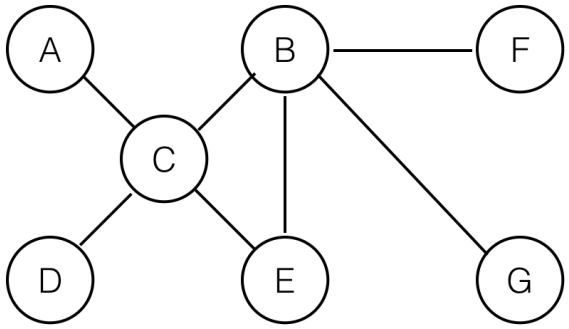
Question 1: Network Measures and Models

It is important to study the measures of a network, such as centrality of nodes in a network and model the structure of a network using a good abstract model.

1. Given the following undirected graph:



   (1) The degree distribution of the graph? **Solution**: (0, 4/7, 1/7, 0, 2/7)
   (2) The radius and diameter of the graph? **Solution**: r = 2, d = 3
   (3) The clustering coefficient of node B? **Solution**: 1/6
   (4) The degree centrality, eccentricity centrality, closeness centrality of node C? **Solution**: 4, 1/2, 1/8

2. What are the key differences between PageRank and HITS algorithms?

   Solution: HITS considers authorities (in-direction) and hubs (out-direction) separately, while PageRank does not.

3. Researchers have been modeling social and/or information networks using several models. What are the similarities and differences among the following three models: (1) Edös-Rényi random graph model, (2) Watts-Strogatz small world model, and (3) Barabasi-Albert scale-free network model?

   Solution: All the three models give few components and small diameter. ER and WS are not inhomogeneous in degree. Also, ER and WS are not heavy- tailed in link number distribution and hence not scale free. WS has high clustering, while ER does not. BA gives heavy-tailed distribution but not high clustering.

Question 2: Clustering and Ranking in Heterogeneous Information Networks

1. RankClus clusters heterogeneous information networks by integrating ranking and clustering in the clustering process. (1) What is the intuition (motivation) of RankClus? (2) Why is RankClus more efficient than SimRank?

   Solution: (1) Better clustering: Rank distributions for clusters are more distinguishing from each other; Better ranking: Better metric for objects is learned from the ranking. (2) For each iteration, SimRank is quadratic to the number of nodes, i.e., linear to the number of node pairs, while RankClus is linear to the

number of links. Since links are usually sparse, the number of links should be much less than the number of nodes.

2. When clustering a heterogeneous information network, different meta-paths carry different semantic meanings and thus lead to different clustering results. (1) Describe three different meta-paths in the DBLP network, and explain the semantic meanings of them. (2) PathSim is one of the popular similarity measures used in clustering a heterogeneous information network. The following table shows an adjacency matrix between authors and venues in the DBLP network, denoting the number of papers published by each author in each venue. Apply PathSim to find the which author is more similar to Mike. Please show the calculation of PlathSim below.

Author\Conf.	SIGMOD	VLDB	ICDM	KDD
Mike	4	2	0	0
Jim	50	20	0	0
Bob	3	2	0	1

Solution: s(Mike, Jim) = $\frac{2\times(4\times50+2\times20)}{(4\times4+2\times2)+(50\times50+20\times20)}$ = 0.1644

s(Mike, Bob) = $\frac{2\times(4\times3+2\times2+0\times1)}{(4\times4+2\times2)+(3\times3+2\times2+1\times1)}$ = 0.9412

Bob is more similar to Jim.

Question 3: Classification and Prediction of Heterogeneous Information Networks

1. RankClass is a ranking-based classification algorithm in heterogeneous information networks. What are the relations and differences between RankClus and RankClass?

Solution: Both use ranking information as features, and let clustering/ranking and classification mutually enhance each other. The difference is RankClass has class labels while RankClus doesn't.

2. One would like to recommend a few papers in the literature that a new CS research paper should cite, given the paper contents (i.e., a set of critical terms), authors and target venues. Suppose we have a bibliographic database such as DBLP plus the citation information and contents of each published paper in the database. Outline and describe a mechanism to do this and reason why your mechanism may yield quality results.

Solution: All reasonable answers are acceptable. One possible mechanism is ClusCite.

Question 4 (Programming Required): Similarity Measure and Classification in Heterogeneous Information Network

This task is to implement one similarity measure (PathSim) and one classification algorithm (GNetMine) in heterogeneous information network. The data input is a heterogeneous information network of academic publications, with 4 types: author, conference, paper and term. The dataset we use contains

14376 papers, 20 conferences, 14475 authors and 8920 terms. Within the dataset, 4057 authors, 100 papers and all 20 conferences are manually labeled to four classes, representing four different research areas: database, data mining, information retrieval and artificial intelligence.

Sub-Task 1. The goal is to implement one similarity measures: PathSim, and study semantic meanings of different meta paths.

Solution: The author himself should be ranked top 1 in the ranked list.

Sub-Task 2. The goal is to implement the classification algorithm: GNetMine, and evaluate accuracies of three types of nodes: author, paper and conference. GNetMine is a graph-based regularization algorithm for classification on heterogeneous information networks. It can be viewed as a process of information propagation, where the label information is propagated from labeled objects to unlabled ones through links until a stationary state is achieved. More details of GNetMine could be found in hw1_tips.pdf.

Solution: Full credits will be given if accuracies are within certain ranges.