CS512 (Spring 2015) "Advanced Data Mining": Midterm Exam II

(Thursday, Apr 30, 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, with no points earned.

Name: NetID: Score:

1 [40]	2 [20]	3 [30]	4 [10]	Total [100]

1. [40] Advanced Clustering

   (a) [15] Compare the following pairs of clustering methods and state their major differences:
   (i) SCAN vs. DBSCAN

(ii) δ-bi-Clustering vs. δ-pClustering

(iii) Fuzzy clustering vs. Gaussian mixture model-based clustering

(b) [18] (i) Why it is often necessary to do constraint-based clustering? How to do clustering by taking "*must-link*" and "*cannot-link*" constraints?

(ii) Besides specifying "*must-link/cannot-link*" constraints, briefly describe *another* mechanism for users to provide guidance on clustering. Outline a clustering algorithm to handle your proposed mechanism.

(c) [7] (i) How to evaluate the quality of a clustering algorithm, if we are given the external ground-truth?

(ii) It is often expensive to obtain ground-truth clustering results for the entire data set, when the data set is extremely large. How to compare clustering algorithms even if we do not have the external ground-truth? (You may be allowed to ask a limited number of any reasonable questions to an external expert, but it is not required)

2. [20] Outlier Analysis

(a) [10] (i) What are the differences between *global outlier* and *local outlier*? Give an example for each respectively.

(ii) Explain why local outlier factor (LOF) can effectively detects local outliers.

(b) [10] (i) In a multidimensional database, outliers can be defined in different ways. Consider a database of all the transactions in Walmart. A transaction consists of *time*, *item*, *store location* etc. Give example and a possible interpretation of outliers defined in the subspace *(time, item)*. Why such an outlier is not necessarily an outlier by merely looking at *time* or *item*?

(ii) When the number of dimensions is extremely large, what are the major challenges of mining outliers in such a high-dimensional database? Outline an approach to detecting high-dimensional outliers.

3. [30] Stream Data Mining

   (a) [10] Outline three essential design ideas of a stream data cube that will facilitate monitoring Twitter data and drill down by *time*, *location*, and *hashtags*?

(b) [10] Suppose one would like to find frequent single items (such as frequent router address) in dynamic data streams, with a minimum support σ and error bound ϵ. Explain why the *lossy counting* algorithm can find the frequent items with a guarantee error bound ϵ.

(c) [10] Outline a method that performs effective clustering in a dynamic evolving data stream.

4. [10] General Application

Suppose we have a system monitoring the Tweet stream, where one can obtain tweets' *user id*, *content*, *location*, *time*. An interesting idea is to develop a "social sensor" based on such a system, which is able to detect events from tweets posted by users even faster than news agents.

(a) Consider when a certain disaster (e.g. earthquake) happens, what metric might be used to effectively and promptly detect such a disaster?

(b) Outline a method to effectively detect such a disaster in real time.