Unit 8
Graph Mining and Social Network Analysis

Prepared By

Arjun Singh Saud, Asst. Prof. CDCSIT

Graph Mining

- The graph mining is the process of extracting patters (sub-graphs) of interest from graphs, that describe the underlying data and could be used further, e.g., for classification or clustering.
- Graph mining has a vast number of applications such as circuits, chemical compounds, protein structures, biological networks, social networks, Web, and XML documents.
- Graph mining is used for Fraud Detection, Community/Cluster detection, Recommending friends, Finding Influential Nodes.

Graph Mining Algorithms

- Subdue algorithm is a greedy algorithm for finding some of the most prevalent subgraphs.
- This method is not complete, i.e. it may not obtain all frequent subgraphs, although it is fast in fast execution.
- This algorithm is based on beam search algorithm, which is a heuristic search algorithm.

Graph Mining Algorithms: *Beam Search*

- Search Algorithms like BFS, DFS and A* etc. are infeasible on large search spaces.
- Beam Search was developed in an attempt to achieve the optimal(or sub-optimal) solution without consuming too much memory.
- Beam search is heuristic approach where only the most promising ß nodes (instead of all nodes) at each step of the search are retained for further branching. ß is called Beam Width.
- Beam search is an optimization of best-first search that reduces its memory requirements.

Graph Mining Algorithms: *Beam Search*

Algorithm

Open = {initial state}

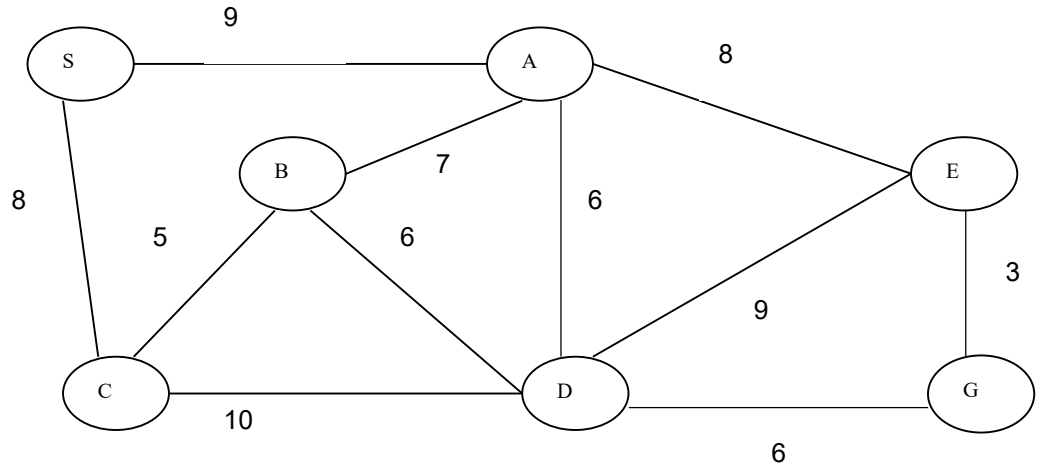
while Open is not empty do

1. Remove the best node from Open, call it n.
2. If n is the goal state, backtrace path to n and return path.
3. Create n's successors.
4. Evaluate each successor, add it to Open.
5. If |Open| > ß , take the best ß nodes and remove the others from the Open.

done

Graph Mining Algorithms: *Beam Search*

Example: Trace beam search for the following graph. Assume beam width=2



Heuristic value to G (goal node) from other nodes is given below:

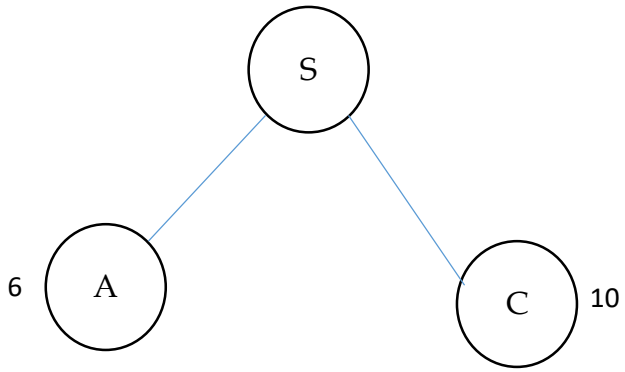
S→ G= 12 A→ G =6 B→ G = 8 C→ G = 10 D→ G = 4

E→ G = 5

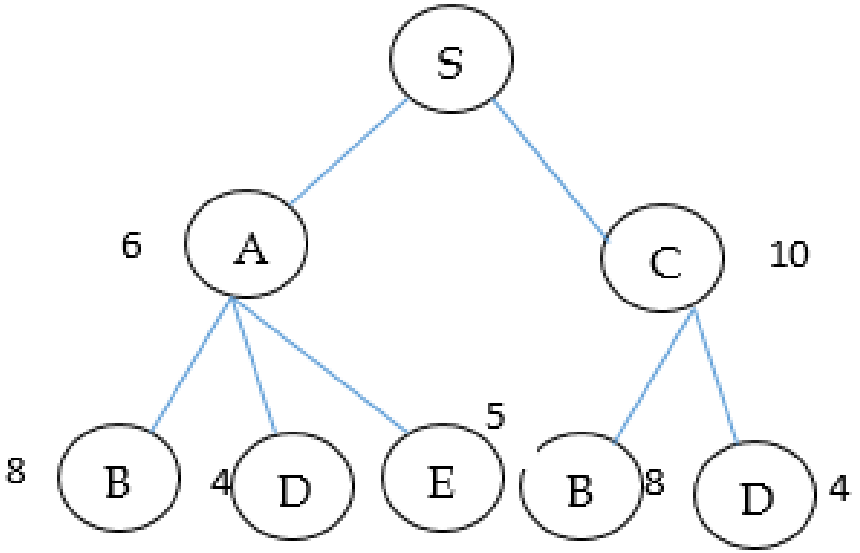
Graph Mining Algorithms: *Beam Search*

Solution

Step 1: Open={s}



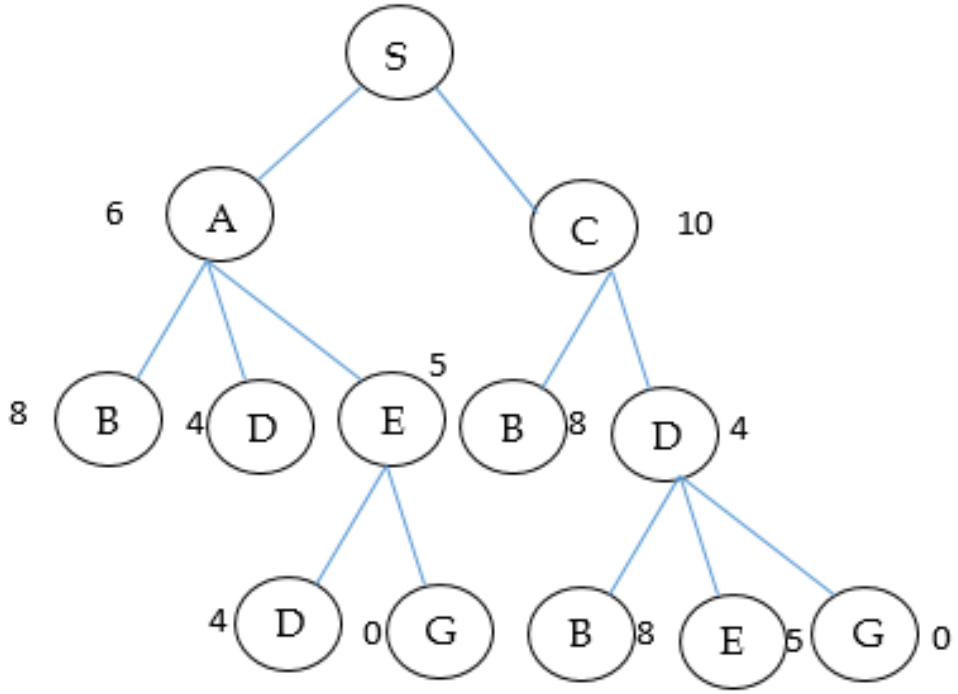
Step 2: Open={A,C}



Graph Mining Algorithms: *Beam Search*

Solution Contd..

Step 3: Open={D,E}



Step 4: Open={G,D}

Solution Found

Path Cost=9+8+3=20

Soln Path=S->A->E->G

Graph Mining Algorithms: *Inductive Logic Programming*

- Inductive Logic Programming (ILP) is a research area formed at the intersection of Machine Learning and Logic Programming.
- ILP systems develop predicate descriptions from examples and background knowledge.
- The examples, background knowledge and final descriptions are all described as logic programs.
- Induction is reasoning from the specific to the general whereas deduction is reasoning from general to specific.

Graph Mining Algorithms: *Inductive Logic Programming*

- In deduction we use rules and some facts to deduce more facts whereas in induction we derive rules from sets of facts.
- Thus we can say that goal of ILP is to find hypotheses expressed in terms of logic programming clauses from a set of positive and negative examples and domain knowledge.

Graph Mining Algorithms: *Inductive Logic Programming*

- Given the following facts:

parent(a,c)	*parent(b,c)*	*father(a,c)*
mother(b,c)	*male(a)*	*female(b)*

- Goal of ILP is to learns following rules from the given dataset.

  father(x,y)←parent(x,y) & male(x)

  mother(x,y)←parent(x,y) & female(x)

- We can represent datalog facts in graphs, then we can use ILP to identify interesting patterns in the graph.

Social Network Analysis

- Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory.
- It characterizes networked structures in terms of nodes (individual actors) and the edges or links (relationships) that connect them.
- Data Mining techniques can assist effectively in dealing with the three primary challenges with social media data.
  - First, social media data sets are large.
  - Second, Social media site's data sets can be noisy.
  - Third, data from online social media platforms are dynamic.

Social Network: Link Analysis

- Recently, link mining is becoming a very popular research area not only for data mining and web mining but also in the field of social network analysis.
- Many researches are focusing on developing new link mining techniques and algorithms, or devoting to improve traditional mining technique
- By considering links (the relationships between objects), more information is made available to the mining process. This brings about several new tasks. Such tasks with examples from various domains for social network analysis are discussed below.

Social Network: Link Analysis

- **Link-based object classification**: Link-based classification predicts the category of an object based not only on its attributes, but also on its links, and on the attributes of linked objects.
- **Object type prediction**: This predicts the type of an object, based on its attributes and its links, and on the attributes of objects linked to it. For example, we may want to know whether an actor in the social network is people or organization.

Social Network: Link Analysis

- **Link type prediction**: This predicts the type of a link, based on properties of the objects involved. We may try to predict whether two people who know each other are family members, coworkers, or acquaintances.
- **Predicting link existence**: We may want to predict whether a link exists between two objects.
- **Link cardinality estimation**: We may wish to predict number of in-links and out-links in a node.

Social Network: Link Analysis

- **Object reconciliation**: In object reconciliation, the task is to predict whether two objects are, in fact, the same, based on their attributes and links.
- **Group detection**: Group detection is a clustering task. It predicts when a set of objects belong to the same group or cluster, based on their attributes as well as their link structure.

Friends of Friend

- Person C is a friend of a friend of person A when there is a person B that is a friend of both A and C.
- Friends of friend are indirect connections in graph representation of a social networks.
- This type of analysis is done in social networks to recommend friends.

Degree Assortativity

- Assortativity, or assortative mixing is a preference for a network's nodes to attach to others that are similar in some way.
- For instance, in social networks, nodes tend to be connected with other nodes with similar degree values. This tendency is referred to as assortative mixing, or assortativity.
- This means, degree assortativity characterizes the tendency for large-degree nodes to connect to other large-degree nodes and low-degree to low-degree.

Degree Assortativity

- The assortativity coefficient is the Pearson correlation coefficient of degree between pairs of linked nodes.
- Positive values of the coefficient indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degree.
- Social networks are typically thought to be distinct from other networks in being assortative (possessing positive degree correlations).
- Well-connected individuals associate with other well-connected individuals, and poorly-connected individuals associate with each other.

Signed Network

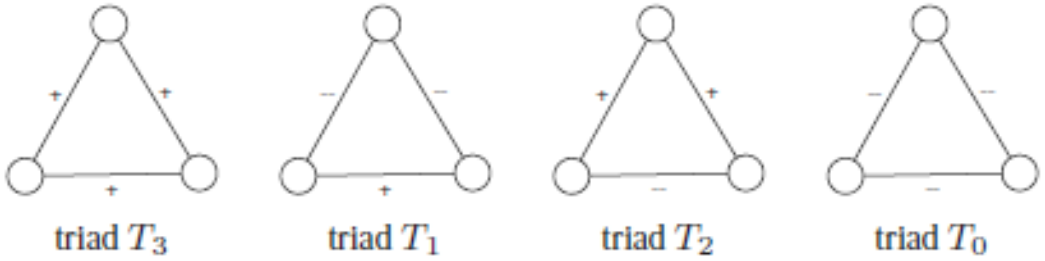
- In a social network analysis, a positive or a negative friendship can be established between two nodes in a network; this results in a signed network.
- As social interaction between people can be positive or negative, so can be links between the nodes.
- When a positive or a negative value is attributed on the relationship between the two nodes, it is called a user evaluation.
- In social groups, people can like or dislike, respect or disrespect other people in their social groups.

Signed Network: *Structural Balance Theory*

- Structural balance considers the possible ways in which triangles on three individuals can be signed
- This theory suggests that triangles with three positive signs (three mutual friends) and those with one positive sign (two friends with a common enemy) are more plausible and hence should be more prevalent in real networks. Such triangles are called balanced.
- Triangles with two positive signs (two enemies with a common friend) or none (three mutual enemies) are not plausible. Such triangles are called unbalanced.

Signed Network: *Structural Balance Theory*

- Balanced triangles with three positive edges exemplify the principle that "the friend of my friend is my friend".
- Whereas those with one positive and two negative edges capture the notions that "the friend of my enemy is my enemy".



Signed Network: *Theory of Status*

- Balance theory can be viewed as a model of likes and dislikes. However, a signed link from A to B can have more than one possible interpretation, depending on A's intention in creating the link.
- In particular, a positive link from A may mean, "B is my friend," but it also may mean, "I think B has higher status than I do."
- Similarly, a negative link from A to B may mean "B is my enemy" or "I think B has lower status than I do."

Signed Network: *Theory of Status*

- In this theory of status, we consider a positive directed link to indicate that the creator of the link views the recipient as having higher status; and a negative directed link indicates that the recipient is viewed as having lower status.
- These relative levels of status can then be propagated along multi-step paths of signed links, often leading to different predictions than balance theory.

Signed Network: *Conflict between the theory of balance and status*

- To give a sense for how the differences between status and balance arise, consider the situation in which a user A links positively to a user B, and B in turn links positively to a user C.
- If C then forms a link to A, what sign should we expect this link to have? Balance theory predicts that since C is a friend of A's friend, we should see a positive link from C to A.
- Status theory, on the other hand, predicts that A regards B as having higher status, and B regards C as having higher status — so C should regard A as having low status and hence be inclined to link negatively to A.
- In other words, the two theories suggest opposite conclusions in this case.

Trust in a Network

- Web of trust is used in network to express or predict trust/distrust between users.
- The webs of trust tend to be relatively sparse: every user has expressed trust values for only a handful of other users.
- The problem is to determine trust values for the remaining user pairs using only those which are explicitly specified.
- There are various ways to infer trust in networks: *Atomic propagation, Propagation of Distrust, and Iterative Propagation*.

Trust in a Network: *Atomic Propagation*

- The atomic propagations are a "basis set" of techniques by which the system may infer that one user should trust or distrust another.
- The set is constructed so that any inference regarding trust should be expressible as a combination of elements of this set. The basis set is as follows:
  - ***Direct propagation***: If i trusts j and j trusts k, then we could conclude that i trusts k.
  - ***Co-citation***: If i1 trusts j1 and j2, and i2 trusts j2, then we conclude that i2 should also trust j1.

Trust in a Network: *Atomic Propagation*

- ***Transpose trust:*** Here i's trust of j causes j to develop some level of trust towards i.
- ***Trust coupling:*** When both i and j trust k, this implies that i trusts j.

Trust in a Network: *Propagation of Distrust*

- The following are three models for the propagation of trust and distrust, given initial trust and distrust matrices T and D respectively:
- **Trust only:** By completely disregarding distrust scores and propagating only the trust scores.
- **One-step distrust:** When a user distrusts someone, they discount all their judgments. Therefore, distrust propagates only one step.
- **Propagated distrust:** When trust and distrust both propagate together, we get: $B = T - D$. Here, B is set of trust and distrust.

Trust in a Network: *Iterative Propagation*

- The end goal is to produce a final matrix F that has the trust or distrust between any pair of users in this universe.
- There are two approaches to computing F from the sequence of propagations: *Eigenvalue Propagation and Weighted Linear Combination*.