Quiz 1

Name:

1. Paul has a trouble. He has a big farm and 10,000 chickens. Every day the chickens range free on the farm and come to cages after sunset except exactly one chickens. Every chicken has a number on its wing. The problem is: he wants to know which chicken is missing without using any computer. He is very good at calculation but cannot memorize 9,999 numbers.

How can he find the missing number?

Let s = 10,000 * (10,000 + 1) / 2

With every chicken of ID i, calculate s = s - i

Finally s is the missing number!

2. In the farm, Paul has 2 cages. Every day after sunset his chickens go into one of the cages. He wants to know the ratio of common chickens in Cage 1 between yesterday and today. For simplicity, suppose we have 5 chickens {1, 2, 3, 4, 5}. The chickens in Cage 1 yesterday and today are Sy = {1, 3, 5} and St = {1, 2, 4, 5} respectively. Approximately calculate the Jaccard coefficient between Sy and St. You may need some permutations below:

P1 = < 4 1 2 5 3 >

P2 = < 5 3 4 1 2 >

P3 = < 1 2 5 3 4 >

P4 = < 3 1 2 5 4 >

P5 = < 2 5 4 3 1 >

Compute signature of Sy ans St respectively.

Sig(Sy) = < min (P1[1], P1[3], P1[5]) = 2,

min (P2[1], P2[3], P2[5]) = 2,

1,

2,

1>

Similarly,

Sig(St) = < 1, 1, 1, 1, 1 >

Jaccard(Sy, St) = $\frac{\#\ of\ common\ min-hash\ values}{\#\ of\ min-hash\ values} = \frac{2}{5}$

Describe a data stream algorithm of computing means and standard deviation with a real number data stream.

```
c = 0
s = 0
ss = 0
For each element v,
    c += 1
    s += v
    ss += v*v

mean = v / c
stddev = sqrt( ss / c - mean * mean)
```

What is difference between Times Series, Cash Register and Turnstile Models?

* Time series: simply, keep the incoming element for a state (e.g., sampling)

* Cash register: update some states with every incoming element and the states always increase (e.g., Count-min sketching)

* Turnstile: update some states with every element and the states sometimes increase, sometimes decrease (e.g., Count sketching)

Suggest an approximate algorithm for the following problem:

Paul sees data stream of positive integers representing A_P and Carole sees data stream representing A_C, both on domain 1,...,N. Design a streaming algorithm to determine certain number of i's with the largest $\frac{A_P[i]}{A_C[i]}$.

S_P = {}, S_C = {}

maxratio = -1

for each input i

Paul computes 10 largest A_P[i]s using a heap and keeps the corresponding i in S_P

Carole computes 10 largest A_C[i]s using a heap and keep the corresponding i in S_C

Among common i S_P and S_C, answer the largest $\frac{A_P[i]}{A_C[i]}$

Paul wants to group a data stream of multidimensional data points into k clusters. Eventually, what he wants compute is k centers such that the sum of minimum distance from every point to one of them becomes the smallest. That is, the optimal solution is as follows:

Given n data points $\{p_1, p_2, \ldots, p_n\}$, the optimal k centers $C = \{c_1, c_2, \ldots, c_k\}$ of clusters are

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{c \in C} \|p_i - c\|_2^2$$

Provide an approximate algorithm to handle streaming data.

Randomly initialize cj (j=1,…,k)

Nj = 1 (j=1,…,k)

For each data point pi

Find the closest center cj among c1,…,ck

cj = (cj*Nj + pi) / Nj

Tell me a problem that requires to handle large streaming data in YOUR research topics (i.e., topics you (or your supervisor) are studying in your lab)