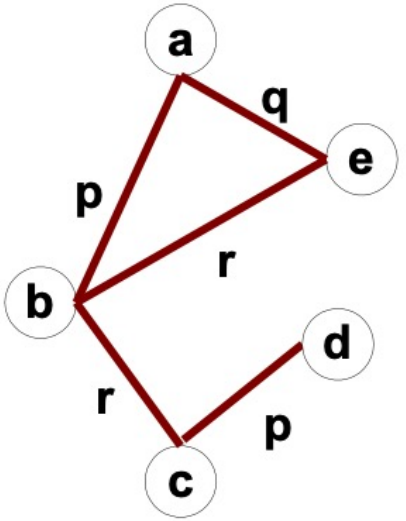
Data Mining

Graph Mining

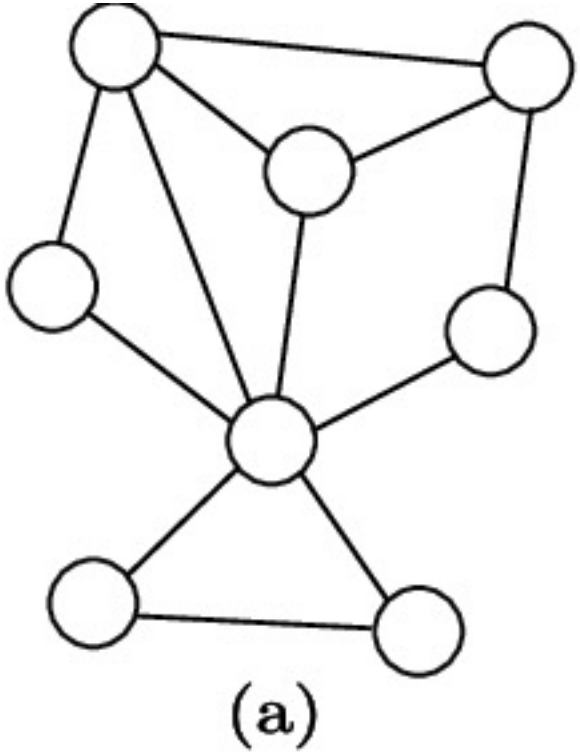
Graph Mining

- The goal of graph mining is to extract interesting subgraphs from a single large graph (e.g., a social network), or from a database of many graphs.
- A graph is a pair G=(V,E) where V is a set of vertices/nodes, and E⊆V×V is a set of edges/links
  - Here, {a, b, c, d, e} are nodes
  - {(a,b), (a,e), (b,e), (b,c), (c,d)} are edges/links
  - <p, q, r, p> are edge attributes
- If context is specified, then the word network is often used. The word graph is used for abstraction.

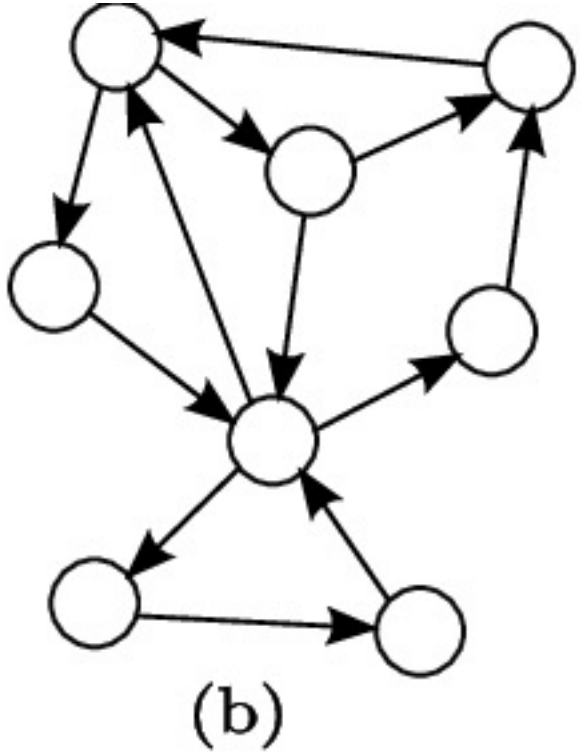


G1

Directed vs Undirected Graphs



Undirected Graph
(e.g. Facebook friendship network)



Directed Graph
(e.g., Twitter follower/following network)

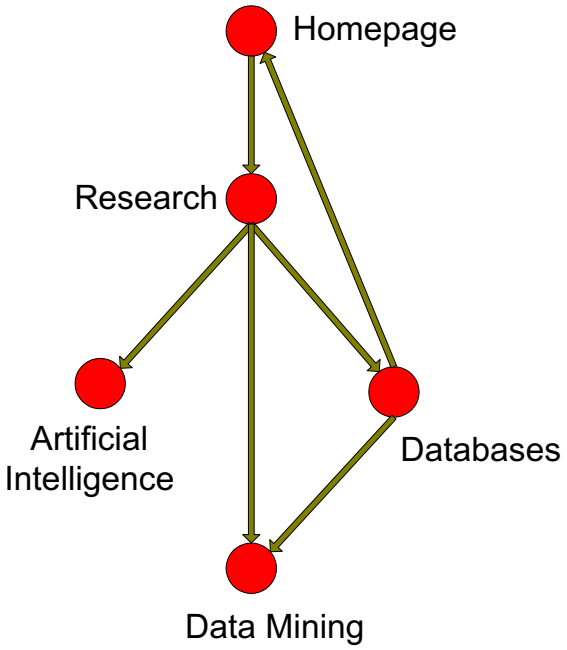
Two problems

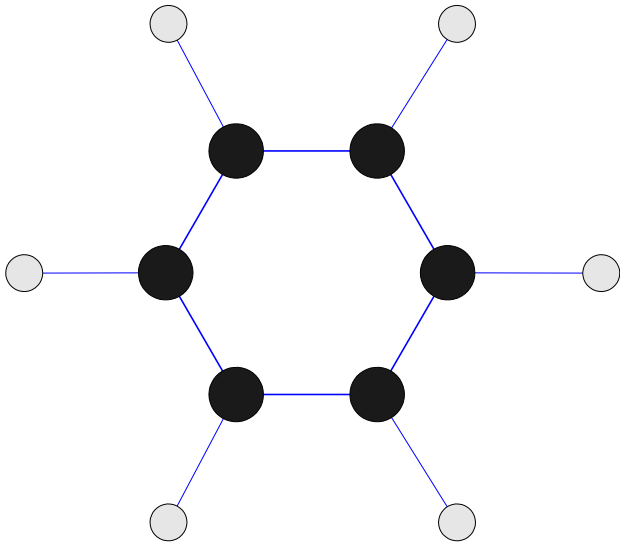
- Type 1: Given a graph, extract interesting properties of the graph
  - Degree distribution
  - Influential nodes
  - Communities

- Type 2: Given a set of graphs, identify recurrent subgraphs across the set (aka frequent subgraph mining)

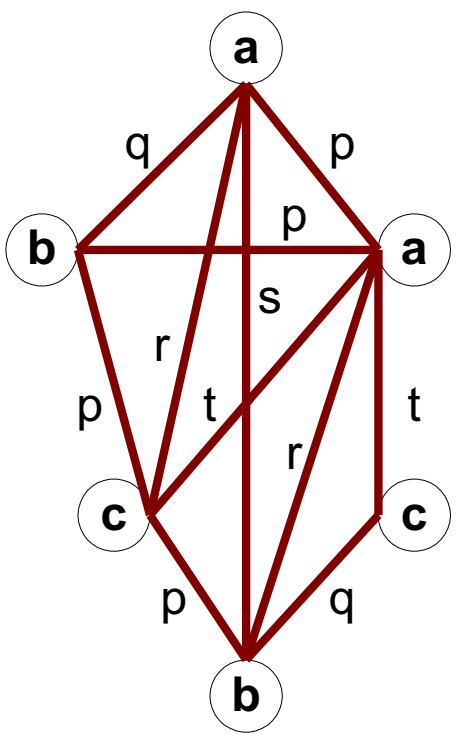
Frequent Subgraph Mining

- Extends association analysis to finding frequent subgraphs
- Useful for Web Mining, computational chemistry, bioinformatics, spatial data sets, etc

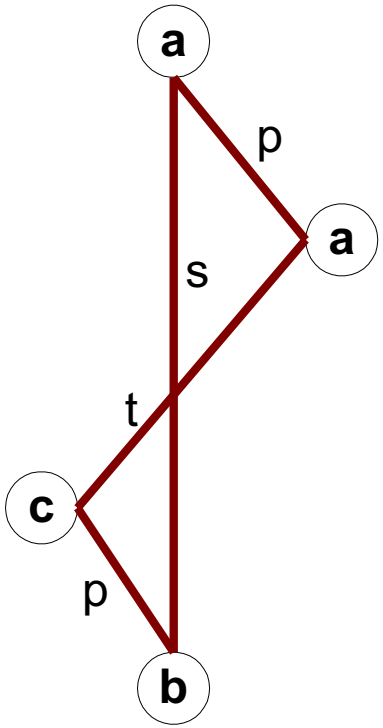




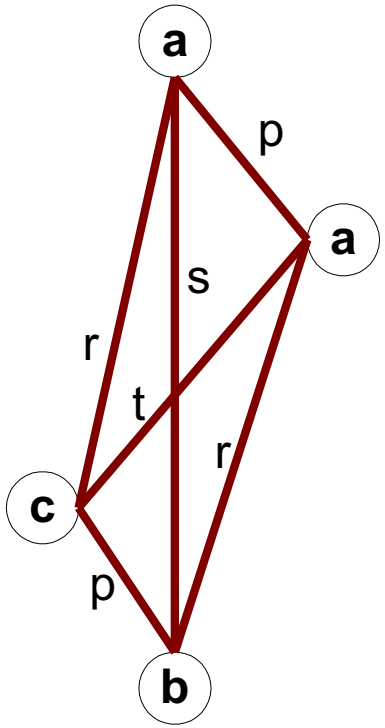
Graph Definitions



(a) Labeled Graph



(b) Subgraph

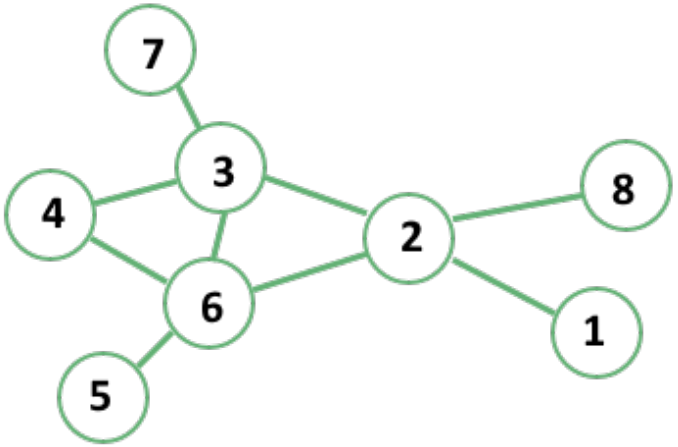


(c) Induced Subgraph

Subgraph: a set of edges and associated nodes form a subgraph
Induced subgraph: a set of nodes and their associated edges form an induced subgraph

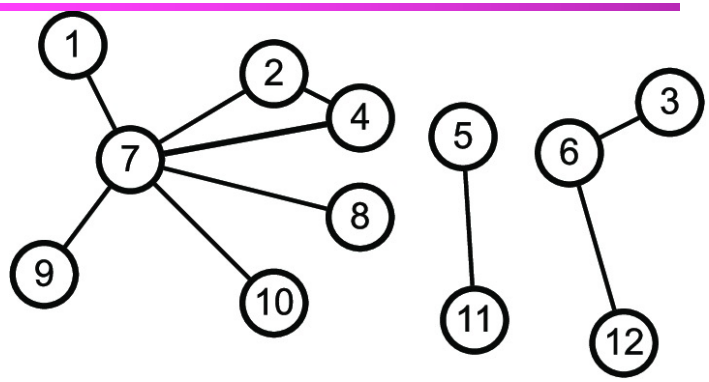
Degree

- The degree of a node is the number of edges that are linked to that node
  - Node 1 and 6 have degrees 4 and 1, respectively
- A directed network has two types of degrees
  - Number of incoming edges (in-degree)
  - Number of outgoing edges (out-degree)

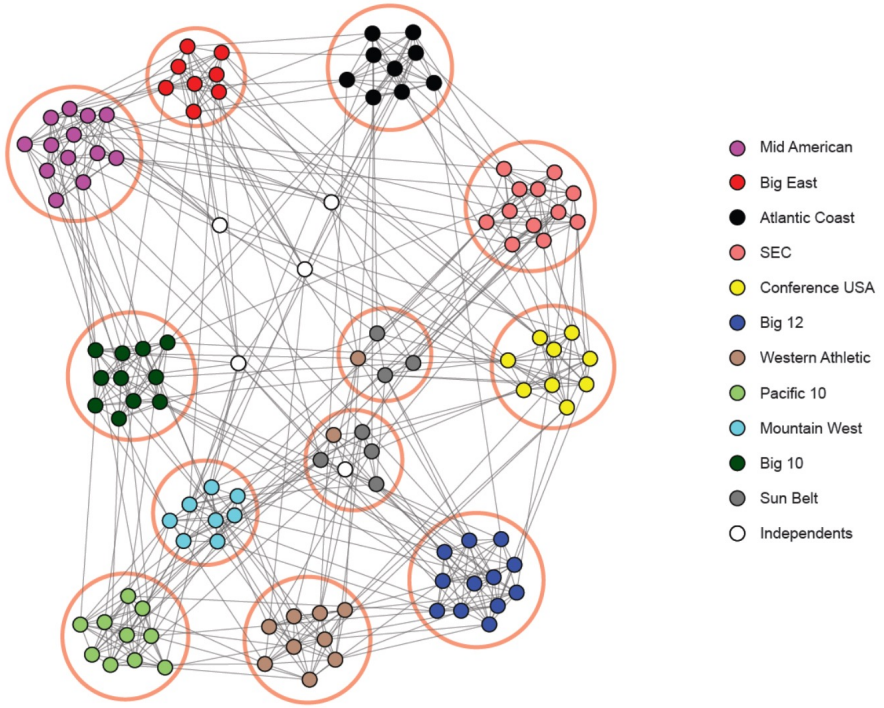


Connected Components

- A path is a sequence of edges that joins a sequence of nodes
- A connected component is a set of nodes where there is a path between each pair
  - This graph has 3 connected components
    - {1, 2, 4, 7, 8, 9, 10}
    - {5, 11}
    - {3, 6, 12}



Communities



- **A community is defined as a subset of nodes within the graph such that connections between the nodes are denser than connections with the rest of the network**
- **There are many algorithms as there are many definitions for assessing inter and intra community similarity/density measures.**
- **Essentially a graph clustering problem**

Communities in NCAA football teams network (https://snap.stanford.edu/)

Girvan–Newman Community Detection

- Betweenness centrality measure
  - For each pair of nodes, there is a shortest path
  - The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex
  - The betweenness of an edge is the number of shortest paths between pairs of nodes that run along it
- Girvan-Newman uses edge betweenness centrality measure

Girvan–Newman Community Detection

- If a network contains communities or groups that are only loosely connected by a few inter-group edges, then all shortest paths between different communities must go along one of these few edges
- The edges connecting communities will have high edge betweenness
- By removing these edges, the groups are separated from one another and so the underlying community structure of the network is revealed