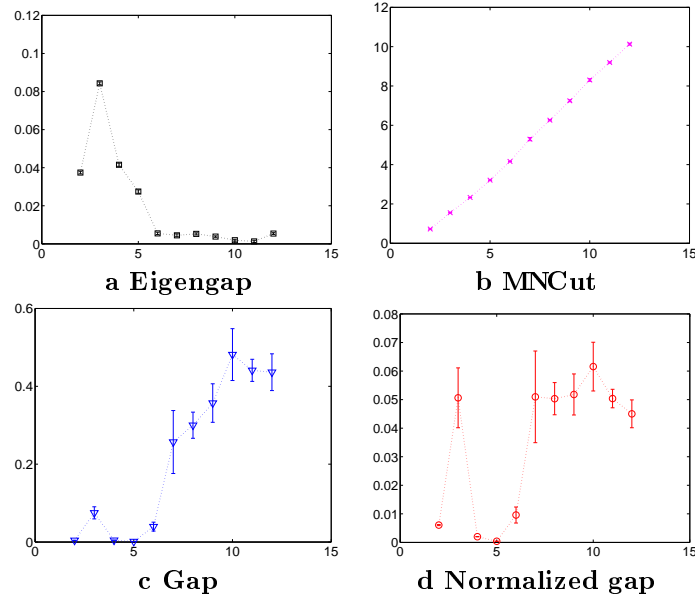


Figure 4: Average values and standard deviations for different cluster selection criteria on the artificial data, at a medium level of noise, over 10 runs.

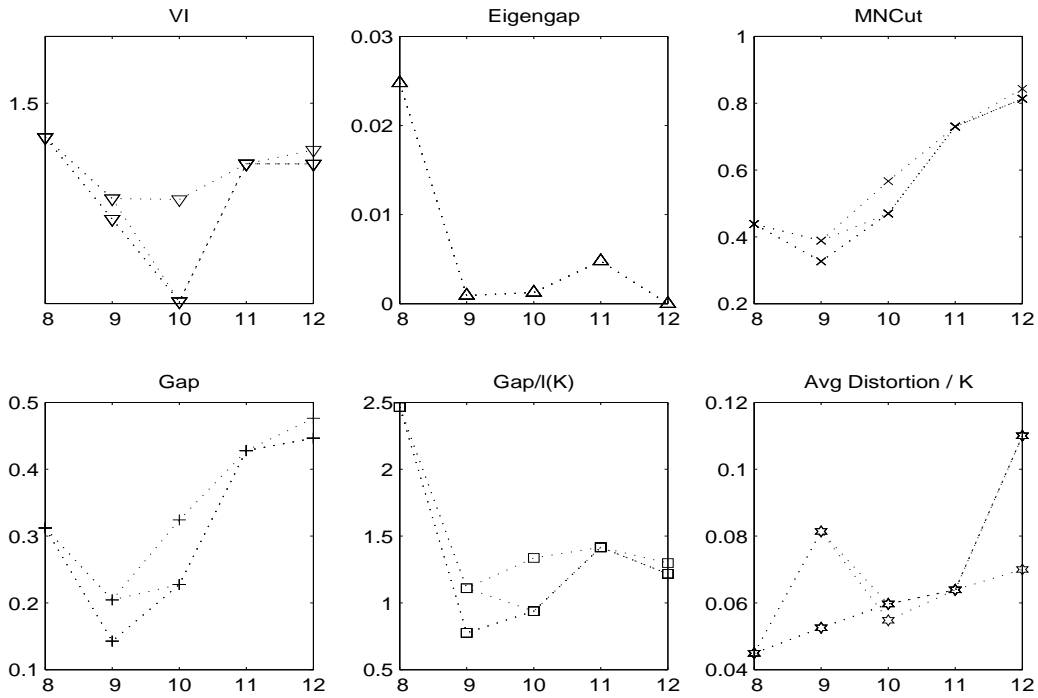


Figure 5: Values for six cluster selection criteria for 5 runs on the 10 digits data. Note that VI always achieves a minimum at 10, *eig* is maximized at 8, *MNCut*, *gap*, *gap/l* are minimized at 9, except for 1 case when *gap/l* is minimized at 10. The value  $K = 10$  is a close second for *gap/l* and a not-so-close second for *gap*. The best performing of the distortion criteria on this data set is *avgdist/K*, minimized at 8, with a second at 10.

across algorithms. This fact is also what allowed us to talk about “spectral clustering” throughout this paper, meaning any of the above mentioned family of algorithms, instead of the particular spectral mapping that we focused on here.

We have introduced the *MNCut* criterion, inspired by the random walks view of spectral clustering, and have shown that it is NP-hard to minimize in general, but that for a nearly perfect  $S$ , it will be optimized by spectral clustering. The case of well separated clusters (block diagonal  $S$ ) is an instance of a perfect clustering situation, but it is not the unique one. As an aside, from the Multicut lemma, one can derive a tractable and tight lower bound  $l(K)$  on the *MNCut* achievable by any clustering.

We also introduced the gap  $MNCut(\Delta) - l(|\Delta|)$  and showed that it measures the internal coherence of the clusters and therefore that spectral clustering is simultaneously optimizing two criteria: the clusters’ separation and their internal coherence.

Our bicriterial view of clustering quality is very close in spirit to a bicriterial framework proposed in [4]. We especially agree with the authors’ view of the number of clusters  $K$ . Quantitatively, however, it can be shown that the two frameworks are not reducible to one another.

In view of this we argued that one can define a variety of cluster selection criteria illustrating different tradeoffs between *MNCut* and *coh*. Our experiments have demonstrated that this approach, and in particular the coherence measures based on the gap, agree better with human intuitions of a “good” clustering than using the *MNCut* alone.

## Acknowledgements

We thank Jianbo Shi, Chris Meek and Andrew Ng for stimulating discussions at the onset of this work. This research was partially supported by University of Washington Royalty Research Fund.

## A Proof of Theorem 5 (NP-hardness of minimizing *MNCut*)

We first prove the following:

**Lemma 10 (Refining a partition increases *MNCut*)** *Let  $\Delta = \{C_1, \dots, C_K\}$  be a partition of  $V$ , and let  $\Delta' = \{C'_1, C''_1, \dots, C_K\}$ , where  $C_1 = C'_1 \cup C''_1$ . Then*

$$MNCut(\Delta) \leq MNCut(\Delta')$$

**Proof** Denote by  $P_{il} = \sum_{j \in C_l} P_{ij}$  where  $i \in V$  and  $l = 1, \dots, K$  and by  $P'_{i1}$  ( $P''_{i1}$ ) the sums  $\sum_{j \in C'_1} P_{ij}$ , ( $\sum_{j \in C''_1} P_{ij}$ ) respectively.

$$MNCut(\Delta) = \sum_{k=1}^K \sum_{k' \neq k} \underbrace{\sum_{i \in C_k} \frac{\pi_i P_{ik'}}{\pi_{C_k}}}_{c_{kk'}} \quad (18)$$

Let us make the notations

$$c'_{1l} = \sum_{i \in C'_1} \frac{\pi_i P_{il}}{\pi_{C'_1}} \quad (19)$$

$$c''_{1l} = \sum_{i \in C''_1} \frac{\pi_i P_{il}}{\pi_{C''_1}} \quad (20)$$

$$c'_{l1} = \sum_{i \in C_l} \frac{\pi_i P'_{i1}}{\pi_{C_l}} \quad (21)$$

$$c''_{l1} = \sum_{i \in C_l} \frac{\pi_i P''_{i1}}{\pi_{C_l}} \quad (22)$$

$$c_{11} = \text{Cut}(C'_1, C''_1) \left( \frac{1}{\text{Vol}C'_1} + \frac{1}{\text{Vol}C''_1} \right) \quad (23)$$

It follows that

$$c_{1l} = \frac{\pi_{C'_1}}{\pi_{C_1}} c'_{1l} + \frac{\pi_{C''_1}}{\pi_{C_1}} c''_{1l} \leq c'_{1l} + c''_{1l} \quad (24)$$

$$\text{and} \quad (25)$$

$$\text{MNCut}(\Delta') = \underbrace{c_{11}}_{\geq 0} + \sum_{l=2}^K \underbrace{(c'_{1l} + c''_{1l})}_{\geq c_{1l}} + \sum_{l=2}^K \underbrace{(c'_{l1} + c''_{l1})}_{c_{l1}} + \sum_{k=2}^K \sum_{k' \neq k, k' > 1} c_{kk'} \quad (26)$$

$$\geq \sum_{l=2}^L c_{1l} + \sum_{l=2}^L c_{l1} + \sum_{k=2}^K \sum_{k' \neq k, k' > 1} c_{kk'} \quad (27)$$

$$= \text{MNCut}(\Delta) \quad (28)$$

Q.E.D.

**Corollary 11** *Equality in lemma 10 is attained only if  $c_{11} = c'_{1l} = c''_{1l} = 0$ , i.e. only if  $S$  is block diagonal with blocks given by the sets of points  $C'_1, C''_1, \cup_{k \geq 2} C_k$ .*

Now we prove theorem 5 by reducing it to the 2-way case for which the proof is given in [13]. To keep the notation simple, we give the detailed proof for  $K = 3$ , then a proof sketch for arbitrary  $K$ . Denote by  $\mathcal{P}_K(S)$  the problem of finding the clustering with  $K$  clusters that minimizes  $\text{MNCut}$  on  $S$ .

**Proof that minimizing  $\text{MNCut}(\Delta)$  over  $\Delta, |\Delta| = 3$  is NP-hard** Assume that  $\mathcal{P}_3(S)$  can be solved in polynomial time for any  $S$ . We will show a contradiction, by reducing  $\mathcal{P}_2$  (known to be NP-hard) to it. Take an  $S$  that is not block-diagonal and let  $\Delta^*$  be the solution to  $\mathcal{P}_2(S)$ . Augment the data set with one point  $x$  of volume  $d$  dissimilar to all other points and denote  $\tilde{V} = V \cup x$ ,

$$\tilde{S} = \begin{bmatrix} S & 0 \\ 0 & d \end{bmatrix} \quad (29)$$

Let  $\tilde{\Delta}^* = \Delta^* \cup \{x\}$ ;  $\tilde{\Delta}^*$  is a 3-way clustering on  $\tilde{V}$ . We show that for  $d$  small enough,  $\tilde{\Delta}^*$  is a solution to  $\mathcal{P}_3(\tilde{S})$ . In the following, we mark partitions of  $\tilde{S}$  with a  $\tilde{\cdot}$  and the corresponding partitions on  $S$  with the same symbol without the  $\tilde{\cdot}$ .

It is easy to see that  $\text{MNCut}(\tilde{\Delta}^*) = \text{MNCut}(\Delta^*)$ . Let  $\tilde{\Delta} = \{A \cup \{x\}, B, C\}$  be any other 3-way partition of  $\tilde{V}$ . Assuming that  $A \neq \emptyset$ ,  $\Delta = \{A, B, C\}$  is a 3-way partition of  $S$  (The case  $A = \emptyset$ , engendering a two-way partition on  $V$  is trivial and left as an

exercise to the reader). It is also easy to see, using lemma 10 and corollary 11 that  $MNCut(\Delta) > MNCut(\Delta^*)$ .

Noting that  $\tilde{P}_{iA} = P_{iA}$ , etc we have that

$$MNCut(\tilde{\Delta}) = \frac{\sum_{i \in A} \tilde{\pi}_i (P_{iB} + P_{iC})}{\tilde{\pi}_A + \tilde{\pi}_x} + \frac{\sum_{i \in B} \tilde{\pi}_i (P_{iA} + P_{iC})}{\tilde{\pi}_B} + \frac{\sum_{i \in C} \tilde{\pi}_i (P_{iA} + P_{iB})}{\tilde{\pi}_B} \quad (30)$$

Where  $\tilde{\pi}_U = \frac{VolU}{VolV+d}$  for  $U = A, B, C, \{x\}$ . When  $d \rightarrow 0$ ,

$$MNCut(\tilde{\Delta}) \rightarrow MNCut(\Delta) > MNCut(\Delta^*) = MNCut(\tilde{\Delta}^*)$$

Therefore, for  $d$  sufficiently small,  $\tilde{\Delta}^*$  is the solution of  $\mathcal{P}_3(\tilde{S})$ , implying that we can solve  $\mathcal{P}_2(S)$  in polynomial time by solving  $\mathcal{P}_3(\tilde{S})$  – a contradiction.

For the general case, we augment  $V$  with  $K - 2$  points  $x_1, \dots, x_{K-2}$  of volume  $d$  and show that the partition  $\tilde{\Delta}^* = \Delta^* \cup \{x_1\} \cup \dots \cup \{x_{K-2}\}$  is the solution to  $\mathcal{P}_K(\tilde{S})$  for small enough  $d$ .

## B Proofs of lemmas 6, 7 (Multicut) and 4

### Proof of lemma 6

$$MNCut(\Delta^*) = \sum_{k=1}^K \left[ 1 - \sum_{j \in C_k^*} P_{i_k j} \right] \text{ for some } i_k \in C_k \quad (31)$$

$$= \sum_{k=1}^K (1 - \hat{P}_{kk}) \quad (32)$$

$$= K - \text{trace} \hat{P} \quad (33)$$

$$= K - \sum_{k=1}^K \lambda_k \quad (34)$$

In the above derivations we used the results of Lemma 1.  
Q.E.D.

**Proof of the Multicut lemma 7** We will show that there is no  $K$ -way cut that achieves a value smaller than  $K - \sum_{k=1}^K \lambda_k$ . Consider an arbitrary partition  $\Delta = \{C_1, \dots, C_K\}$ . Denote by  $x^k$  the indicator vector of cluster  $C_k$  for  $k = 1, \dots, K$ .

Let us massage the expression of  $MNCut(\Delta)$  into a convenient form:

$$MNCut(\Delta) = K - \sum_{k=1}^K Pr[C_k \rightarrow C_k | C_k] \quad (35)$$

$$= K - \sum_{k=1}^K \frac{\sum_{i \in C_k} D_i \sum_{j \in C_k} P_{ij}}{\sum_{i \in C_k} D_i} \quad (36)$$

$$= K - \sum_{k=1}^K \frac{\sum_{i, j \in C_k} S_{ij}}{\sum_{i \in C_k} D_i} \quad (37)$$

Noting that

$$\sum_{i \in C_k} D_i = \sum_{i \in V} (x_i^k)^2 D_i \quad (38)$$

and

$$\sum_{i,j \in C_k} S_{ij} = \sum_{i,j \in V} S_{ij} x_i^k x_j^k \quad (39)$$

$$= \frac{1}{2} \sum_{i,j \in V} S_{ij} [(x_i^k)^2 + (x_j^k)^2 - (x_i^k - x_j^k)^2] \quad (40)$$

$$= \sum_{i \in V} (x_i^k)^2 D_i - \sum_{ij \in E} S_{ij} (x_i^k - x_j^k)^2 \quad (41)$$

we obtain that

$$MNCut(\Delta) = \sum_{k=1}^K \frac{\sum_{ij \in E} S_{ij} (x_i^k - x_j^k)^2}{\sum_{i \in V} (x_i^k)^2 D_i} \quad (42)$$

$$= \sum_{k=1}^K R(x^k) \quad (43)$$

In the sums above,  $i, j \in V$  means summation over the cartesian product  $V \times V$  while  $ij \in E$  means summation over all “edges”, i.e all unordered pairs  $(i, j)$  with  $i \neq j$ .

The expression  $R(x)$  represents the Rayleigh quotient for the Laplacian of the weighted graph described by  $S$  c.f.[2] equation (1.13). Therefore, if we denote

$$y^k = D^{1/2} x^k \quad (44)$$

we have that

$$R(x^k) = \frac{(y^k)^T L y^k}{(y^k)^T y^k} = \tilde{R}(y^k) \quad (45)$$

and

$$MNCut(\Delta) = \sum_{k=1}^K \tilde{R}(y^k) \quad (46)$$

Now we turn to the problem of minimizing  $MNCut$ . It is easy to see that minimizing  $MNCut(\Delta)$  over all partitions  $\Delta$  is equivalent to finding

$$\min \sum_{k=1}^K R(x^k) \text{ s.t. } x_i^k \in \{0, 1\} \forall k, i \text{ and } x^k \perp D x^l \text{ for } k \neq l \quad (47)$$

This minimum is greater or equal to

$$\min \underbrace{\sum_{k=1}^K R(x^k)}_{J(x^1, \dots, x^K)} \text{ s.t. } x^k \perp D x^l \text{ for } k \neq l \quad (48)$$

Now we focus on  $J$  and show that its minimum in (48) is equal to  $MNCut(\Delta^*)$  which will prove that the latter is the smallest achievable  $K$ -way normalized cut.

With  $y^1, \dots, y^K$  defined as in (44) we have that

$$x^k \perp D x^l \Leftrightarrow y^k \perp y^l. \quad (49)$$

In addition, we can assume w.l.o.g. that the vectors  $y^k$  have unit length. Therefore, minimizing  $J$  is equivalent to finding

$$\min \tilde{J}(y^1, \dots, y^K) = \min \sum_{k=1}^K (y^k)^T L y^k \text{ s.t. } (y^l)^T y^k = \delta_{kl} \quad (50)$$

By Lemma 4, the minimum is  $\sum_{k=1}^K \mu_k$ .

Now we use Lemma 2 which says that  $\mu_k = 1 - \lambda_k$ . This proves that

$$\min \tilde{J} = \min J = \text{MNCut}(\Delta^*) \quad (51)$$

In addition, note that the minimizing vectors  $y^k$  are linear combinations of the first  $K$  eigenvectors of  $L$ . Again, by Lemma 2 we have that  $x^1, \dots, x^K$  are linear combinations of the first  $K$  eigenvectors of  $P$  which are piecewise constant.

**Proof of lemma 4** Denote by  $u^1, \dots, u^n$  the orthonormal eigenvectors of  $L$ . The  $y$ 's are linear combinations of these vectors, i.e

$$y^k = \sum_{j=1}^n a_{jk} u^j \quad (52)$$

Denote by  $A$  the  $n \times K$  matrix  $[a_{jk}]_{j,k}$ . Because both  $\{u^j\}$  and  $\{y^k\}$  are orthonormal, the vectors  $\{a_{\cdot k}\}$  (i.e the columns of  $A$ ) are orthonormal as well. We minimize  $\tilde{J}$  w.r.t  $\{a_{jk}\}$  by the Lagrange multiplier method.

$$\tilde{J}_\beta = \sum_{j=1}^n \sum_{k=1}^K \mu_j a_{jk}^2 - \sum_{k>l} \beta_{kl} a_{\cdot l}^T a_{\cdot k} - \sum_{k=1}^K \beta_k (a_{\cdot k}^T a_{\cdot k} - 1) \quad (53)$$

Because

$$(y^k)^T L y^k = \sum_{j=1}^n \mu_j a_{jk}^2 \quad (54)$$

$$\frac{\partial \tilde{J}}{\partial a_{jk}} = 2\mu_j a_{jk} - \beta_{jk} \sum_{l \neq k} a_{jl} - 2\beta_k a_{jk} \quad (55)$$

Equating the above partial derivative with 0, and defining the  $K \times K$  matrix  $B$  to be  $B_{kl} = B_{lk} = 1/2\beta_{kl}$  for  $k \neq l$  and  $B_{kk} = \beta_k$  we obtain

$$B a_{\cdot j} = \mu_j a_{\cdot j} \text{ for } j = 1, \dots, n \quad (56)$$

Thus the vectors  $a_{\cdot j}$  are either eigenvectors of  $B$  or 0. There are  $n$  eigenvalues  $\mu_j$  while the matrix  $B$  is  $K \times K$  so it can have only  $K$  eigenvalues and corresponding independent eigenvectors. Hence, at most  $K$  rows of  $A$  are non-zero. The non-zero rows are orthogonal, because  $B$  is symmetric. On the other hand,  $A$  has orthonormal columns, therefore there will be at least  $K$  non-zero rows in  $A$ . Denote by  $H$  the  $K \times K$  matrix obtained from  $A$  by eliminating the null rows. The columns of  $H$  are orthonormal, therefore its rows must be orthonormal as well. Denote by  $j_1, \dots, j_K$  the indices of these rows in  $A$ .

The value of  $\tilde{J}$  under the conditions above becomes

$$\tilde{J} = \sum_{i=1}^K \mu_{j_i} \quad (57)$$

Thus,  $\tilde{J}$  has several local minima, one for each possible subset of indices  $j_1, \dots, j_K$ . For each of them, the value of the minimum is the sum of the selected eigenvalues and the minimizing vectors  $y^k$  are linear combinations of the eigenvectors  $u^{j_1}, \dots, u^{j_K}$ . The lowest of the minima corresponds to choosing the  $K$  smallest eigenvalues of the Laplacian, i.e.  $\mu_1, \dots, \mu_K$ .

## C Proof of lemma 9

Write  $[v^1, \dots, v^K] = [u^1, \dots, u^K] + E$  where  $u^k$  is piecewise constant w.r.t. to  $\Delta^*$ ,  $u_{ij} = c_{kj}$  if  $x_i \in C_k^*$ , and  $E_{ij} = v_i^j - u_{ij}$ . By our assumption,  $\|E_{ij}\|_F < \epsilon$ . Write  $E = [r^1, \dots, r^K]$ . Write  $C = (c_{ij})_{K \times K}$ . We have

$$P[v^1, \dots, v^K] = P[u^1, \dots, u^K] + P[r^1, \dots, r^K].$$

Since  $Pv^k = \lambda_k v^k$ , then

$$[\lambda_1 v^1, \dots, \lambda_K v^K] = [Pu^1, \dots, Pu^K] + [Pr^1, \dots, Pr^K].$$

Therefore we have  $Pu^l = \lambda_l(u^l + r^l) - Pr^l = \lambda_l(u^l) + (\lambda_l r^l - Pr^l)$ . Thus

$$P[u^1, \dots, u^K] = \text{diag}(\lambda_1, \dots, \lambda_K)[u^1, \dots, u^K] + (\text{diag}(\lambda_1, \dots, \lambda_K, \dots, 0) - P)E.$$

By computation,  $[u^1, \dots, u^K] * C^{-1} = [x^1, \dots, x^K]$  where  $x^k$  is the indicator vector of the cluster  $C_k^*$ . Let  $\lambda_{K+1}$  be the  $K+1$  largest eigenvalue. Let  $\sigma$  be the smallest singular value of  $C$ . Then

$$P[x^1, \dots, x^K] = \text{diag}(\lambda_1, \dots, \lambda_K)[x^1, \dots, x^K] + (\text{Diag}(\lambda_1, \dots, \lambda_K, \dots, 0) - P)EC^{-1}.$$

Since  $\|(\text{diag}(\lambda_1, \dots, \lambda_K, \dots, 0) - P)EC^{-1}\|_F \leq \lambda_{K+1} * \|E\|_F * 1/\sigma \leq \lambda_{K+1}\epsilon/\sigma$ , we have for any  $i, i' \in C_i^*$  and for any  $k$

$$|\sum_{j \in C_k^*} P_{ij} - \sum_{j \in C_k^*} P_{i'j}| \leq \lambda_{K+1}\delta/\sigma.$$

Let  $\beta = \lambda_{K+1}\epsilon/\sigma$ .

Compute  $MNCut(\Delta)$ , we get

$$MNCut(\Delta^*) = K - \sum_{k=1}^K Pr(C_k^* \rightarrow C_k^* | C_k^*) \quad (58)$$

$$\leq K - \sum_{k=1}^K \frac{\sum_{i \in C_k^*} D_i(\lambda_i + \beta)}{\sum_{i \in C_k} D_i} \quad (59)$$

$$\leq K - \sum_{k=1}^K \lambda_k + K * \beta \quad (60)$$

$$= K - \sum_{k=1}^K \lambda_k + K\lambda_{K+1}\epsilon/\sigma \quad (61)$$

## References

- [1] Yossi Azar, Amos Fiat, Anna R. Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *ACM Symposium on Theory of Computing*, pages 619–626, 2001.
- [2] Fan R. K. Chung. *Spectral Graph Theory*. Number Regional Conference Series in Mathematics in 92. American Mathematical Society, Providence, RI, 1997.
- [3] Eigencluster, [www-math.mit.edu/cluster](http://www-math.mit.edu/cluster), 2000.
- [4] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: good, bad and spectral. In *Proc. of 41st Symposium on the Foundations of Computer Science, FOCS 2000*, 2000.

- [5] C. Kaynak. Methods of combining multiple classifiers and their applications to handwritten digit recognition. Master's thesis, Institute of Graduate Studies in Science and Engineering, Bogazici University, 1995.
- [6] Laszlo Lovász, Ravi Kannan, and M. Simonovits. Isoperimetric problems for convex bodies. *Discrete Computational Geometry*, 13:541–559, 1995.
- [7] Marina Meila and Liang Xu. A unicity theorem for spectral clustering. Technical report, University of Washington, 2003. (in preparation).
- [8] Marina Meilă. Comparing clusterings. Technical Report 419, University of Washington, 2002. [www.stat.washington.edu/reports](http://www.stat.washington.edu/reports).
- [9] Marina Meilă and Jianbo Shi. A random walks view of spectral segmentation. In T. Jaakkola and T. Richardson, editors, *Artificial Intelligence and Statistics AIS-TATS*, 2001.
- [10] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.
- [11] P. Perona and W. Freeman. A factorization approach to grouping. In *European Conference on Computer Vision*, 1998.
- [12] G.L. Scott and H. C. Longuet-Higgins. Feature grouping by relocation of eigenvectors of the proximity matrix. In *Proceeding of the British Machine Vision Conference*, 1990.
- [13] J. Shi and J. Malik. Normalized cuts and image segmentation. *PAMI*, 2000.
- [14] Daniel A. Spielman and Shang-Hua Teng. Spectral partitioning works: Planar graphs and finite element meshes. In *IEEE Symposium on Foundations of Computer Science*, pages 96–105, 1996.
- [15] Deepak Verma and Marina Meilă. A comparison of spectral clustering algorithms. TR 03-05-01, University of Washington, May 2003. (submitted).
- [16] Y. Weiss. Segmentation using eigenvectors: a unifying view. In *International Conference on Computer Vision*, 1999.