Data Mining

Graph Mining – Graph Clustering

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

/

Graph Clustering

- Definition: Cutting a graph into pieces where each piece is a cluster.

- Goal: Vertices in the same cluster are well connected and vertices in different clusters are *not necessarily* well connected.

Graph Clustering Applications

- (Search Engines) Find hubs and authoritative web pages on the web, which is treated as a graph

- (Social Network Analysis) Identify communities in a social network

- (Bioinformatics) Detect functional modules in a protein interaction network

Cuts and Clusters

- Consider a graph $G = (V, E)$ where V is the set of vertices and E is the set of edges.
- A cut $C = (S, T)$ is a partitioning of V,

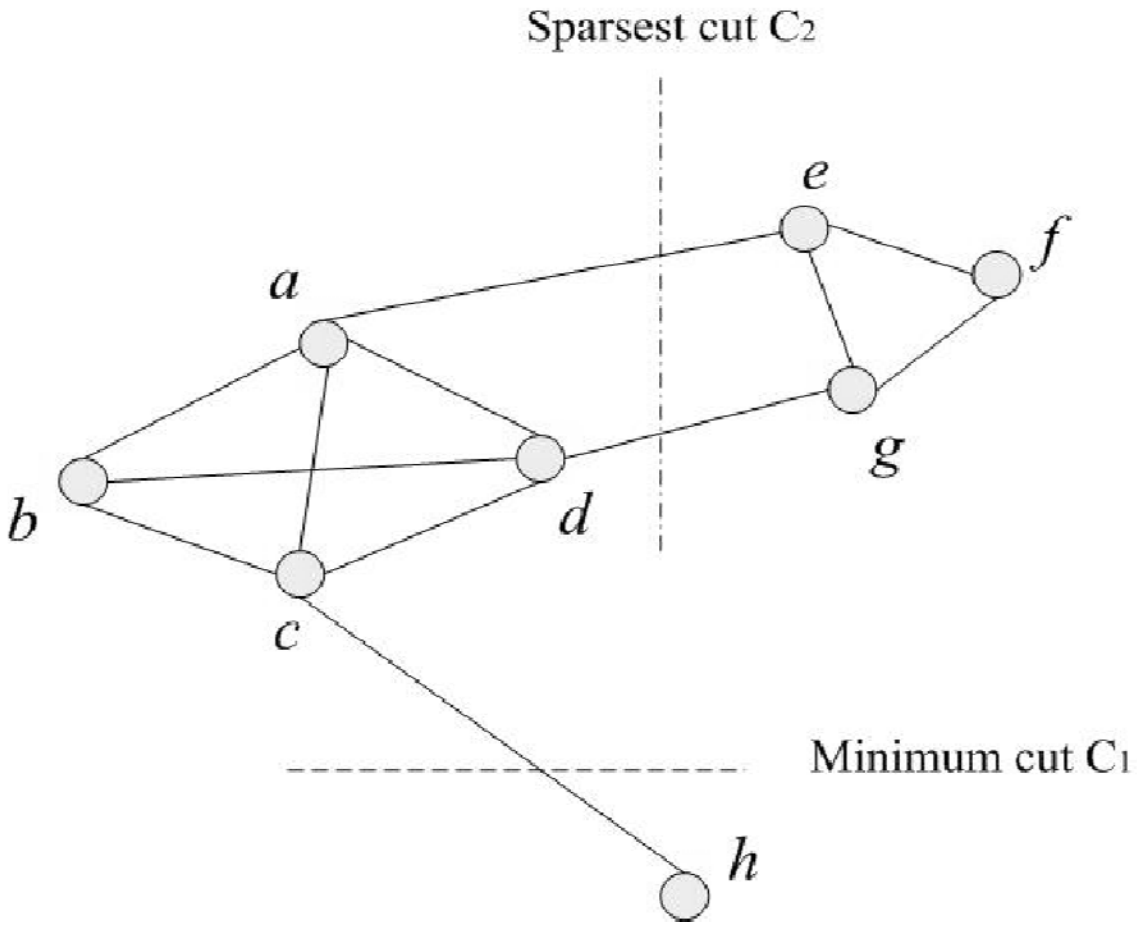
  i.e. $V = S \cup T$ and $S \cap T = \emptyset$.
- The cut set is $\{(u, v) \in E \mid u \in S, v \in T\}$.
- The number of edges in the cut set is called the size of the cut (cut size).

Minimum Cut and Sparsest Cut

Minimum cut: a cut with the smallest cut size

$$\text{sparsity} = \frac{\text{cut size}}{\min\{|S|, |T|\}}$$

Sparsest cut: a cut with the smallest sparsity


Sparsest cut C2
e
f
a
g
b
d
c
Minimum cut C1
h

Example

Graph G has two clusters: $\{a, b, c, d\}$, $\{e, f, g\}$ and an outlier vertex, h.

Cut $C_1 = (\{a, b, c, d, e, f, g\}, \{h\})$

Cut set of C_1 is $\{(c, h)\}$, cut size of C_1 is 1 and sparsity of C_1 is $1/1 = 1$. C_1 is a minimum cut.

Cut $C_2 = (\{a, b, c, d, h\}, \{e, f, g\})$

Cut set of C_2 is $\{(a, e), (d, g)\}$, cut size of C_2 is 2, and sparsity of C_2 is $2/3 = 0.67$. C_2 is a sparsest cut.

Graph Clustering Challenges

- A minimum cut does not yield a good clustering.

- A sparsest cut leads to a good clustering, but the sparsest cut problem is NP-hard.

- Need heuristics to solve the sparsest cut problem.

- Need heuristics to partition graph G into multiple clusters.

End of
Graph Clustering Module