CPS196.03 Information Management and Mining - Spring 2009
Assignment 1

- Due date: Thursday, Jan. 29, 2009, in class (2.50 PM). Late submissions will not be accepted.
- Submission: In class, or email solution in pdf or plain text to shivnath@cs.duke.edu.
- Do not forget to indicate your name on your submission.
- State all assumptions. For questions where descriptive solutions are required, you will be graded both on the correctness and clarity of your reasoning.
- Email questions to shivnath@cs.duke.edu.
- Total points = 100.

Question 1 **Points 15**

Question 2 on Page 404 (Chapter 6) of the reference textbook by Pang-Ning Tan and others. Chapter 6 of the reference textbook is available at: http://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf

Question 2 **Points 15**

Question 3 on Page 405 (Chapter 6) of the reference textbook by Pang-Ning Tan and others.

Question 3 **Points 15**

Question 8 on Page 406 (Chapter 6) of the reference textbook by Pang-Ning Tan and others.

Question 4 **Points 20**

Consider Table 6.22 on Page 404 (Chapter 6) of the reference textbook by Pang-Ning Tan and others. We want to run the Apriori algorithm on this market-basket dataset. Each transaction represents a basket. Answer the following questions. The support threshold is 40%.

1. How many passes will Apriori make in this case?
2. Give the frequent itemsets found in each of Apriori's passes.
3. Recall the two methods that we discussed in class on how counters can be maintained in Apriori. (See Slide# 31 on Notes 2 posted on the class Web site.) For both these methods, give the amount of memory that Apriori will use in its first two passes over the data.

Question 5 **Points 20**

This question asks you to give a scenario where the Park-Chen-Yu (PCY) algorithm performs better that the Apriori algorithm. For the scenario you come up with, write down the market-basket dataset and the hash function, and explain clearly why PCY is better than Apriori in this case.

Question 6 **Points 15**

This question asks you to give a scenario where the Savasere-Omiecinski-Navathe (SON) algorithm performs better that the Apriori algorithm. For the scenario you come up with, write down the market-basket dataset, and explain clearly why SON is better than Apriori in this case.