CSC4524: Streaming algorithms

Reservoir sampling

Mini-project 4

- **Sub-sample the stream to a fixed size representative sub-set**
  - Input: k the size of the desired sub-set
  - Output: representative sub-set of size k
  - Work due to 09-12-2019 23:59

Solution

Reservoir sampling

Reservoir sampling

Reservoir sampling is a family of randomized algorithms for choosing a simple **random sample** without replacement of k items from a **population of unknown size** n in a single pass over the items. The size of the population n is not known to the algorithm and is typically too large to fit all n items into memory. The population is revealed to the algorithm over time, and the algorithm cannot look back at previous items. At any point, the current state of the algorithm must permit extraction of a simple random sample without replacement of size k over the part of the population seen so far.

Reservoir sampling

Data stream

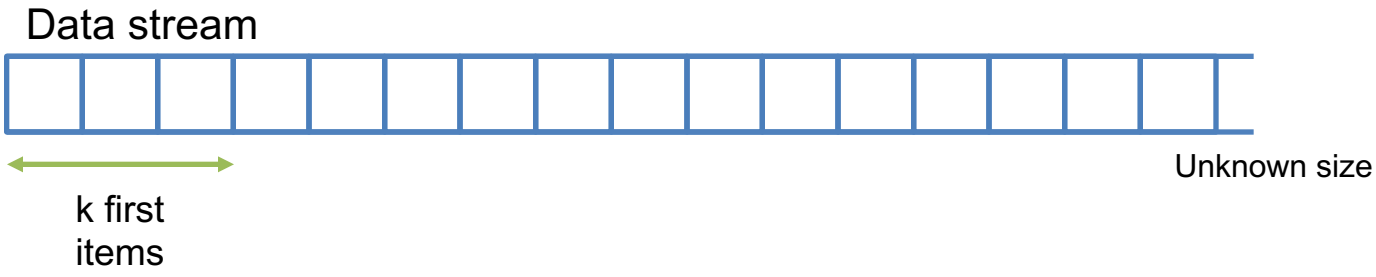
Unknown size

Sample

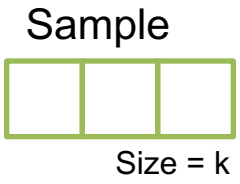
Size = k

Goal : Select each item in the stream with equiprobability.

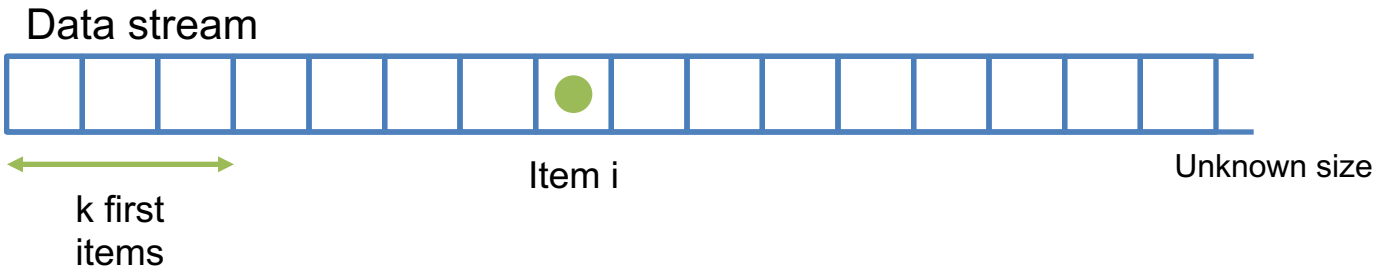
Reservoir sampling



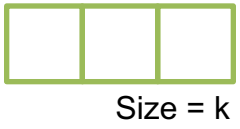
Step 1 : The first k items are selected with probability 1



Reservoir sampling



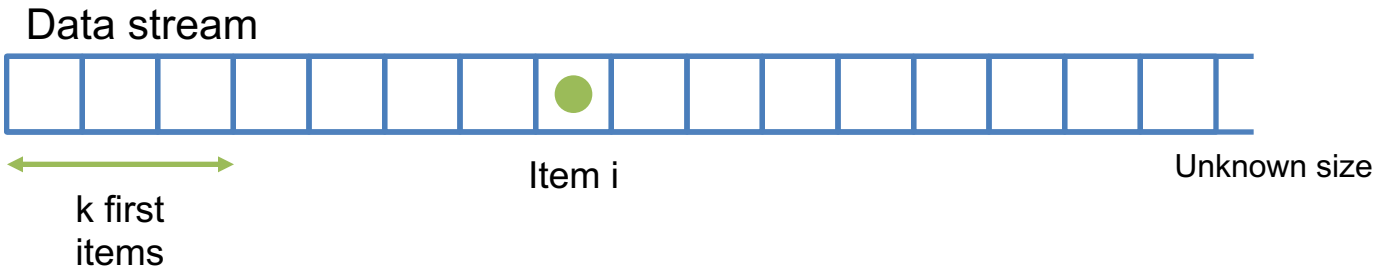
Step 1 : The first k items are selected with probability 1



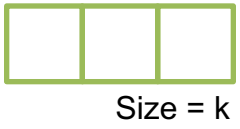
Step 2 : Each item I such as i > k is selected with probability $\frac{k}{i}$.

It replaces any item already in the sample with probability $\frac{1}{k}$.

Reservoir sampling



Step 1 : The first k items are selected with probability 1



Step 2 : Each item I such as i > k is selected with probability $\frac{k}{i}$.
It replaces any item already in the sample with probability $\frac{1}{k}$.

Conclusion : At any given time this process guaranties that each item has $\frac{k}{n}$ chance to be selected, given n the number of examined items.

Complexity

- **Time**

$$O(\#elements)$$

- **Memory**

$$O(k)$$

Next step

Paper lectures