Question 1: Network Measures and Models

1. (a) (A, B, C, D, E, F, G) = (1, 4, 4, 1, 2, 1, 1)
   → Degree sequence: (4, 4, 2, 1, 1, 1, 1)
   → Degree frequency distribution: (0, 4, 1, 0, 2)
   → Degree distribution: (0, 4/7, 1/7, 0, 2/7)

   (b) Eccentricity: A:3 B:2 C:2 D:3 E:2 F:3 G:3
   →Radius: r(G) = 2
   →Diameter: d(G) = 3

   (c) $G_B = (4,1)$
   Clustering coefficient of node B: $C(V_B) = 2* 1 / (4*(4-1)) = 1/6$

   (d) Degree centrality: 4
   Eccentricity centrality: 1/2
   Closeness centrality: 1/ 1+1+1+1+2+2 = 1/8

2. PageRank computes one value but HITS computes two values for a page (Hub and Authority)
   The matrix of HITS is a multiplication of the adjacency matrix and its transpose.
   HITS is query-based.

3. Similarities: They all give few components and small diameter.
   Differences:
   (1) Edös-Rényi random graph model:
   It does not give high clustering and heavy-tailed degree distributions
   It is the mathematically most well-studied and understood model
   (2) Watts-Strogatz small world model:
   It gives high clustering
   It does not give heavy-tailed degree distributions
   (3) Barabasi-Albert scale-free network model:
   It gives heavy-tailed distribution
   It does not give high clustering

Question 2: Clustering and Ranking in Heterogeneous Information Networks

1. (1) Clustering and ranking are mutually enhanced: Rank distributions for clusters are more distinguishing from each other; Better metric for objects is learned from the ranking
   (2) SimRank will be at least quadratic at each iteration since it evaluates distance between every pair in the network. As for RankClus, Clustering and ranking are mutually enhanced.

2. (1) Author – Paper -> Paper – Author: Author cites another author's paper
   Author – Paper – Venue – Paper – Author: Two author publish paper in same venue
   Author – Paper – Topic – Paper – Author: Two Authors write paper in same topic
   (2) We use A-P-C-A-P meta-path.
   s(Mike, Jim) = 2*(4*50+2*20) / ((4*4+2*2) + (50*50 + 20*20)) = 0.164
   s(Mike, Bob) = 2*(4*3+2*2) / ((4*4+2*2)+(3*3+2*2+1*1)) = 0.941

Question3: Classification and Prediction of Heterogeneous Information Networks

1.(1)

Relations: They both are enhanced. RankClus: Clustering and ranking are mutually enhanced; RankClass: Highly ranked objects will play more role in classification. They both use rank to improve themselves.

Differences:

After all, one is classification and another one is clustering. RankClus can have no training, no available class labels, no expert knowledge. But RankClass need more information.

(2) ClusCite

First: Organize given information (paper contents, authors and target venues) into different groups, each having its own behavioral pattern to identify references of interest.

Second: Derive group membership groups for our new CS research paper; For different interest groups, learn distinct models on finding relevant papers and judging authority of papers.

Third: Integrate learned models of its related interest Derive group membership groups

Reason: Since different meta-paths may have different importance, we handle different groups (different types of meta-paths) with different model. Comparing with most recommendation methods use one model, the mechanism is heterogeneous network modeling. We can capture paper-paper relevance of different semantics and enable authority propagation between different types of objects.

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

(1) Top-5 ranked results (i.e., similar researchers) for author id 7696 (except itslef):
APVPA: [7479 3227 4780 1760 3230]
APTPA: [3230 392 1759 4780 1760]

(2) Accuracy of author: 0.92874937718
Accuracy of paper: 0.859649122807
Accuracy of conference: 1.0