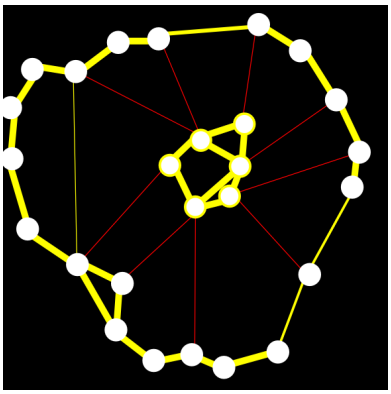

Figure 20.6: A cut involving the set of nodes in the center and the nodes on the perimeter.

The above concepts play a crucial role in the theory of normalized cuts. This beautiful and deeply original method first published in Shi and Malik [160], has now come to be a "textbook chapter" of computer vision and machine learning. It was invented by Jianbo Shi and Jitendra Malik and was the main topic of Shi's dissertation. This method was extended to $K \geq 3$ clusters by Stella Yu in her dissertation [191] and is also the subject of Yu and Shi [193].

Given a set of data, the goal of clustering is to partition the data into different groups according to their similarities. When the data is given in terms of a similarity graph G, where the weight $w_{i\,j}$ between two nodes v_i and v_j is a measure of similarity of v_i and v_j, the problem can be stated as follows: Find a partition $(A_1, \ldots, A_K)$ of the set of nodes V into different groups such that the edges between different groups have very low weight (which indicates that the points in different clusters are dissimilar), and the edges within a group have high weight (which indicates that points within the same cluster are similar).

The above graph clustering problem can be formalized as an optimization problem, using the notion of cut mentioned earlier. If we want to partition V into K clusters, we can do so by finding a partition $(A_1, \ldots, A_K)$ that minimizes the quantity

$$\mathrm{cut}(A_1, \ldots, A_K) = \frac{1}{2}\sum_{i=1}^{K} \mathrm{cut}(A_i) = \frac{1}{2}\sum_{i=1}^{K} \mathrm{links}(A_i, \overline{A}_i).$$

For $K = 2$, the mincut problem is a classical problem that can be solved efficiently, but in practice, it does not yield satisfactory partitions. Indeed, in many cases, the mincut solution separates one vertex from the rest of the graph. What we need is to design our cost function in such a way that it keeps the subsets A_i "reasonably large" (reasonably balanced).

An example of a weighted graph and a partition of its nodes into two clusters is shown in Figure 20.7.