# HW3 - Data Mining - due by 10 am 12 March 2019

**Homework Policy:**

1. You can discuss your HW with at most one other group. However, discussions are restricted to oral only. Written similarity is regarded as plagiarism. If you group discusses with other groups, you must acknowledge the discussion in your written solution.
2. Homework submission: one person in the group submitting the homework is ok, but all the names must appear on the paper of the written homework.
3. The written homework must be submitted by the due date. Please turn in a hard copy in class. The hard copy will be graded.
4. This list may be updated further.

**Problem 1 (5 points)** (Exercise 5.1.2 MMDS book ) Compute the PageRank of each

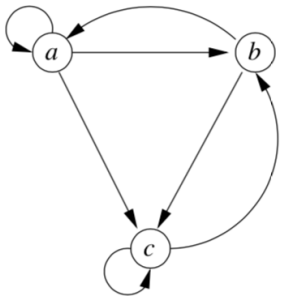


Figure 1: The graph for Problem 1.

page in Fig 1, assuming $\beta = 0.8$.

**Problem 2 (5 points)** (Exercise 5.1.5 MMDS book ) Show by induction on $n$ that if the second, third, and fourth components of a vector $v$ are equal, and $M$ is the transition matrix of the graph in Figure 2, then the second, third, and fourth components are also equal in $M^n v$ for any $n \geq 0$.

**Problem 3 (10 points)** (Exercise 5.4.2) For the graph in Figure 2, assuming only $B$ is a trusted page. Compute the TrustRank and the spam mass of each page.

**Problem 4 (5 points)** (Exercise 5.5.1 MMDS book ) Compute the hubbiness and authority of each of the nodes in the graph in Figure 2.

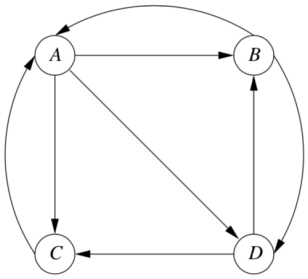


Figure 2: The graph for Problem 2 , 3 and 4.

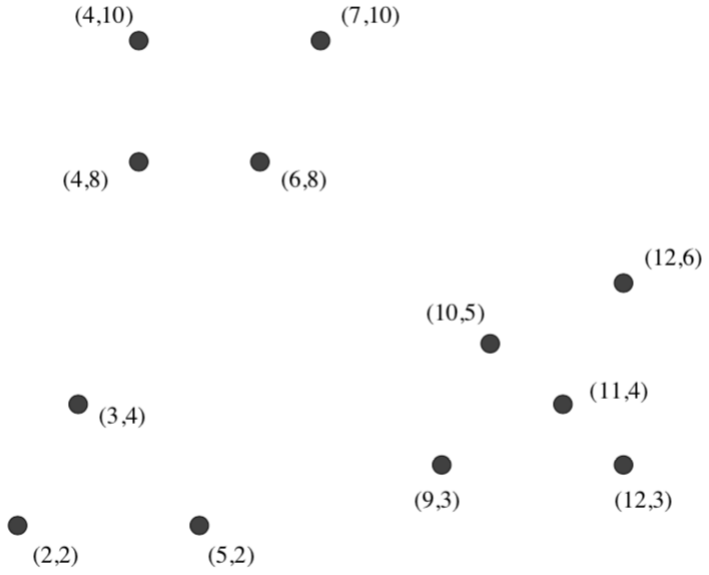


Figure 3: An example point set for Problem 5 and 6.

**Problem 5 (5 points)** (Exercise 7.2.3 MMDS Book): Construct the cluster tree of the example point set in Figure 3 if we choose to merge the two clusters whose resulting cluster has the smallest radius.

**Problem 6 (10 points)** For the points of Figure 3, if we select three starting points using the method presented in class (Section 7.3.2 in the book), and the first point we choose is $(3, 4)$, which other points are selected.

**Problem 7 (10 points)** Suppose that there are three advertisers, $A, B$, and $C$. There are three queries, $x, y$, and $z$. Each advertiser has a budget of 2. Each query cost 1 unit. Advertiser $A$ bids only on $x$; $B$ bids on $x$ and $y$, while $C$ bids on $x, y$, and $z$. Note that on the query sequence $xxyyzz$, the optimum off-line algorithm would yield a revenue of 6, since all queries can be assigned.

(a) [5 points] Show that the greedy algorithm will assign at least 4 of these 6 queries.

(b) [5 points] Find another sequence of queries such that the greedy algorithm can assign as few as half the queries that the optimum off-line algorithm assigns on that sequence.