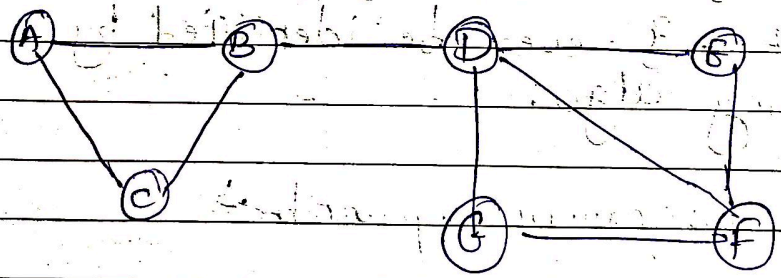
Camlin Page
Date / /

DS Unit 5

Q.1 - A social networking service is an online platform which people use to build social networks or social relations with other people who share similar personal or career interests, activities, backgrounds or real-life connections.

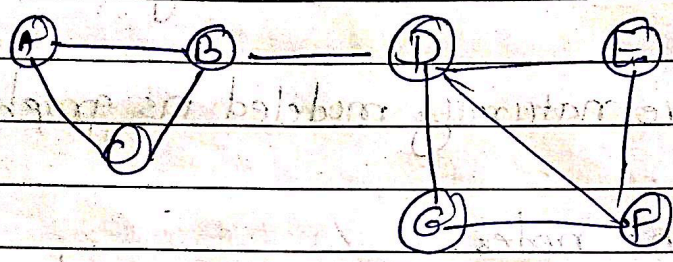
- SNA is the process of investigating social structures through the use of networks and graph theory.
- Social networks are naturally modeled as graphs; called as social graphs.
- Entities are the nodes.
- An edge connects two nodes if the nodes are related by the relationship that characterizes the network.
- Degree is represented by labelling the edges.
- Often, social graphs are undirected.



Q.2 - There are two general approaches to clustering:
(i) hierarchical
(ii) Point-assignment

(1) Hierarchical clustering
- Hierarchical clustering of a social-network graph starts by combining some two nodes that are connected by an edge.

- Successively, edges that are not between two nodes of the same cluster would be chosen randomly to combine the cluster the cluster to which their two nodes belong.
- The choices would be random, because all distances represented by an edge are the same



- At highest level it appears that there are two communities {A, B, C} and {D, E, F, G}.
- {D, E, F} and {D, F, G} subcommunities of {D, E, F, G} - ever be identified by a pure clustering algo.

(2) Point-assignment approach :

- Suppose we try a k-means approach, pick k=2.
- Pick two starting nodes at random.
- Start with one randomly chosen node and then pick another as far away as possible.
- Suppose we get two starting nodes, B and F.
  - Assign A & C to the cluster of B.
  - E and G to the cluster of F.
  - But D is as close to B as it is to F.
- Deferred decision about D, until all are assigned.

- shortest average distance to all the nodes of the cluster, then D should be assigned to cluster of F.

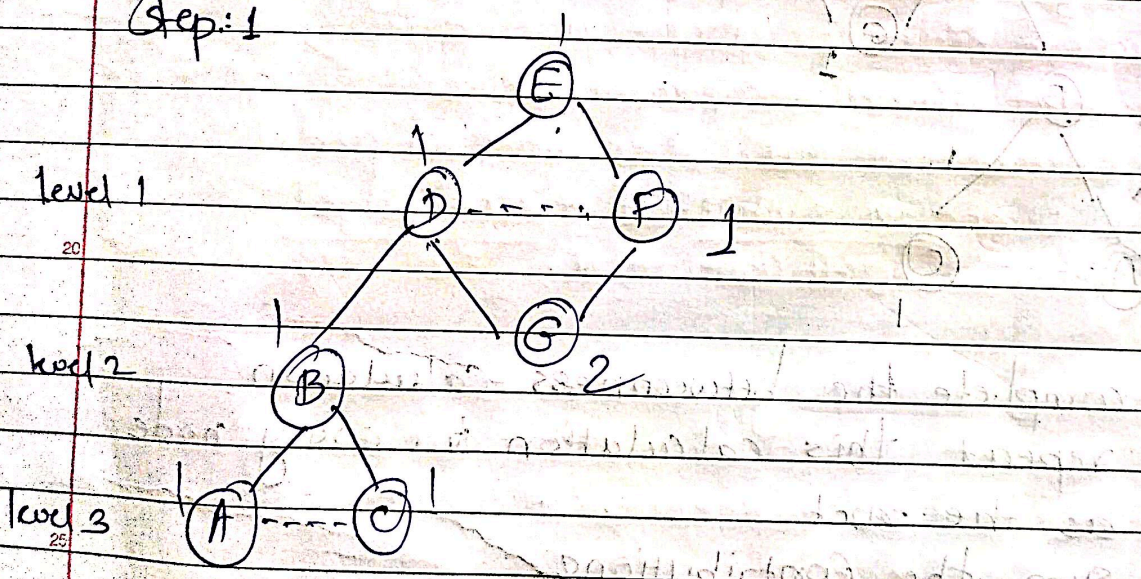
Q-3 Betweenness

Betweenness of an edge (a, b) to be the no. of pairs of nodes x and y such that the edge (a, b) lies on the shortest path between x and y.

GN algorithm:-

- GN algorithm visits each node X once and computes the number of shortest paths from X.

Step: 1

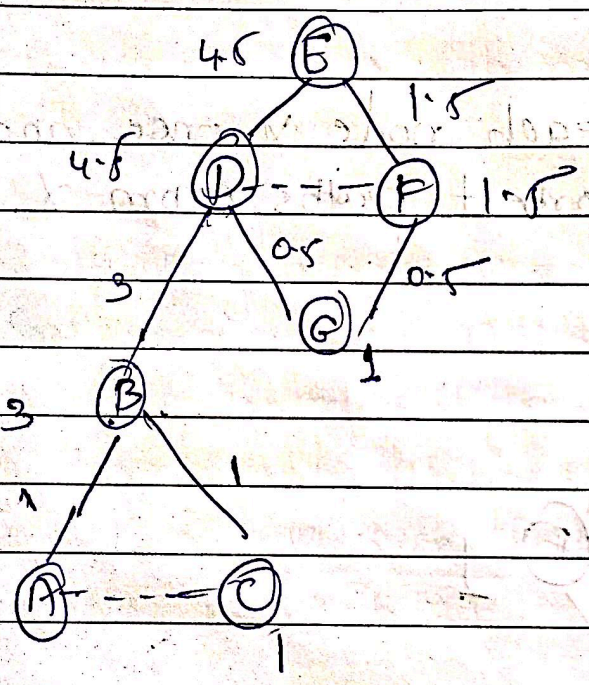


Step: 2 → label each node by the no. of SPs that reach it from root.

- Calculate for each edge e sum over all nodes Y of the fraction of sps from root X to Y that go through e.

- Calculation involves computing their sum for both nodes and edges, from the bottom.

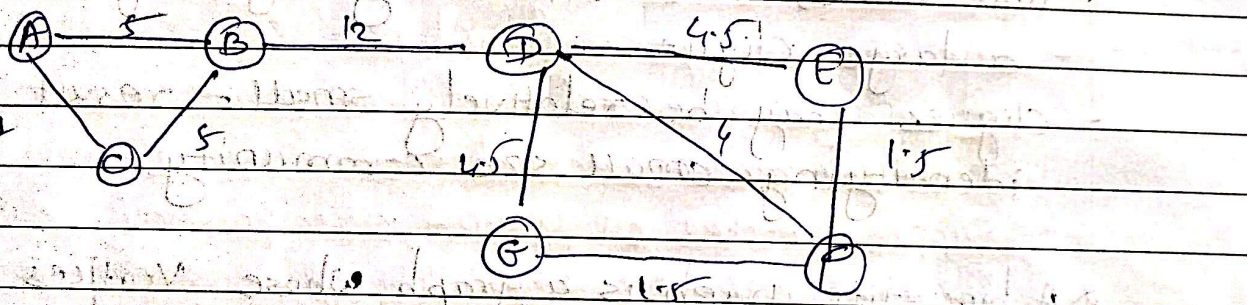
- Each Nodes and Edges are given credit.



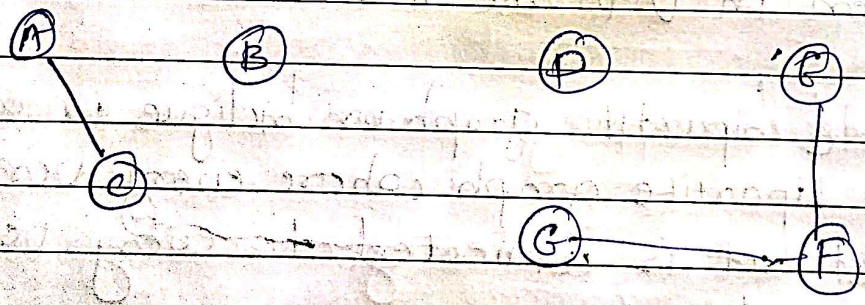
- To complete the betweenness calculation
  - repeat this calculation for every node as the root.
  - Sum the contributions.
  - Finally, divide by 2 to get true betweenness.

Q-4 Betweenness to find communities.

- Betweenness scores for the edges of a graph behave something like a distance measure,
- Not exactly a distance measure, because it is not defined for pairs of nodes that are unconnected by an edge; and might not satisfy triangle inequality.
- Idea is expressed as a process of edge removal
- Removal edges with the highest betweenness.



Remove edges with betweenness four or more



All the edges with betweenness 4 or more have been removed.

- B is a "traitor" to the community {D, B, C}
- D can be seen as a "traitor" to the group {D, E, F, G}

Q.5. Discover Communities in social-Network Graph directly.

Clique :

In the mathematical area of graph theory, a clique is a subset of vertices of an undirected graph such that every two distinct vertices in the clique are adjacent, that is its induced subgraph is complete.

- Find sets of nodes with many edges by finding a large clique.
- Cliques may be relatively small - result in identifying small size community.

* A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V.

* A comple bipartite graph or biclique is a special kind of bipartite graph where every vertex of first set is connected to every vertex of second set.

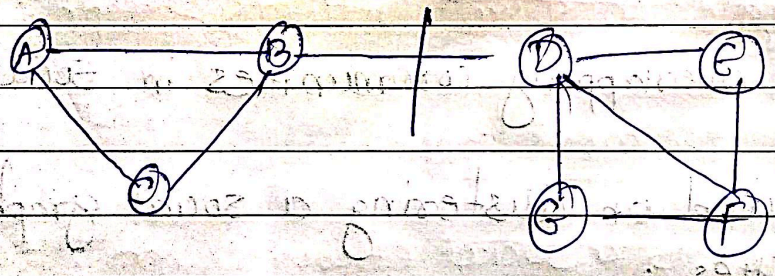
- We can use Complete bipartite subgraph for community in ordinary graphs.

- In Given graph G, find instances of $K_{s,t}$.
- For instance $K_{s,t}$, assume $t \leq s$.
- Baskets corresponds to nodes on right side of G.
- Each basket corresponds to a node that is connected to all t of nodes in F.
- Thus, frequent itemset of size t and s of baskets in which all those items appears form an instance of $K_{s,t}$.

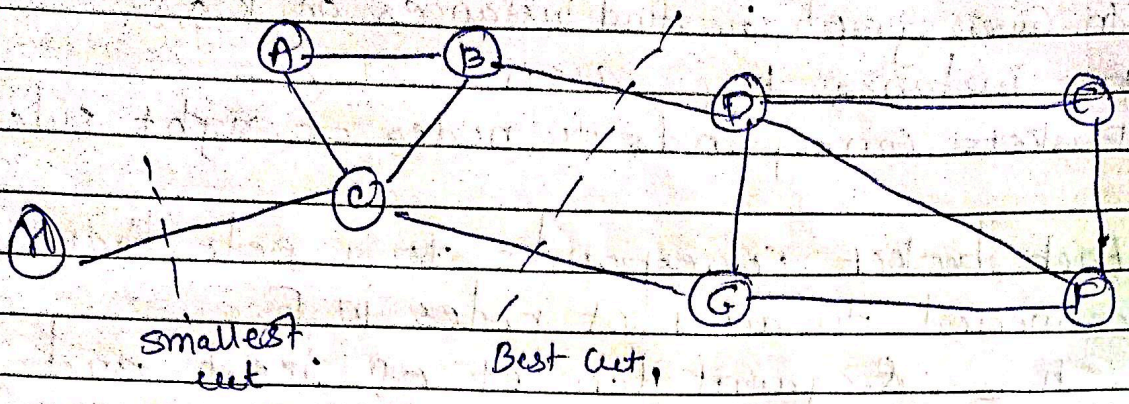
Q.6 How social Network graph can be partitioned to identify Communities?

→

* To partitioning a graph, we will use some important tools from matrix theory.
* Minimize number of edges that connect different components.
* Goal of minimizing the 'cut' size needs to be understood carefully.



* Divide nodes into two sets so that Cut, or set of edges that connect nodes in different sets is minimized.
* Two sets are approximately equal in size.



Normalized cuts:

- Good cut must balance the size of the cut against size of the sets.
- One choice - "normalized cut"
- Volume of a set (Vol(S)) - to be number of edges with at least one end in S.
- Suppose nodes of graph are partitioned into two disjoint sets S & T.

* Normalised cut value for S and T:

$$\frac{cut(S,T)}{Vol(S)} + \frac{cut(S,T)}{Vol(T)}$$

7 How to find overlapping communities in Social Network Graph

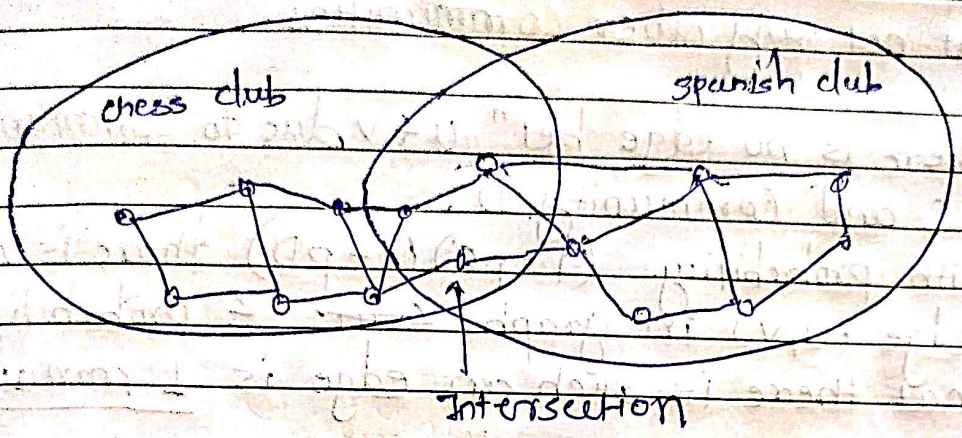
⇒

* Concentrated on clustering a social graph to find communities.
* In practice communities are rarely disjoint.
* We explain a method for taking a social graph and fitting model to it that best explains how it could have been generated by a mechanism.
* This assumes that probability that two individuals are connected by an edge increases as they become members of more communities in common.

* An important tool in this analysis is "maximum-likelihood estimation" (MLE).

Nature of Communities:

- Nodes are people & there is an edge betⁿ two nodes if the people are "friends".



(overlapping of communities is denser than non-overlapping)

Q.8 Affiliation-Graph model to find overlapping communities:

- In given number of communities, each community can have any set of individuals as members. That is, it is a parameters.
- Each community C has probability p^C associated with it, probability that two members of community C are connected by an edge because they are both members of C.
- If a pair of nodes is in two or more communities, then there is an edge between them.

- The key observation is how the edge probabilities are computed, given an assignment of individual to communities and values of the p_C's.

- consider an edge (u,v) betⁿ nodes u and v.

- suppose u and v are members of communities C & D, but not any other communities.

- There is no edge betⁿ u & v due to community C and community D.

- With probability $(1-p_C)(1-p_D)$ there is no edge (u,v) in graph ~~Color~~ & probability that there is such an edge is 1 minus that.

p_{uv} (probability of an edge betⁿ u and v :-

$$p_{uv} = 1 - \prod_{C \text{ in } M} (1-p_C)$$

Q Why triangle in social-Network graph are counted ? Explain algorithm for finding triangles in Social Network Graph.

→ - If a graph is a social network with n participants and m pairs of "friends" then number of △les to be much greater than value of random graph

- The reason is that if A and B are friends, and A is also a friend with C, there should be a much greater chance than average B & C are also friends.

- Counting the no. of triangles helps us to measure extend to which a graph looks like a social network.

- The age of a community is related to the density of triangles.

* An algorithm for finding triangles:

- Consider a graph of n nodes and $m \geq n$ edges and nodes are integers $1, 2, \ldots, n$.
- Call a node a heavy hitter if its degree is at least $\sqrt{m}$.
- Note that the number of heavy-hitter nodes is no more than $2\sqrt{m}$.
- since each edge contributes to degree of only two nodes, there would then have to be more than m edges.

- We shall order nodes as follows:
  * First order nodes by degree.
  * If v and u have the same degree, recall that both v and u are integers, so order them numerically.

* That is, we say $v < u$ if and only if:
  ① The degree of v is less than degree of u or
  ② The degree of u and v are the same and $v < u$.