TEL AVIV UNIVERSITY
Department of Computer Science
0368.3239 – Leveraging Big Data
Fall Semester, 2013/2014

Homework 1, October 31, 2013

- **Due on November 17.**
- **You are allowed to consult any sources or other humans, but you must write and submit the solution yourself and understand it in detail.**
- **Please submit a PDF file to the email address that will appear on the Web site. The file should be named `firstname_lastname_HW1.pdf` . Keep a copy. Include your full name and ID in the file.**

1. **Maximum Likelihood Estimator (MLE) for Morris counters**: (Use a computer) We had seen that the estimator $\hat{n}_U(x) = 2^x - 1$, when the Morris counter is x, is an unbiased estimator of the true count n.

a. Compute the MLE, $\hat{n}_{\text{MLE}}(X)$, when the Morris counter value is $x = 1, \ldots, 10$.

b. For $n = 1, \ldots, 10$, compute and show the following in a table.

(a) The bias of $\hat{n}_{\text{MLE}}$, $(\text{Bias}(\hat{n}_{\text{MLE}}, n) = E[\hat{n}_{\text{MLE}}] - n)$.

(b) The Mean Square Error (MSE) of $\hat{n}_{\text{MLE}}$, which is the expectation of the square of the difference between the estimate and n.

(c) The Normalized Root Mean Square Error (NRMSE) of the MLE, which is the ratio of the square root of the MSE to n.

(d) The Coefficient of Variation (CV) of the unbiased estimator (which is also its NRMSE...)

c. Prove the relation $\text{MSE}[\hat{\theta}] = \text{Var}[\hat{\theta}] + \text{Bias}(\hat{\theta}, \theta)^2$, for every estimator $\hat{\theta}$ of θ.

d. Consider k independent Morris counters $x_1, \ldots, x_k$ for $n = 8$. Write the NRMSE of $\frac{1}{k}\sum_{i=1}^{k} \hat{n}_{\text{MLE}}(x_i)$ and $\frac{1}{k}\sum_{i=1}^{k} \hat{n}_U(x_i)$ as a function of k and using the values in the table for $n = 8$.

2. **Misra Gries with decrements (MG±):**

Consider a stream that includes both increments and decrements to the count of each item. Stream elements have the form (i, Δ), where i is an item ID and $\Delta \in \{+1, -1\}$.

The true count c_i of an item i is as follows: Initially, $c_i \leftarrow 0$. After processing stream element (i, Δ), $c_i \leftarrow \max\{0, c_i + \Delta\}$. Note that only decrements that can be matched with previous increments matter.

The MG± algorithm is identical to MG when processing elements that are increments. When we see $(i, -1)$, MG± does as follows: if there is a counter for i, we decrement it by 1 (and remove the counter altogether if its value is 0). Otherwise, if there is no counter, we do nothing.

The estimator $\hat{c}_i$ for c_i with MG± is the same one used with MG: If there is a counter for item i, we return its value. Otherwise, we return 0.

a. Can the estimate $\hat{c}_i$ ever exceed c_i ?

b. What is the tightest upper bound you can give on $c_i - \hat{c}_i$? (in terms of an easy to track property of the stream such as the total number of increments and/or decrements). Prove your answer.

3. **Cisco's sampled NetFlow**: The collection of flow statistics at Internet routers is essential for good network management (and other applications). The IP protocol used for Internet routing breaks each flow to many packets. The router handles a stream of intermixed IP packets from different flows. The statistic collection aims to present an aggregate view over individual IP flows and their respective packet counts. One standard for statistics collection is Cisco's Netflow, which counts the number of packets in each flow. A version of it, *Sampled Netflow* addresses resource constraints on the time to process each packet and on the total number of counters (each counted distinct flows uses a counter). With sampled NetFlow, each packet is sampled with probability p. The sampled packets are then aggregated into flows.

More precisely, the streaming algorithm samples each packet with probability p. The flow key (which includes source and destination IP addresses, protocol) is identified. If a counter exists already for the flow key then it is incremented. Otherwise, a new counter is initiated with value 1.

a. Give an unbiased estimator $\hat{c}$ for the size c (number of packets) of a flow. (You need to specify the estimate if the flow key is not sampled at all and if it is sampled and the count is x.). Prove that your estimator is indeed unbiased.

b. What is the variance of $\hat{c}$ as a function of the flow size c ? (**hint:** consider the c packets individually)

c. What is the Coefficient of Variation of $\hat{c}$?

d. Are sampled NetFlow summaries mergeable ? Explain.

e. Consider a stream of m packets. How would you split it into flows as to maximize the expected number of counters? Prove your answer (hint: consider each packet of the m packets individually. what is the probability that it creates a new counter as a function of its place in its flow). What would the expected size be in this case. Explain.