The Graph Laplacian and Data Clustering

Derek Smith

October 3, 2015

Consider the social network of seven individuals with the unimaginative names A, B, C, D, E, F and G. An *edge* connects each pair of friends. This network or graph consists of two smaller, distinct graphs or components.

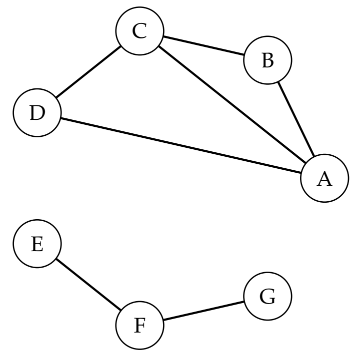

Figure 1: A 7-person social network.

Question 1 *Who would you suggest person B befriend?*

It seems reasonable to ignore people in the component $\{E, F, G\}$. Person B is contained in the component $\{A, B, C, D\}$ and is friends with both A and C. Thus you would suggest that B befriend D.

Question 2 *How to write an algorithm which suggests that B befriend D?*

By the above reasoning, we require a program which produces the components of the social network. Then we could simply check if B is already friends with everyone in her component.

Question 3 *How to algorithmically produce the components of the graph?*

Linear algebra provides a solution. In the above graph, the individual named A has three friends. In the language of graph theory, the degree of vertex A equals 3. The computer will be fed the graph Laplacian, a matrix defined via the formula:

$$L = (a_{ij}) = \begin{cases} \text{degree of vertex } i \text{ along the diagonal} \\ -1 \text{ when an edge connects vertices } i \text{ and } j. \end{cases}$$

For the network of seven friends, the Laplacian matrix looks like:

$$L = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 \\ -1 & 0 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \tag{1}$$

where rows are in alphabetical order. Note the symmetric, block-diagonal structure of the matrix. Additionally, the nullspace of L must have dimension at least two. The following calculation exhibits one vector in the nullspace.

$$L\vec{v} = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 \\ -1 & 0 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{2}$$

Exercise 1 *Can you find a second independent vector in the nullspace of L?*

Exercise 2 *Suppose $\vec{v}$ is a vector in the nullspace of L whose components are either zero or one. Describe how to find the corresponding connected component of the graph.*

Extending these ideas leads to the following theorem.

Theorem 1 *The dimension of the nullspace of the Laplacian matrix equals the number of components of the corresponding graph. There exists a basis for the nullspace with each vector having components either zero or one.*

This theorem provides a computational method to find the components of a graph: simply find a basis for the nullspace consisting of "binary" basis vectors as above. There exist computationally more efficient methods to find the components of a graph, so why approach the problem using linear algebra?

- A real-life social graph probably looks like the figure on the right, which lacks any distinct components; the previous analysis would fail. However, this graph exhibits "approximate components" or clusters which have been colored for visualization purposes.
- The magnitude of the off-diagonal entries of the Laplacian are either 0 or 1. We may wish to replace these integers with any number between 0 and 1 to indicate an affinity between two people.

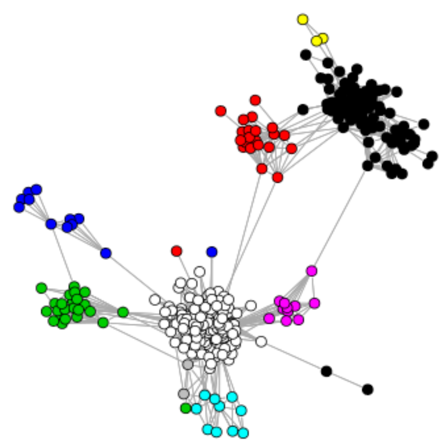

Figure 2: Cluster your Facebook friends

A technique called spectral clustering uses the Laplacian matrix to detect approximate clusters in these situations. The cluster detection problem is called cluster analysis and there are many approaches to the problem besides spectral clustering.[1]

If we zero-out the diagonal of the Laplacian matrix, we get (the negative of) a related object called the adjacency matrix of the graph. In the case of non-integer values, this is called the transition matrix. The spectral analysis of these matrices finds application in Markov chains, a multi-billion-dollar example of which is Google's PageRank algorithm.

[1] The word spectral in this context refers to the analysis of eigenvectors and eigenvalues of a matrix. Recall that vectors in the nullspace are eigenvectors associated with eigenvalue zero.