University of Ioannina,
Computer Science Department

Data Mining – CS 059
Fall 2012

Assignment 1

This is the first part of Assignment 1. The deadline for this part of the assignment is at the beginning of the class on October 30th. You should turn in the code for question 1, and submit the remaining questions either electronically, or on paper. You are encouraged to submit your assignment on Friday October 26th, since the second part of Assignment 1 will be given then. For late submissions the late policy on the page of the course will be applied. The details for the turn-in are on the page of the course.

Question 1 (Reservoir Sampling)

In class we described the Reservoir Sampling algorithm for sampling a single item from a stream of N items. In this question you are required to modify the algorithm to sample $\boldsymbol{k}$ items.

1. Describe the algorithm for sampling k items uniformly at random from a stream of N items. The algorithm should work in a single pass over the data, reading the items one by one, without prior knowledge of the size of the stream N, and using $O(k)$ of memory (assume the size of an item is fixed).
   (**Hint**: In a random sample each element should have probability k/N to appear in the sample).
2. Prove that your algorithm produces a uniform sample, that is, prove that for every $i, 1 \leq i \leq N$, the i-th element has probability k/N to appear in the sample.
   (**Hint**: What is the probability that after the i-th item is selected, it is later replaced in the j-th item, for $j > i$?)
3. Write a program that implements the sampling algorithm (in any language you want). Your program should sample k lines from a file. Use the standard input for input, and the standard output for output. Your program should work as follows:
   "`sample <inputfile >outputfile`".

Question 2 (Sampling edges)

We have a graph G, with N nodes and M edges, and degree d_x for each node x. We want to sample an edge from the graph uniformly at random (i.e., each edge has probability $1/M$ to be sampled).

Consider the following algorithm for sampling an edge:

- Select a node uniformly at random from the set of nodes (i.e., with probability $1/N$)
- Select one of the edges incident on the selected node x uniformly at random (i.e., with probability $1/d_x$) and output the edge.

Answer the following questions:

1. Give the probability of selecting edge (x, y) as a function of the number of nodes N, and the degrees d_x, d_y of nodes x and y respectively.
2. Give a counter example that shows that the algorithm does not produce a uniform sample (i.e., it does **not** hold that the probability of every edge to be selected is $1/M$).
3. Modify one of the steps of the algorithm above to produce a uniform sample.
   (**Hint**: What is the relationship between M and the degrees of the nodes?)