T-61.5060: Algorithmic Methods of Data Mining

Lecturer: Juho Rousu

Course Exam

December 18, 2012

Instructions:

You have **three (3)** hours to complete this exam. You are allowed to use one **two-sided cheat-sheet** (A4 page, both sides hand-written), which you have to submit together with the exam paper. No additional material can be used. The total score that can be obtained is **50 points**. As described in the course requirements, you need to score **at least 25/50 points** to pass this exam.

Question 1 (Concept definitions) **[10 points]**

Define the following concepts (a few sentences each):

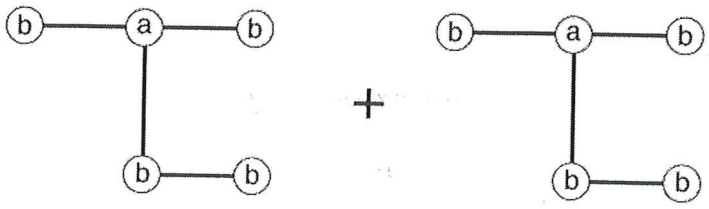
a) Association rule
b) FP-Tree
c) Graph automorphism
d) Co-location pattern
e) Maximal frequent itemset
f) Kendall's distance
g) RankSVM
h) Fagin's algorithm
i) PageRank
j) Independent cascade model

Question 2 (Frequent itemsets and association rules) **[10 points]**

Describe in detail the Apriori algorithm for frequent itemset and association rule mining.

Question 3 (Graph mining) **[10 points]**

a) Describe the edge-growing approach for candidate generation in graph mining **[5 points]**
b) Determine the set of candidate subgraphs generated by the merger of the below two graphs using edge-growing **[5 points]**



[turn the page]

Question 4 (Object ranking) **[10 points]**

a) Describe Cohen's greedy ordering algorithm for object ranking
b) Simulate Cohen's greedy ordering algorithm on the following set of preference scores (assume PREF(x,y) = 1-PREF(y,x) holds):
PREF(A,B) = 0.1, PREF(A,C) = 0.5, PREF(A,D) = 0.9, PREF(B,C) = 0.6, PREF(B,D) = 0.7, PREF(C,D) = 0.5

Question 5 (Rank aggregation) **[10 points]**

Consider rank aggregation using Threshold algorithm with the two aggregation functions given below. Give a pseudocode for the algorithm in both cases. Discuss the correctness of the algorithm (why your algorithm gives the correct top-k) and its optimality with respect to number of sequential and random accesses to items in the ranked lists.

a) Maximum score: $f(u_1,\ldots,u_m) = \max(u_{1,\ldots},u_m)$ **[5 points]**
b) Minimum score: $f(u_1,\ldots,u_m) = \min(u_{1,\ldots},u_m)$ **[5 points]**