CS535 BIG DATA

PART B. GEAR WORKSHOP I
ADVANCED BIG DATA ANALYTICS
CASE STUDY

Sangmi Lee Pallickara
Computer Science, Colorado State University
http://www.cs.colostate.edu/~cs535

FAQs
- Programming Assignment 2
  - 5:00PM March 29 (Adjusted)
  - Team submission

Today's topics
- Lossy Counting Algorithm
- Workshop presentations

Programming Assignment 2
Lossy Counting Algorithm

Algorithm
- Divide the incoming stream into buckets with the bucket size of \( w = \frac{1}{\varepsilon} \)
- Each buckets are labeled with integer starting from 1
- Current bucket number = \( b_{current} \)
  - \( b_{current} = \frac{N}{w} \)
  - Where \( N \) is the total count so far
- "True frequency of an element \( e' \) = \( f_e \)
- Data structure
  - \( (e, f, \Delta) \)
  - \( e \) is an element in the stream
  - \( f \) is an integer representing its estimated frequency
  - \( \Delta \) is a maximum possible error of \( f \)

http://www.cs.colostate.edu/~cs535
Delete phase

- When an element arrives
  - Lookup to see if there is an entry for that element already exists
  - If there is an entry, increase its frequency by one
  - Otherwise, create a new entry of the form (x, f, Δ)

- When the new elements fill up the bucket
  - N mod k == 0

- Prune elements
  - (x, f, Δ) is deleted if f + Δ > threshold

- When user request a list of item with threshold s
  - Outputs are items that (x, f, Δ)N

New elements added has maximum count error of 0

Example (ε = 0.2, w = 1/ε = 5), 1st bucket

- Insert b (Bucket 2)
  - New elements added: 1 2 3 4
  - Delete phase: delete elements with f + Δ > threshold
  - Outputs: (x=1; f=1; Δ=0)

New elements added has maximum count error of 0

Example (ε = 0.2, w = 1/ε = 5), 2nd bucket

- Insert b (Bucket 2)
  - New elements added: 3 4 5 6
  - Delete phase: delete elements with f + Δ > threshold
  - Outputs: (x=1; f=1; Δ=0)

New elements added has maximum count error of 0

Example (ε = 0.2, w = 1/ε = 5), 3rd bucket

- Insert b (Bucket 2)
  - New elements added: 7 3 3 6 1
  - Delete phase: delete elements with f + Δ > threshold
  - Outputs: (x=1; f=1; Δ=0)

New elements added has maximum count error of 0

When user request a list of item with threshold s

- Outputs are items that (x, f, Δ)N

- When user request a list of item with threshold s
  - Outputs are items that (x, f, Δ)N
  - Elements with frequencies ≤ 1 are deleted
  - Elements with frequencies ≤ εN are deleted
  - Elements with frequencies ≤ 2 are deleted

New elements added has maximum count error of 0

http://www.cs.colostate.edu/~cs535
GEAR Workshop I | Advanced Big Data Analytics Case Study
Workshop 1 Advanced Big Data Analytics Case Study
Workshop 2 Scalable Computing Models
Workshop 3 Large Scale Graph Analysis
Workshop 4 Scalable Data Storage, retrievals and analytics