CS512 (Spring 2013) "Advanced Data Mining": Midterm Exam I

(Tuesday, Feb. 26, 90 minutes, 100 marks **brief answers** directly written on the exam paper)
Note: Closed book and notes but one reference sheet allowed, basic calculator permitted but other electronic devices are not allowed, scratch paper not need to be returned. The last question is opinion collection: whatever answer will receive three bonus points.

Name: NetID: Score:

1. [30] Introduction to Networks

   (a) [10] What are the differences between *eigenvector centrality* and *Katz centrality*? Explain why they cannot be directly used to discover Web authoritative pages?

(b) [10] What are the differences between PageRank and HITS algorithms? Why do people say that HITS explores both *co-citation* and *bibliographic coupling*?

(c) [10] Why is it that WWW cannot be modeled by the Erdös-Réyni model, nor by the Watts-Strogatz model?

2. [37] Mining heterogeneous information networks: Part I

   (a) [9] Explain why integration of classification and ranking may improve the quality of classification in heterogeneous information networks.

(b) [9] Using one simple example, show why the PathSim measure may help find peer objects whereas random-walk and pair-wise random walk may not.

(c) [9] Explain how user guidance may help find appropriate meta-paths for rank-based clustering in heterogeneous information networks.

(d) [10] The Computer Science bibliographic database DBLP contains information about authors, papers, publication venues, paper titles, and paper publication year (assuming there is no paper citation information), forming a heterogeneous information network. Outline a method that may predict what topics (a set of terms) that an author may work on in coming years.

3. [30] Mining heterogeneous information networks: Part II

   (a) [10] Taking two constraints as examples in the analysis of advisor-advisee relationship, explain why constraints are critical at roles discovery in heterogeneous information networks.

(b) [10] Algorithm DISTINCT (that distinguishes authors with identical names) was developed when there were no concepts like Meta-Path and RankClus introduced. Explain how you will enhance the DISTINCT method using those new concepts introduced in class.

(c) [10] Explain why LTM (Latent Truth Model) is more powerful than TruthFinder on truth finding.

4. [3] (Opinion). [3 bonus points]

   (a) I ☐ like ☐ dislike the exams in this style.

   (b) In general, the exam questions are ☐ too hard ☐ too easy ☐ just right.

   (c) I ☐ have plenty of time ☐ have just enough time ☐ do not have enough time to finish the exam questions.