Bachelor of Science (Third Year)
CSD107- Introduction to Data Analytics
Title of the Unit : Mining Social-Network Graphs
Module Name : Partitioning of graphs, Neighbourhood properties in graphs
Module Number : 19

Partitioning of Graphs

In this section, we examine another approach to organizing social-network graphs. We use some important tools from matrix theory to formulate the problem of partitioning a graph to minimize the number of edges that connect different components.

The goal of minimizing the "cut" size needs to be understood carefully before proceeding. For instance, if you just joined Facebook, you are not yet connected to any friends. We do not want to partition the friends graph with you in one group and the rest of the world in the other group, even though that would partition the graph without there being any edges that connect members of the two groups. This cut is not desirable because the two components are too unequal in size.

What Makes a Good Partition?

Given a graph, we would like to divide the nodes into two sets so that the cut, or set of edges that connect nodes in different sets is minimized. However, we also want to constrain the selection of the cut so that the two sets are approximately equal in size.

Recall our running example of the graph in Fig. 10.1. There, it is evident that the best partition puts {A, B, C} in one set and {D, E, F, G} in the other. The cut consists only of the edge (B, D) and is of size 1. No nontrivial cut can be smaller.

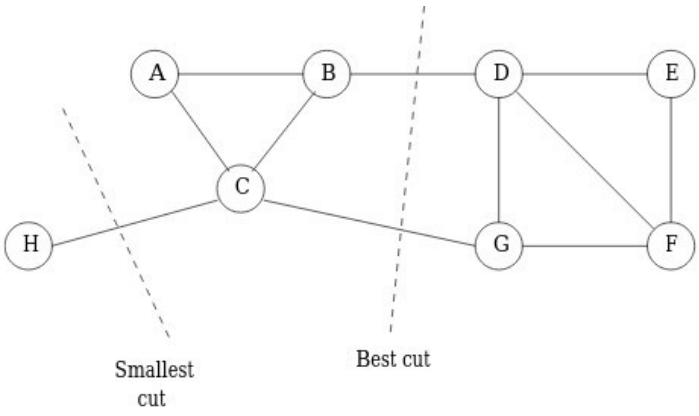

Figure 10.11: The smallest cut might not be the best cut

In Fig. 10.11 is a variant of our example, where we have added the node H and two extra edges, (H, C) and (C, G). If all we wanted was to minimize the size of the cut, then the best choice would be to put H in one set and all the other nodes in the other set. But it should be apparent that if we reject partitions where one set is too small, then the best we can do is to use the cut consisting of edges (B, D) and (C, G), which partitions the graph into two equal-sized sets {A, B, C, H} and {D, E, F, G}.

Neighborhood Properties in Graphs

There are several important properties of graphs that relate to the number of nodes one can reach from a given node along a short path.

In this section we look at algorithms for solving problems about paths and neighborhoods for very large graphs. In some cases, exact solutions are not feasible for graphs with millions of nodes. We therefore look at approximation algorithms as well as exact algorithms.

Directed Graphs and Neighborhoods

In this section we shall use a directed graph as a model of a network.

A directed graph has a set of nodes and a set of arcs; the latter is a pair of nodes written u→v. We call u the source and v the target of the arc. The arc is said to be from u to v.

Many kinds of graphs can be modeled by directed graphs.

- The Web is a major example, where the arc u→v is a link from page u to page v.
- Or, the arc u→v could mean that telephone subscriber u has called subscriber v in the past month.
- For another example, the arc could mean that individual u is following individual v on Twitter.
- In yet another graph, the arc could mean that research paper u references paper v.

Moreover, all undirected graphs can be represented by directed graphs. Instead of the undirected edge (u, v), use two arcs u→v and v→u. Thus, the material of this section also applies to graphs that are inherently undirected, such as a friends graph in a social network.

A path in a directed graph is a sequence of nodes v0, v1, . . . , vk such that there are arcs vi→vi+1 for all i= 0,1, . . . , k−1. The length of this path is k, the number of arcs along the path. Note that there are k+ 1 nodes in a path of length k, and a node by itself is considered a path of length 0.

The neighborhood of radius d for a node v is the set of nodes u for which there is a path of length at most d from v to u. We denote this neighborhood by N(v, d).

For example, N(v,0) is always{v}, and N(v,1) is v plus the set of nodes to which there is an arc from v.

More generally, if V is a set of nodes, then N(V, d) is the set of nodes u for which there is a path of length d or less from at least one node in the set V.

The neighborhood profile of a node v is the sequence of sizes of its neighbor-hoods |N(v,1)|,|N(v,2)|, ...

We do not include the neighborhood of distance 0, since its size is always 1.

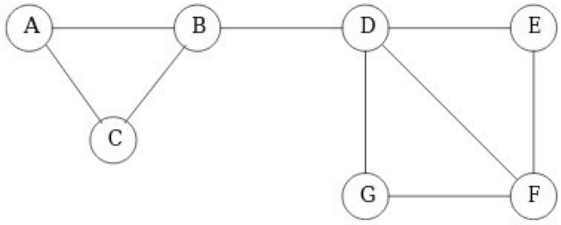

Figure 10.23: Our small social network; think of it as a directed graph

Consider the undirected graph of Fig. 10.1, which we re-produce here as Fig. 10.23. To turn it into a directed graph, think of each edge as a pair of arcs, one in each direction. For instance, the edge (A, B) becomes the arcs A→B and B→A.

First, consider the neighborhoods of node A.

We know N(A,0) ={A}.

Moreover, N(A,1) ={A, B, C}, since there are arcs from A only to B and C.

Furthermore, N(A,2) ={A, B, C, D}and N(A,3) ={A, B, C, D, E, F, G}.

Neighborhoods for larger radius are all the same as N(A,3).

On the other hand, consider node B. We find N(B,0) ={B},N(B,1) ={A, B,C, D}, and N(B,2) ={A, B, C, D, E, F, G}.

We know that Bis more central to the network than A, and this fact is reflected by the neighborhood profiles of the two nodes. Evidently, B is more central than A, because at every distance, its neighborhood is at least as large as that of A.

In fact, D is even more central than B, because its neighborhood profile dominates the profile of each of the nodes.

The Diameter of a Graph

The diameter of a directed graph is the smallest integer d such that for every two nodes u and v there is a path of length d or less from u to v.

In a directed graph, this definition only makes sense if the graph is strongly connected; that is, there is a path from any node to any other node.

If the graph is undirected, the definition of diameter is the same as for directed graphs, but the path may traverse the undirected edges in either direction. That is, we treat an undirected edge as a pair of arcs, one in each direction. The notion of diameter makes sense in an undirected graph as long as that graph is connected.

For the graph of Fig. 10.23, the diameter is 3. There are some pairs of nodes, such as A and E, that have no path of length less than 3. But every pair of nodes has a path from one to the other with length at most 3.