Automatic Determination of Clusters

Bettina Hoser and Jan Schröder

Chair for Information Services and Electronic Markets, Universität Karlsruhe(TH), Germany
Bettina.Hoser@em.uni-karlsruhe.de
Jan.Schroeder@em.uni-karlsruhe.de

Summary. In this paper we propose an automatic method for spectral clustering of weighted directed graphs. It is based on the eigensystem of a complex Hermitian adjacency matrix $H_{n\times n}$. The number of relevant clusters is determined automatically. Nodes are assigned to clusters using the inner product matrix $S_{n\times n}$ calculated from a matrix $R_{n\times l}$ of the l eigenvectors as column vectors which correspond to the positve eigenvalues of H. It can be shown that by assigning the vertices of the network to clusters such that a node i belongs to cluster p_c if $Re(S_{i,p_c}) = max_j Re(S_{i,j})$ an good partitioning can be found. Simulation results are presented.

1 Spectral Clustering Method

Spectral clustering of a graph by means of the eigensystem has a long tradition in graph theory. In 1970 Fiedler showed, that the the second smallest eigenvalue λ_{n-1} of a Laplace matrix depicts a measure for the algebraic connectivity of a graph [6]. Within the field of spectral clustering since then several algorithms were proposed that can be distinguished by the determination of the partitions. Some derive the partitions by doing recursive bipartitioning (e.g. [5, 8]), others directly derive the partitions by predefining the number of clusters (e.g. [4, 9, 1]) and Zien, Chan and Schlag [7] have proposed a hybrid method. The crucial point is always the definition of the initial clusters centers and their number. The second relevant step is the assignment of vertices to clusters. Especially for large graphs the predefinition of the number of clusters is difficult. If the partition into k clusters (k-partition) is derived directly, it is a question whether to use k eigenvectors which Chan et al. as well as Ng et al. [4, 9] did, or as Alpert and Yao suggested [1] to use many or even all eigenvectors.

We use an approach in between by taking the eigenvectors corresponding to all positive eigenvalues of the eigensystem (see Section 4). The spectral decomposition algorithm was proposed by Hoser and Geyer-Schulz [2] which is used to find structural details of the given directed and weighted network. The eigensystem of the graph is subsequently used to determine the clusters. Compared to many algorithms in the spectral partitioning field which work on undirected graphs like [4, 9, 8, 5] this algorithm and therefore the clustering works also on directed, weighted graphs.

2 Notations and Definitions

First we introduce the notation and some basic facts about complex numbers, Hilbert space and eigensystems of Hermitian matrices.

A complex number z can be represented in algebraic form or equivalently in exponential form as $z = a + ib = |z|e^{i\phi}$ with the real part of z being denoted as $Re(z) = a$, the imaginary part as $Im(z) = b$, the absolute value as $|z| = \sqrt{a^2 + b^2}$, and the phase as $0 \leq \phi = \arccos \frac{Re(z)}{|z|} \leq \pi$, with i as the imaginary unit ($i^2 = -1$). $\overline{z} = a - ib$ denotes the complex conjugate of z. Vectors will be complex valued column vectors unless otherwise stated.

Hilbert space is a complete normed inner product space. With the inner product of $\mathbf{x}, \mathbf{y} \in \mathit{\Gamma}^n$ being defined as $\langle \mathbf{x} \mid \mathbf{y} \rangle = \mathbf{x}^* \mathbf{y} = \sum_{k=1}^{n} \overline{x_k} y_k$, the norm is given defined as in Equation(1):

$$\sqrt{\langle \mathbf{x} \mid \mathbf{x} \rangle} = \| \mathbf{x} \| \tag{1}$$

From this it is clear that to minimize the distance between two points, characterized by vectors $\mathbf{x}, \mathbf{y}$ in an n-dimensional vector space is equivalent to maximizing the real part of the inner product of both. Let furthermore $\mathbf{x}, \mathbf{y}$ be normalized as in Equation(2),

$$\|\mathbf{x}\| = \|\mathbf{y}\| = 1 \tag{2}$$

then the distance between the two vectors is zero if the real part is equal to one, which is the maximum of the inner product (Equation(3)) between the vectors $\mathbf{x}$ and $\mathbf{y}$

$$\begin{aligned} \|\mathbf{x} - \mathbf{y}\|^2 &= \langle \mathbf{x} - \mathbf{y} | \mathbf{x} - \mathbf{y} \rangle \\ &= \langle \mathbf{x} | \mathbf{x} \rangle + \langle \mathbf{y} | \mathbf{y} \rangle - \langle \mathbf{x} | \mathbf{y} \rangle - \langle \mathbf{y} | \mathbf{x} \rangle \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \langle \mathbf{x} | \mathbf{y} \rangle - \overline{\langle \mathbf{x} | \mathbf{y} \rangle} \\ &= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2Re(\langle \mathbf{x} | \mathbf{y} \rangle) \\ &= 2 - 2Re(\langle \mathbf{x} | \mathbf{y} \rangle). \end{aligned} \tag{3}$$

A matrix H is called Hermitian, if and only if $H^* = H$ with H^* representing the conjugate complex transpose of H. Hermitian matrices are also normal $HH^* = H^*H$. All eigenvalues of a Hermitian matrix are real and all eigenvectors are orthogonal and can be chosen as to define a complete orthonormal basis. Thus a matrix H can be written as the weighted sum of eigenprojectors $H = \sum_{k=1}^{n} \lambda_k P_k$ with λ_k representing the eigenvalues and $P_k = \mathbf{x_k} \mathbf{x_k}^*$ the orthogonal projector corresponding to λ_k.

The outer product of two column vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as in Equation(4):

$$\mathbf{x}\mathbf{y}^* = \begin{pmatrix} x_1\overline{y_1} & \dots & x_1\overline{y_n} \\ \dots & \dots & \dots \\ x_n\overline{y_1} & \dots & x_n\overline{y_n} \end{pmatrix} \tag{4}$$

Since the eigenvalues are real they can be sorted $\lambda_1 \geq \dots \geq \lambda_n$ and thus can help to identify the dominant substructures in a network, since the eigenvalues are the weights of the eigenspaces.

Following the concept of eigenvector centrality as described by Wasserman and Faust [10] we identify the most central vertex v_m in a graph $G(V,E)$ by its absolute value $|x_{1,m}|$ of the eigenvector component corresponding to the largest eigenvalue λ_1. This also holds for the most central vertices in each substructure identified by the