Networks and Complexity

Solution 21-2

This is an example solution from the forthcoming book Networks and Complexity.
Find more exercises at https://github.com/NC-Book/NCB

Ex 21.2: Spectral methods in the wild [3]

A supermarket chain approaches you. Their goal is to identify different groups in their customer base. They have a large dataset that containing 20 million customers, for each customer the dataset describes which products they have bought in the past. Describe in bullet points how you would approach this question.

Solution

- We can represent the customers previous purchases as a table. This table could state the rate at which the customer buys an item or just whether the customer has ever bought the item. For example it could be something line this

	Product 1	Product 2	Product 3	...
Customer 1	1	0	0	...
Customer 2	1	1	0	...
Customer 3	0	0	1	...
Customer 4	1	1	0	...
⋮				

- We standardize the products such that every column in the table has mean zero and variance one
- Now we regard the rows as vectors and define a distance d_{ij} between customers as the euclidian distance between their respective vectors
- We define a similarity $s_{ij} = 1/d_{ij}$.
- We threshold the similarities. For example a similarity s_{ij} is retained if it is among the top-10 similarity scores for either customer i or j. All other s_{ij} are set to zero.
- We now regard the s_{ij} as elements of a weighted adjacency $\mathbf{s}$ and construct the corresponding Laplacian $\mathbf{L} = -\mathbf{s} + \mathbf{D}$, where $\mathbf{D} = \delta_{ij} \sum_k s_{ik}$.
- We compute the second smallest eigenvalue of $\mathbf{L}$ and the corresponding eigenvector $\boldsymbol{v}$.
- The customers who correspond to a positive entry in the eigenvector are identified as a cluster, the customers who correspond to a negative entry in the eigenvector form another cluster.
- Additional eigenvectors can be used to split the clusters further.