Spectral Clustering Survey

HU, Pili*

May 14, 2012†

Abstract

Abstract. Sources can be found in [13].

*hupili [at] ie [dot] cuhk [dot] edu [dot] hk

†Last compile:May 17, 2012

Contents

1 Introduction

Spectral Clustering(SC) was used in several disciplines long ago. For example, computer vision[25]. Spectral Embeding(SE) was also widely discussed in the community[7]. Outside spectral community, the machine learning community also developed many linear or non-linear Dimensionality Reduction(DR) methods, like Principal Component Anslysis (PCA), Kernel PCA (KPCA)[24], Locally Linear Embedding (LLE)[21], etc. Other technique like Multi-Dimensional Scaling(MDS) was successfully used in computational psychology for a very long time[6], which can be viewed as both "embedding" or "dimensionality reduction".

According to our survey, although those methods target at different problems and are derived from different assumptions, they do share a lot in common. The most significant sign is that, the core procedure involves Eigen Value Decomposition(EVD) or Singular Value Decomposition(SVD), aka "spectral". They all involve an intermidiate step of embedding high-dimensional / non-Euclidean / non-metric points into a low-dimensional Euclidean space (although some do not embed explicitly). In this case, we categorize all these algorithms as Spectral Embedding Technique(SET).

1.1 A Sample Spectral Clustering Algorithm

There are many variations of SC. They all work under certain conditions and researchers don't have a rule of thumb so far. Before we analyze their procedure and justification, we present a simple but workable sample algorithm(**Alg 1**).

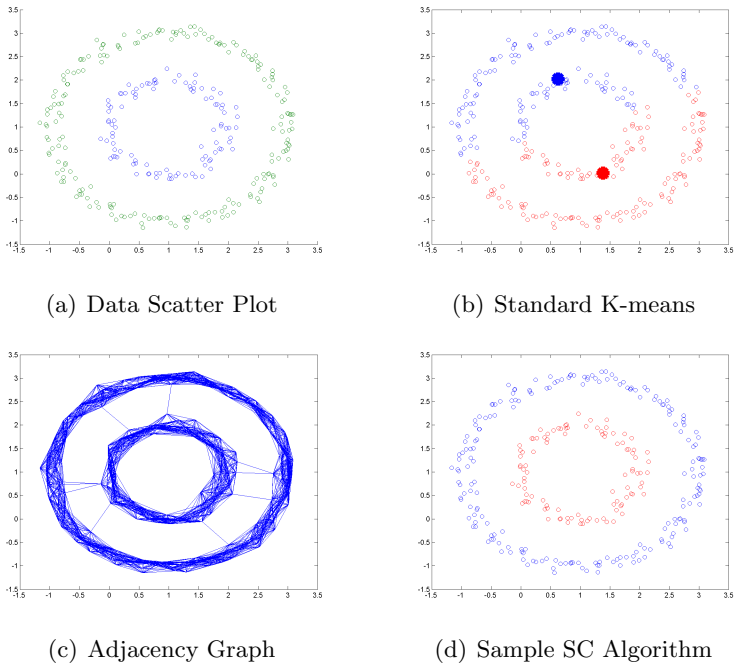
Algorithm 1 Sample Spectral Clustering

Input: Data matrix $X = [x_1, x_2, \ldots, x_N]$; Number of Clusters K.
Output: Clustering $\{C_i\}$: $C_i \in V$ and $\cap_i C_i = \emptyset$ and $\cup_i C_i = V$.

1: Form adjacency matrix A within ϵ-ball.
2: Solve $A = U\Lambda U^{\mathrm{T}}$, indexed according the eigenvalue's magnitude.
3: $Y \leftarrow$ first K columns of U.
4: Cluster Y's rows by K-means.

In **Alg 1**, the ϵ-ball adjacency graph is constructed as follows. First create one vertex for each data point. If for two points i, j satisfy $||x_i - x_j|| < \epsilon$, connect them with an edge. In this simple demonstration, we consider an unweighted graph, i.e. all entries of A are 0(disconnected) or 1(connected).

Fig 1 demonstrates the result of our sample SC algorithm, compared with standard K-means algorithm. **Fig 1(a)** shows the scatter plot of data. It is composed of one radius 1 circle and another radius 2 circle, both centered at (1,1). **Fig 1(b)** shows the result of standard K-means working on Euclidean distance. **Fig 1(c)** shows the graph representation, where the



(a) Data Scatter Plot (b) Standard K-means

(c) Adjacency Graph (d) Sample SC Algorithm

Figure 1: Demonstration of Sample SC Algorithm

adjacency graph is formed by taking a ϵ-ball and $\epsilon = 0.7$ in the example. **Fig 1(d)** shows the output of **Alg 1**. It's obvious that standard K-means algorithm can not correctly cluster the two circles. This is a known major weakness of K-means(in Euclidean): When clusters are not well separated spheres, it has difficulty recovering the underlying clusters. Although K-means works for this case if we transform the points into polar coordinate system(see [13] for code), the solution is not universal. However, in this example, our sample SC algorithm can separate the two clusters, probably because the eigenvectors of adjacency matrix convey adequate information.

A precaution is that **Alg 1** does not always work even in this simple case. Nor have we seen this algorithm from formally published works (so far), let alone justifications. This algorithm is only to show the flavour of spectral clustering and it contains those important steps in other more sophisticated algorithms. Readers are recommended to learn von Luxburg's tutorial[28] before reading the following sections. Since that paper is very detailed, we'll present overlapping topics concisely.

1.2 Spectral Clustering Taxonomy

When the readers start to survey SC related topics, they will soon find that there are two streams of study:

- **Embedding Like**. An example is presented in **Alg 1**. One of the core

procedure is an embedding of points into lower dimensional Euclidean space. After that, hard cut algorithm like K-means is invoked to get the final result. This stream has a lot in common with SE and DR. In the current article, we focus on this line of research.

- **Partitioning Like**. One early example is the 2-way cut algorithm presented by Shi and Malik in[25]. Later Kannan et al. analyzed the framework further[16]. The core procedure of this type of algorithm is a 2-way partitioning subroutine. By invoke this subroutine on resultant subgraph repeatedly, we can get final K clusters. When the subroutine can guarantee certain quality, the global resultant clustering quality can be guaranteed [16]. The attracting aspect of Kannan's framework is that, we can plugin any subroutine as long as they have quality gurantee in one cut. For example, the eigen vector of left normalized graph Laplacian can induce good 2-way cut in terms of the Normalized Cut[25] (**Section 3.1.3**). This line of research is more close to Spectral Graph Partitioning(SGP), although those algorithms also get the name of Spectral Clustering.

It's worth to note some work of SGP. In Spielman's[26] and Lau's[18] lecture notes, they both presented a graph partitioning algorithm said to result from Lovász[19]. Chung[8] made this framework clear: First define a function $f(v)$ on all vertices; Then set threshold T to cut vertices into two groups, $\{v|f(v) \leq T\}$ and $\{v|f(v) > T\}$. The question now becomes to find a good heuristic function f. Note that the original bi-partition problem is of complexity $O(2^N)$ and the heuristic based bi-partition problem is of $O(N)$(plus the time to get f). If we define f as the second eigen vector of left normalized graph Laplacian, it is just the algorithm of Shi and Malik[25]. Besides, there can be multiple f_i and the best cut can still be searched in polynomial time (given f_i). For example, in [18] Lau used random walk probability $P_1, P_2, \ldots, P_t$ as f. Some recent research of the heuristic function are Personalized Pagerank[2], Evolving Sets[3], etc.

In the rest of this article, we refer Spectral Clustering to the first type, i.e. the embedding like algorithms.

1.3 Linear Algebraic Properties

2 Spectral Clustering Framework

We propose the following framework to absorb currently surveyed variations of SC(and other SET):

1. **Metric Formulation**. This step forms a pairwise metric, upon which an adjacency matrix of (weighted) graph can be constructed. There

are several kinds of input: high-dimensional data (usual case); pairwise proximity input (like MDS, see **Section 4.1**); (weighted) graph (like the input of SGP). If the input of SC is already a graph, this step is omitted. For proximity measures, especially dissmilarity, it is usually first converted to approximte pairwise inner product in Euclidean space. The pairwise inner product is a positive related quantity with similarity(e.g. Jaccard's coefficient for 0/1 vector[29]), and thus suits the notion of weights of graph edges. For high-dimensional data, there are more freedom in the metric formulation stage, like similarity graph[28], geodesic distance[27], etc.

2. **Sepctral Embedding**. With the adjacency matrix built in last stage, this stage embeds all vertices into a lower dimensional Euclidean space. For SC community, this embedding makes clustering structure more clear, so that simple algorithms working in Euclidean space can detect the clusters. For DR community, this embedding reveals the shape of manifolds in their parametric space. The two goals are essentially the same. The core procedure is to do EVD and differences lie before and after EVD:

   - **Matrix Type**. Some authors use graph Laplacian [28, 4, 25] ; others use [20, 7, 16] adjacency matrix.
   - **Normalization**. Both Laplacian and adjacency matrix can be unnormalized, symmetrically normalized, or left(row) normalized[28]. They corresponds to different interpretation and will be explored later.
   - **Scaling**. As is shown in **Alg 1**, we can directly embed vertices using the corresponding row of Y (like [20]). Other alternatives are to scale by square root eigenvalue (like [7]) and scale by eigenvalue (like PCA[5]).
   - **Projection**. For many algorithms, the Y (after scaling) provides an Euclidean space embedding. There are others which further project the rows of Y onto a unit sphere, like [20] and [7].

3. **Clustering**. Based on the embedding result, simple algorithms can be invoked to perform a hard cut on the points. Traditional methods from data mining community are K-means and hierarchical clustering like single/complete linkage[15]. Among those techniques, K-means are the most widely used. A variation of K-means will be proposed later in this article, in order to better fit some angle preserving SET. Simpler hard cut techniques are also possible, e.g. looking at the largest entry of the embedding coordinates[16].

Note that not all of the combinations are justified in published works. We organize them in this way to reveal possibilities from a practitioners

perspective. If some combinations yield good results in practice, we can seek for justifications using tools from spectral graph theory or machine learning.

2.1 Metric Formulation

We rate great importance of metric formulation step, although it is not the main body of SET. There are always many ways of metric formulation given a practical problem. With poorly constructed metric matrix, even the best embedding technique helps little.

2.1.1 High Dimensional Data

For high dimensional data, the following ways can be applied to obtain a graph representation:

- ϵ-ball[28]. If $||x_i - x_j|| < \epsilon$, we connect vertices i, j with an edge.
- k-Nearest-Neighbour(kNN)[28]. For each vertex, we connect it with its k nearest neighbours based on Euclidean distance.
- Mutual kNN(m-kNN)[28]. Note the set of kNN is not symmetric. Multual kNN connects those points who are kNN to each other.
- Complete graph[28]. All vertices are connected with each other. This construction is ususally used with Gaussian kernel weighting below.

After the construction of graph, edges can be weighted in several ways:

- Unweighted[4]. The adjacency matrix is mere 1(connected) or 0(disconnected).
- Gaussian kernel[28], also called heat kernel[4]. Each edge is weighted by $A_{ij} = \exp\{-||x_i - x_j||^2/t\}$, where t is a super parameter controling the decaying rate of similarity.

Heuristics on selecting ϵ, k, t are proposed by many authors, e.g. ch8 of [28], but there is no real rule of thumb. Since the focus of SET study is on the embedding part, most work do not try hard to tweak the construction of similarity graph. We propose other possibilities, which may be helpful to target different practical problems:

- Mahalanobis distance[30]. The connection condition $||x_i - x_j|| < \epsilon$ is substituted by $(x_i - x_j)^{\mathrm{T}}\Sigma^{-1}(x_i - x_j) < \epsilon^2$. Accodingly, when Gaussian kernel is used, the edge weight is given by $A_{ij} = \exp\{-0.5(x_i - x_j)^{\mathrm{T}}\Sigma^{-1}(x_i - x_j)\}$.

- Jaccard's coefficient[29]. It computes the ratio of the intersection and union two sets. It is useful when the high dimensional input coordinates can be interpreted as sets.
- Cosine similarity[31]. It computes the angle between two vectors and is widely used in text mining context.

2.1.2 Proximity

Many real problems has proximity as input. Proximity can be described by similarity or dissimilarity. With pairwise similarity input, we can directly fit the data into following SE procedure. A more interesting problem is to transform dissimilarity into similarity, or at least an equivalent quantity. Decomposing the transformed matrix should be able to yield reasonable embedding(under certain justification).

Denote data matrix by $X = [x_1, x_2, \ldots, x_N]$. Every column x_j corresponds to an n dimensional point. We calculate the pairwise squared distance by $d_{ij}^2 = ||x_i - x_j||^2 = x_i^{\mathrm{T}} x_i + x_j^{\mathrm{T}} x_j - 2x_i^{\mathrm{T}} x_j$. Grouping the N^2 entries into matrix form, we have:

$$D^{(2)} = c\vec{1}^{\mathrm{T}} + \vec{1}c^{\mathrm{T}} - 2X^{\mathrm{T}}X \tag{1}$$

where c is a column vector with $x_i^{\mathrm{T}} x_i$ being the entries. We'll see later (**Section 4.1**) that once a matrix B can be written in the form $B = X^{\mathrm{T}}X$, we have standard ways to decompose it and recover the low dimensional embedding. That is, given dissimilarity measures, we can construct the corresponding inner product form. A standard approach is double centering: [6]

$$J = I - \frac{1}{n}\vec{1}\vec{1}^{\mathrm{T}} \tag{2}$$

$$X^{\mathrm{T}}X = -\frac{1}{2}JD^{(2)}J \tag{3}$$

There are yet two problems left:

- First, not all dissimilarities are distance. One reason is that the real world data is with noise. In such case, random noise will be reduced by SET. Another reason, maybe more frequently seen, is data inconsistency. This is quite often seen in computational psychology when people try to rate the degree of dissimilarity of objects[6]. Those empirical data may break distance laws, like triangle inequality.
- Second, in order to make double centering work, the low dimensional data are required to zero mean, namely $\sum_i x_i = 0$. Actually, we can choose any points as the origin(rather than sample mean), and corresponding forms can be derived. Since our input in this case is

pairwise distance(D or $D^{(2)}$), their is no explicit definition of origin. Or in another word, the effect of embedding is invariant under translation. We can safely locate the embedded sample mean at the origin.

In total, given dissimilarity matrix D, we take the element wise square $D^{(2)}$ and then pass the double centered version $-\frac{1}{2}JD^{(2)}J$ to next stage.

2.1.3 Enhancements

The above sections discussed the basic formulation of an adjacency matrix A. They can be fit direcly into SE procedures. At the same time, people propose some enhancement techniques based on certain assumptions.

Geodesic distance, as is in isomap[27], computes all pair shortest path (geodesic distance) of the adjacency graph first. Then the pairwise geodesic distance is treated as normal distance, D. The standard MDS (**Section 4.1**) can construct an embedding for D. For details, please see **Section 4.2**.

Although we have not found other widely used enhancements in literature, the discussion here do reveal other possibilities. For example, what will be the result if we plug in effective resistance as distance and fit in later SE stage? The effective resistance is closely related with commute time[18]. If it yield good results in practice, justification will not be hard.

2.2 Spectral Embeding

This stage takes a weighted graph adjacency matrix as input. The matrix can be derive from several starting points as described in the last section. Besides, enhancements may have already been performed. Regardless of their origins, we treat them as the adjacency matrix of similarity graph.

The output of SE stage is usually an $N \times d$ matrix Y. The columns of Y are d eigen vectors(or scaled version). The N rows of Y provides a d-dimensional Euclidean embedding.

2.2.1 Diagonal Shifting

Although off-diagonals of A are defined to be affinity / similarity / inner product, the definition of diagonal varies in different contexts.

In spectral community, the diagonals are interpreted as self-loops, and they play little role in objective definition. For example, in Normalized Cut (**Section 3.1.3**), self-loops does not influence the cut and their influence on the volume of each cluster is mighty, which could be absorbed into more general framwork (**Section 3.1.4**). So a natural operation is to zero out those self-loops as is in most SC literature.

In machine learning community, A is more probably expected to be a inner product matrix. In other words, it is Positive SemiDefinite(PSD). In

the language of kernels, it is a valid Gram matrix. For example, in **Section 2.1.1**, we use the Gaussian kernel $A_{ij} = \exp\{-||x_i - x_j||^2/t\}$ to weight edges. If we stick to this equation, $A_{ii} = 1$. That means a vertex has highest similarity with itself, which is naturally right. Most importantly, A may now be PSD and fits some DR techniques like MDS (**Section 4.1**), kernel PCA (**Section 4.7**), etc. If we zero out the diagonals, those algorithms can also be invoked and probably still output good results. However, the justification will be different.

Despite this operational difference, we notice that they yield essentially the same result following a diagonal shifting procedure [9]:

$$A' = \sigma I + A \tag{4}$$

where $\sigma = 1$ in our Gaussian kernel example. With large enough σ, any A can be transformed to an equivalent PSD version by linear algebraic argument. The only difference between A and A' is that they have different eigenvalues. Their eigenvectors are essentially the same.

2.2.2 Adjacency or Laplacian

We define the degree matrix R to be a diagonal matrix with $R_{ii} = \sum_j A_{ij}$. The Laplacian is defined as:

$$L = R - A \tag{5}$$

The use of adjacency or Laplacian is closely related with the justification. For adjacency matrix, the normalized version has analogy to a transition matrix of a Markov process. Besides, the adjacency matrix series SET often has angule preserving justification. For Laplacian, its variations may appear in cut or conductance like objectives.

Moreover, the Laplacian is PSD:

$$\begin{aligned} x^{\mathrm{T}}Lx &= \sum_i (\sum_j A_{ij}) x_i^2 - \sum_{ij} A_{ij} x_i x_j \\ &= \frac{1}{2} \sum_i \sum_j A_{ij}(x_i^2 + x_j^2) - \sum_{ij} A_{ij} x_i x_j \\ &= \frac{1}{2} \sum_i \sum_j A_{ij}(x_i - x_j)^2 \geq 0 \end{aligned} \tag{6}$$

and Laplacian has $\vec{1}$ as an eigen vector with smallest eigen value 0:

$$L\vec{1} = 0\vec{1} \tag{7}$$

In practice, the first eigen vector of Laplacian (or variations) is ignored, and successive smallest eigen vectors are used. For adjacency matrix, the first largest eigen vectors are used. This is natural due to the negation in **Eq 5**.

We leave further discussion of the use of adjacency or Laplacian to **Section 3** where the context is more clear.

2.2.3 Normalization

The intuition of normalization comes in two folds:

- In spectral graph theory justifications, like Ratio Cut (**Section 3.1.2**), Normalized Cut (**Section 3.1.3**), Weighted Cut (**Section 3.1.4**), the normalization can be interpreted as assigning different importance to nodes. For example, all 1's in Rcut and degree in Ncut. The general normalization can be described by a weighting matrix W with corresponding weights on the diagonal.
- In random walk series justifications, the walk matrix is required to be row-stochastic. Thus $W = R$ and left normalization is the standard operation.

We follow similar notations as [28] and collect possible normalization operations in **Tbl 2.2.3**.

Table 1: Normalization

Normalization	Adjacency	Laplacian
Unnormalized	A	L
Symmetric normalized	$A_{\text{sym}} = W^{-\frac{1}{2}}AW^{-\frac{1}{2}}$	$L_{\text{sym}} = W^{-\frac{1}{2}}LW^{-\frac{1}{2}}$
Left normalized	$A_{\text{rw}} = W^{-1}A$	$L_{\text{rw}} = W^{-1}A$

Assume $\lambda, \lambda_{\text{sym}}, \lambda_{\text{rw}}$ are eigenvalues of $A, A_{\text{sym}}, A_{\text{rw}}$ with corresponding eigen vectors $v, v_{\text{sym}}, v_{\text{rw}}$, their relationship can be shown through the following array of equations:

$$
\begin{aligned}
Av &= \lambda v && (8)\\
W^{-1}Av_{\text{rw}} = A_{\text{rw}}v_{\text{rw}} &= \lambda_{\text{rw}}v_{\text{rw}} && (9)\\
Av_{\text{rw}} &= \lambda_{\text{rw}}Wv_{\text{rw}} && (10)\\
W^{-\frac{1}{2}}W^{-\frac{1}{2}}AW^{-\frac{1}{2}}W^{\frac{1}{2}}v_{\text{rw}} = A_{\text{rw}}v_{\text{rw}} &= \lambda_{\text{rw}}v_{\text{rw}} && (11)\\
A_{\text{sym}}(W^{\frac{1}{2}}v_{\text{rw}}) &= \lambda_{\text{rw}}(W^{\frac{1}{2}}v_{\text{rw}}) && (12)\\
A_{\text{sym}}v_{\text{sym}} &= \lambda_{\text{sym}}v_{\text{sym}} && (13)
\end{aligned}
$$

Comparing **Eq 8** and **Eq 10**, we can see that left normalized version corresponds to solving a generalized eigen problem (A, W). It's obvious normalization makes a big difference in practice. On the other hand, there is only slight difference between two types of normalization. Comparing **Eq 12** and **Eq 13**, we know that they have the same eigen values and their eigen vectors differ by a scaling of $W^{\frac{1}{2}}$.

Similar relationship holds for $L, L_{\text{sym}}, L_{\text{rw}}$.

In the literatures, e.g. [4] [25], L_{sym} often appears as one intermediate step due to variable substitution and eigen vector of L_{rw} is usually the final result.

It's worth to note that which normalization to use is coupled with stages before and after and is also related to justification angle. It's very hard to reach a general conclusion. For example, Ng[20] actually uses a symmetric normalization and Shi[25] uses a left normalization. In [20], Ng claimed superior performance. On the other hand, Luxburg recommend left normalization in general([28] ch8). Some comparisons may be ill-posed because involved algorithms may use different matrix type and do different post-processing on eigenvectors. All the details contribute to performance differences on varied application scenarios.

2.2.4 Eigen Value Decomposition

Eigen Value Decomposition (EVD) and Singular Value Decomposition (SVD) are two important subroutines in SET. They are algebraically related [32]:

$$X = U\Sigma V^{\mathrm{T}} \tag{14}$$
$$XX^{\mathrm{T}} = U\Sigma V^{\mathrm{T}}V\Sigma U^{\mathrm{T}} \tag{15}$$
$$(XX^{\mathrm{T}})U = U\Sigma^2 \tag{16}$$

Similarly,

$$(X^{\mathrm{T}}X)V = V\Sigma^2 \tag{17}$$

Now suppose $X = [x_1, x_2, \ldots, x_N]$ is our data matrix. Some authors view PCA(**Section 4.5**) as an SVD on X ([6] ch24). Others view PCA (**Section 4.5**) EVD on $X^{\mathrm{T}}X$. **Eq 17** shows that they yield the same embedding V.

In our framework, the current SE step take adjacency matrix (or Laplacian) as input. It is essentially a matrix containing pairwise information. So EVD is performed in this stage. The comparison of EVD and SVD in this section is to show that under some SET's settings, the explicit construction of an equivalent similarity matrix (metric formulation in our framework) can be omitted.

2.2.5 Scaling and Projection

The scaling and projection described earlier are both regarded as post-processing of eigen vectors. For example, MDS(**Section 4.1**) and Brand's algorithm [7] scale the eigenvectors by square root eigenvalues; Ng's algorithm [20] projects the coordinates given by eigenvectors to unit sphere. A per case justification and discussion is better and we leave it to **Section 3**.

Despite of their own rationale, we provide one possible interpretation of scaling. Consider 0/1 version adjacency matrix A. If we raise it to the power A^k, the entry $(A^k)_{ij}$ just counts the number of k-hop paths from i to j. We call the graph corresponding to A^k as the "k-th order connectness graph".

The original matrix power is only defined for integer k. We notationally generalize it in the following way: (for symmetric A)

$$A^k = (U\Lambda U^{\mathrm{T}})^k = U\Lambda^k U^{\mathrm{T}} \tag{18}$$

$$A^x = U\Lambda^x U^{\mathrm{T}} \tag{19}$$

where x is a real value and $\Lambda^x = \mathrm{diag}(\lambda_1^x, \lambda_2^x, \ldots, \lambda_N^x)$. With this generalization, we can define eigen value scaling as a standard procedure of post-processing. For non-scaled algorithms, just let $x = 0$. This can be interpreted as working on the x-th power of adjacency matrix. It has the same effect if we raise the matrix to the x-th power in the enhancement stage(**Section 2.1.3**).

The justification for x-th power is: x controls the degree of graph distance we want to capture. For example, when $x \to \infty$, by the argument of power method, only the eigenvector with largest eigenvalue (principal eigenvector) is "kept". So the principal eigenvector provides the embedding which captures very "distant" relationships. The successive eigenvectors provides embedding which captures nearer and nearer relationships.

This justification has an analogy. Katz[17] provided an index to measure similarities of vertices in graphs: ([1] ch2.2)

$$\mathrm{Katz}(i,j) = \sum_{k=1}^{\infty} \beta^k (A^k)_{ij} \tag{20}$$

The Katz Index takes all "k-th order connectness graph" into consideration and the decaying factor β imposes a preference between near and distant connectness relationship.

One may think to scale the eigen vectors by a polynomial or more generally a function of eigen values, i.e.

$$Uf(\Lambda) \tag{21}$$

This operation risk losing good justifications. In our survey, we have only seen three choice of f:

- 0. It corresponds to the non-scaled case.
- $\Lambda^{\frac{1}{2}}$. This operation often roots from an error energy minimization / low-rank approximation view point(**Section 3.4**).
- Katz polynomial(**Eq 20**). On one hand, it has good justification by considering all length's connectness. On the other hand, we have a closed form for **Eq 20**:[1]

$$\mathrm{Katz} = (I - \beta A)^{-1} - I \tag{22}$$

Nevertheless, seeking for an application tailored $f(\Lambda)$ with good justification is still open work for practitioners.

2.3 Clustering

For K-means and hierarchical clustering, readers can consult standard data mining texts like [15]. For simpler hard cut algorithms, like using the largest entry of embedded coordinates as the cluster index, we omit the discussion in this section because their operation and justification are closely binded. As an example to stimulate exploring more hard cut techniques, we propose a variation of K-means. This variation should work better with Ng's [20] and Brand's [7] SE procedure.

Algorithm 2 Angular K-means

Input: output of SE: $Y = [y_1, y_2, \ldots, y_N]$; Number of Clusters K.
Output: Clustering $\{C_i\}$: $C_i \in V$ and $\cap_i C_i = \emptyset$ and $\cup_i C_i = V$.

1: Random initilize cluster centers $t_1^{(new)}, t_2^{(new)}, \ldots, t_K^{(new)}$.
2: **repeat**
3: $t_i^{(old)} \leftarrow t_i^{(new)}$
4: $l_i \leftarrow \arg\max_j < y_i, t_j^{(old)} >, \forall i = 1, 2, \ldots, N$
5: $t_j^{(new)} \leftarrow \sum_{i=1}^{N} [l_i = j] x_i, \forall j = 1, 2, \ldots, K$ {[.] is indicator function}
6: Normalize $t_j^{(new)}, \forall j = 1, 2, \ldots, K$
7: **until** $(\sum_{j=1}^{K} (1 - < t_j^{(new)}, t_j^{(old)} >)) < \epsilon$
8: $C_j = \{i | l_i = j\}$

In Ng's [20] and Brand's [7] work, their low dimensional points are projected onto a unit sphere. It is more reasonable to cluster according to the angles in this case. Our variation of K-means (**Alg 2**) takes the embedding property into consideration. Again the hard cut stage is not the focus of SC or SE study, so little work has been found to explore suitable hard cut algorithms. As long as the SE stage induces well located clusters, Euclidean K-means can easily detect them. As reported in [20], with special initialization of K-means centroids, only one recursion is needed for convergence.

3 Spectral Clustering Justification

In this section, we collect justifications from different authors. We will see that not all of the combinations provided in **Section 2** have corresponding justifications.

3.1 Combinatoric Justification

The traditional and most widely study is combinatoric justification. The idea is centered on the concept of "cut". This section is adapted from [25] [28] [9].

3.1.1 Cut

Suppose C_1 and C_2 are two subsets of V. The cut between C_1 and C_2 is defined as:

$$\text{cut}(C_1, C_2) = \sum_{u \in C_1, v \in C_2, (u,v) \in E} A_{uv} \tag{23}$$

The volume of C_1 is defined as:

$$\text{vol}(C_1) = \text{cut}(C_1, V) = \sum_{u \in C_1, v \in V, (u,v) \in E} A_{uv} = \sum_{v \in C_1} d(v) \tag{24}$$

where $d(v) = \sum_{u \in V} A_{vu}$ is the degree of vertex v.

The most straightforward trial to obtain a good clustering is given by the following optimization problem:

$$\underset{\{C_1, C_2, \ldots, C_k\}}{\text{minimize}} \sum_{i=1}^{k} \text{cut}(C_i, V - C_i) \tag{25}$$

Using **Eq 6**, it can be converted to an equivalent form:

$$\underset{\{C_1, C_2, \ldots, C_k\}}{\text{minimize}} \sum_{i=1}^{k} \chi_{C_i}^{\mathrm{T}} L \chi_{C_i} \tag{26}$$

where χ_{C_i} is the characteristic vector of C_i, defined as:

$$\chi_{C_i}(v) = \begin{cases} 1 & v \in C_i \\ 0 & \text{Else} \end{cases} \tag{27}$$

A more compact form of the above objective is:

$$\underset{\{C_1, C_2, \ldots, C_k\}}{\text{minimize}} \ \text{Tr}\left[\chi^{\mathrm{T}} L \chi\right] \tag{28}$$

where $\chi = [\chi_{C_1}, \chi_{C_2}, \ldots, \chi_{C_k}]$. This kind of combinatoric problem is shown to be NP-Hard by previous authors. Using standard spectral argument[28], we have one important observation that if the graph is diconnected and contains k connected components, the first k eigen vectors of Laplacian (count from smallest eigen value) are just the linear combination of the characteristic vectors of those connected components. In other words, in such ideal case, those eigen vectors are piecewise linear. Standard clustering algorithms in Euclidean space can easily give the right clustering. So one heuristic is to relax the problem to permit real values, i.e. substitute χ_{C_i} by v_i:

$$\underset{v_i \in \mathbb{R}^N, v_i^{\mathrm{T}} v_j = 0 V = (v_1, v_2, \ldots, v_N)}{\text{minimize}} \ \text{Tr}\left[V^{\mathrm{T}} L V\right] \tag{29}$$

The orthogonal condition $v_i^{\mathrm{T}} v_j = 0$ is inherited from the fact that characteristic vectors are orthogonal.

Now we find the problem of **Eq 29** is poorly-defined. Without further constraints, the choice $V = 0$ yield the minimum but it's obviously contrary to our real objective. Note that the unrelaxed version is well-defined, for the value of χ_{C_i} is constrained to $\{0, 1\}$. Certain "normalization" helps to well define our objective. In the following part of this section, we'll see how different normalization induce different objectives.

3.1.2 Ratio Cut

The first attempt of normalization is to constrain V to unit vectors. In this way, the following optimization problem can absorb proportional scaling of V and also prevent the trivial solution mentioned above:

$$\begin{aligned} \underset{v_i \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{i=1}^{k} v_i^{\mathrm{T}} L v_i \\ s.t. \quad & v_i^{\mathrm{T}} v_i = 1, \forall i = 1, 2, \ldots, k \\ & v_i^{\mathrm{T}} v_j = 0, \forall i \neq j \text{ and } i, j = 1, 2, \ldots, k \end{aligned} \tag{30}$$

This problem is equivalent to the following one:

$$\begin{aligned} \underset{v_i \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{i=1}^{k} \frac{v_i^{\mathrm{T}} L v_i}{v_i^{\mathrm{T}} v_i} \\ s.t. \quad & v_i^{\mathrm{T}} v_i = 1, \forall i = 1, 2, \ldots, k \\ & v_i^{\mathrm{T}} v_j = 0, \forall i \neq j \text{ and } i, j = 1, 2, \ldots, k \end{aligned} \tag{31}$$

Suppose we have a solution given by $\{v_i^*\}$. Then $\{t_i v_i^*\}, \forall t_i \in \mathbb{R}^N$ induce the same objective by linear algebraic argument. Then the problem can be tackled in two steps:

1. Solve:

$$\begin{aligned} \underset{v_i \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{i=1}^{k} \frac{v_i^{\mathrm{T}} L v_i}{v_i^{\mathrm{T}} v_i} \\ s.t. \quad & v_i^{\mathrm{T}} v_j = 0, \forall i \neq j \text{ and } i, j = 1, 2, \ldots, k \end{aligned} \tag{32}$$

2. Scale v_i so that $v_i^{\mathrm{T}} v_i = 1$.

The second step is easy, so we focus on the first step. By applying the result of Rayleigh Quotient repeatedly[18], it can be shown the solution is given by the k smallest eigenvectors of L.

Like the above section, **Eq 32** is a relaxed version of some problem. By reverting the relaxation process, we can find the combinatoric justification it corresponds to. Now we substitue orthogonal $\{v_i\}$ in **Eq 32** with

characteristic vector $\{\chi_{C_i}\}$ to see its combinatoric version:

$$\begin{aligned} & \sum_{i=1}^{k} \frac{\chi_{C_i}^{\mathrm{T}} L \chi_{C_i}}{\chi_{C_i}^{\mathrm{T}} \chi_{C_i}} \\ = & \sum_{i=1}^{k} \frac{\sum_{(u,v)\in E} A_{uv}(\chi_{C_i}(v) - \chi_{C_i}(u))^2}{|C_i|} \\ = & \sum_{i=1}^{k} \frac{\sum_{(u,v)\in E, u\in C_i, v\in V-C_i} A_{uv}}{|C_i|} \\ = & \sum_{i=1}^{k} \frac{\mathrm{cut}(C_i, V - C_i)}{|C_i|} \end{aligned} \tag{33}$$

Note that the last line in **Eq 33** is right the definition of Ratio Cut(RCut), given a clustering $\{C_1, C_2, \ldots, C_k\}$. Comparing it to our first objective, Cut(**Eq 25**), we find that RatioCut takes cluster size into consideration. This is more reasonable if we consider the existence of outliers. In the extreme case, one outlier may have no connection with other points. The Cut objective will induce a singleton cluster for this outlier. However, RCut may tend to partition the graph into bigger clusters, which is more close to our goal.

3.1.3 Normalized Cut

Normalized Cut(NCut) is another widely studied combinatoric objective. Following the same approach as last section, we can derive its combinatoric version and relaxed version.

The definition of NCut is:

$$\mathrm{NCut} = \sum_{i=1}^{k} \frac{\mathrm{cut}(C_i, V - C_i)}{\mathrm{vol}(C_i)} \tag{34}$$

Its corresponding relaxed version is:

$$\begin{aligned} \underset{v_i \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{i=1}^{k} v_i^{\mathrm{T}} L v_i \\ s.t. \quad & v_i^{\mathrm{T}} R v_i = 1, \forall i = 1, 2, \ldots, k \\ & v_i^{\mathrm{T}} v_j = 0, \forall i \neq j \text{ and } i, j = 1, 2, \ldots, k \end{aligned} \tag{35}$$

where R is the degree matrix.

From the optimization perspective, the difference between RCut and NCut is that the normalization constraints are $v_i^{\mathrm{T}} I v_i = 1$ and $v_i^{\mathrm{T}} R v_i = 1$, respectively. From the combinatoric perspective, NCut takes the "importance" of vertices into consideration while RCut does not distinguish them.

The "importance" here is expressed using degree of nodes. In **Section 3.1.4**, we'll see a generalzation.

3.1.4 Weighted Cut

The Cut defined in **Eq 25** can capture the linkage between clusters. Minimizing it should result in reasonable clustering. However, in real applications, vertices may be of different importance. If we denote the vertex weight by diagonal entries of W, the normalization constraint $v_i^{\mathrm{T}} W v_i = 1$ has a more general combinatoric correspondance:

$$\mathrm{WCut} = \sum_{i=1}^{k} \frac{\mathrm{cut}(C_i, V - C_i)}{W(C_i)} \tag{36}$$

where $W(C_i) = \sum_{v \in C_i} W_{vv}$.

By letting $W = I$ or $W = D$, it degrades to RCut and NCut, respectively.

3.1.5 Ratio/Normalized/Weighted Association

Recall the Cut series objectives try to minimize the inter cluster linkage. Likewise, one may think to maximize the intra cluster linkage, given by the association:

$$\mathrm{assoc}(C_1) = \mathrm{cut}(C_1, C_1) = \sum_{u,v \in C_1, (u,v) \in E} A_{uv} \tag{37}$$

The Ratio Association(RAssoc) is defined as:

$$\mathrm{RAssoc} \quad = \quad \sum_i \frac{\mathrm{assoc}(C_i)}{|C_i|} \tag{38}$$

We can derive its corresponding relaxed optimization:

$$\begin{aligned} \underset{v_i \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{i=1}^{k} v_i^{\mathrm{T}} A v_i \\ s.t. \quad & v_i^{\mathrm{T}} v_i = 1, \forall i = 1, 2, \ldots, k \\ & v_i^{\mathrm{T}} v_j = 0, \forall i \neq j \text{ and } i, j = 1, 2, \ldots, k \end{aligned} \tag{39}$$

The Normalized Association(NAssoc) is defined as:

$$\mathrm{NAssoc} \quad = \quad \sum_i \frac{\mathrm{assoc}(C_i)}{\mathrm{vol}(C_i)} \tag{40}$$

We can derive its corresponding relaxed optimization:

$$\begin{aligned} \underset{v_i \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{i=1}^{k} v_i^{\mathrm{T}} A v_i \\ s.t. \quad & v_i^{\mathrm{T}} D v_i = 1, \forall i = 1, 2, \ldots, k \\ & v_i^{\mathrm{T}} v_j = 0, \forall i \neq j \text{ and } i, j = 1, 2, \ldots, k \end{aligned} \tag{41}$$

Of course, the Weighted Association(WAssoc) is defined as:

$$\text{WAssoc} = \sum_i \frac{\text{assoc}(C_i)}{W(C_i)} \tag{42}$$

We can derive its corresponding relaxed optimization:

$$\begin{aligned} \underset{v_i \in \mathbb{R}^N}{\text{minimize}} \quad & \sum_{i=1}^{k} v_i^{\mathrm{T}} A v_i \\ s.t. \quad & v_i^{\mathrm{T}} W v_i = 1, \forall i = 1, 2, \ldots, k \\ & v_i^{\mathrm{T}} v_j = 0, \forall i \neq j \text{ and } i, j = 1, 2, \ldots, k \end{aligned} \tag{43}$$

The difference between association series and cut series is that they use A and L, respectively. This can justify why decomposing adjacency and Laplacian are both reasonable to get good embedding. However, in our survey, the use of Laplacian is obviously dominant. There are other properties making Laplacian superior that can not be justified using the combinatoric framework in this section(see **Section 4.3** for example).

3.1.6 Conductance and Expansion

For the last part of combinatoric justification, we quick note two other criteria: ([18], W2)

$$\text{Expansion} = \sum_i \frac{\text{cut}(C_i, V - C_i)}{\min\{|C_i|, |V - C_i|\}} \tag{44}$$

$$\text{Conductance} = \sum_i \frac{\text{cut}(C_i, V - C_i)}{\min\{\text{vol}(C_i), \text{vol}(V - C_i)\}} \tag{45}$$

The notion of expasion and conductance is studied in many spectral graph theory problems. They have close analogy to RCut and NCut. However, due to the min operator, it becomes more difficult to establish their relaxed version. In our survey so far, there are no sophisticated approach to tackle with the two objectives.

3.2 Random Walk

[von]

3.3 Matrix Perturbation

[andrew ng]

3.4 Low Rank Approximation

[matthew brand]

3.5 Density Estimation View

[mo chen, 2010]

3.6 Commute Time

[jihun ham, kernel]. view pseudo inverse of graph Laplacian by commute times on graphs.

3.7 Polarization

[m. brand] unifying view...

References

[1] C.C. Aggarwal. *Social network data analytics*. Springer-Verlag New York Inc, 2011.

[2] R. Andersen and F. Chung. Detecting sharp drops in pagerank and a simplified local partitioning algorithm. *Theory and Applications of Models of Computation*, ¡¡missing¿¿:1–12, 2007.

[3] R. Andersen and Y. Peres. Finding sparse cuts locally using evolving sets. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 235–244. ACM, 2009.

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[5] C.M Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

[6] I. Borg and P.J.F. Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Verlag, 2005.

[7] M. Brand and K. Huang. A unifying theorem for spectral embedding and clustering. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[8] F. Chung. Random walks and local cuts in graphs. *Linear Algebra and its applications*, 423(1):22–32, 2007.

[9] I. Dhillon, Y. Guan, and B. Kulis. A unified view of kernel k-means, spectral clustering and graph cuts. Technical report, Univ. of Texas at Austin, 2004.

[10] S.W. Hadley, B.L. Mark, and A. Vannelli. An efficient eigenvector approach for finding netlist partitions. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 11(7):885–892, 1992.

[11] B. Hendrickson and R. Leland. Multidimensional spectral load balancing. *Report SAND93-0074, Sandia National Laboratories, Albuquerque, NM*, ¡¡missing¿¿:¡¡missing¿¿, 1993.

[12] Pili Hu. Matrix calculus. GitHub, https://github.com/hupili/tutorial/tree/master/matrix-calculus, 3 2012. HU, Pili's tutorial collection.

[13] Pili Hu. Spectral techniques for community detection on 2-hop topology. GitHub, https://github.com/hupili/Spectral-2Hop, 4 2012. course project of CUHK/CSCI5160.

[14] Pili Hu. Tutorial collection. GitHub, https://github.com/hupili/tutorial, 3 2012. HU, Pili's tutorial collection.

[15] H. Jiawei and M. Kamber. *Data mining: concepts and techniques*, volume 5. San Francisco, CA, itd: Morgan Kaufmann, 2001.

[16] R. Kannan, S. Vempala, and A. Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.

[17] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[18] Lap Chi Lau. Spectral algorithm lecture notes. http://www.cse.cuhk.edu.hk/ chi/csc5160/index.html, 2012.

[19] L. Lovász and M. Simonovits. The mixing rate of markov chains, an isoperimetric inequality, and computing the volume. In *Foundations of Computer Science, 1990. Proceedings., 31st Annual Symposium on*, pages 346–354. IEEE, 1990.

[20] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[21] S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[22] L.K. Saul and S.T. Roweis. An introduction to locally linear embedding. Technical report, NYU, 2000.

[23] L.K. Saul and S.T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, 4:119–155, 2003.

[24] B. Schölkopf, A. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[25] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[26] DA Spielman. Spectral graph theory lecture notes. http://www.cs.yale.edu/homes/spielman/561/, 2011.

[27] J.B. Tenenbaum, V. De Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[28] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[29] Wikipedia, Jaccard Coefficient, `http://en.wikipedia.org/wiki/Jaccard_index`

[30] Wikipedia, Mahalanobis Distance, `http://en.wikipedia.org/wiki/Mahalanobis_distance`

[31] Wikipedia, Cosine Similarity, `http://en.wikipedia.org/wiki/Cosine_similarity`

[32] Wikipedia, Singular Value Decomposition `http://en.wikipedia.org/wiki/Singular_value_decomposition`

Appendix