A new algorithms class!

• Why do we need a new algorithms class?
  – Unprecedented amount of data containing a wealth of information.
    • Example: Twitter receives 6000 tweets per second which amounts to 500 million tweets per day with a storage requirement of ~640 gigabytes.
  – Traditional algorithms process data in RAM, sequentially and may have high time-complexity
    • Not suitable for processing Twitter data
Characteristics of Big Data

• VOLUME
  – Can not store the entire data in the main memory

• VELOCITY
  – Data changes frequently. Needs highly efficient processing, often parallel processing.

• VARIETY & VERACITY
  – Data coming from many different sources, and often contains noise-adds to the complexity of data processing
This Course

• Develop algorithms to deal with such data
  – Space and Time Efficient
  – Parallel Processing
  – Approximation & Randomization
• Theoretical course with main focus on algorithm analysis
  – Relevant applications will be discussed, and there will be plenty of coding exercises
  – But no software tools will be covered
• Background in basic algorithms (311) and probability (240) are strictly required.
Personnel

• Instructors & Teaching Assistants
  – Barna Saha
    • Email: barna@cs.umass.edu
    • Office Hour: Thur 12:45-1:45, CS336
  – David Tench
    • Email: dtench@cs.umass.edu
    • Office Hour: Wed 2:00-3:00 pm, CS 207
  – Raghavendra Addanki
    • Email: raddanki@cs.umass.edu
    • Office Hour: Mon 4:00-5:00 pm, CS207
Grading

• Homeworks (3-4) in a group of 2 to 4
  – Will consist of mathematical problems and/or programming assignments
  – Find your partners early and wisely. Do not come to me with complaints about your partner.
  – 30%

• Midterm [March 22\textsuperscript{nd}, in class]
  – 20%

• Final [University schedule, May 3\textsuperscript{rd}]
  – 30%

• Mini Coding/Programming Assignments
  – Few simple exercises to be done \textit{individually}
  – Roughly 4
  – 20%
Communication

• All class related discussions should be done through piazza.
  – Sign up from the course page.

• Course website
  – http://www-edlab.cs.umass.edu/cs590d/

• Homework submission
  – Must be submitted via moodle—no hardcopy submission
  – All codes must be submitted via Moodle
  – Absolutely no submission by email
Books

• Text Book: We will use reference materials from the following books. **Both can be downloaded for free.**
  
• **Mining of Massive Datasets**, Jure Leskovec, Anand Rajaraman and Jeff Ullman.
  
• **Foundations of Data Science**, a book in preparation, by John Hopcroft and Ravi Kannan
An Interesting Problem

• Suppose we see a sequence of items, one at a time.
• We want to keep a single item in memory.
• We want it to be selected at random from the sequence.
• Easy if we know the number of items “n”
  – Just draw a random number in between 1 and n
• What if we do not know n?
Reservoir Sampling

- Upon seeing the **first** element—keep it.
  - The first element is chosen with probability $\frac{1}{2}$.

- Upon seeing the **second** element—select the second element with probability $\frac{1}{2}$. If the second element is selected discard the first element.
  - The probability that the second item is sampled $= \frac{1}{2}$
  - The probability that the first item is sampled $= \frac{1}{2}$

- Upon seeing the **third** element—select it with probability $\frac{1}{3}$, if selected then discard the element that was previously selected.
  - The probability that the third item is sampled $= \frac{1}{3}$.
  - The probability that the second item is sampled $= \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$
  - The probability that the first item is sampled $= \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$
Reservoir Sampling

- Upon seeing the **fourth** element—select it with probability $\frac{1}{4}$, if selected then discard the element that was previously selected.
  - The probability that the fourth item is sampled $= \frac{1}{4}$.
  - The probability that the third item is sampled $= \frac{3}{4} \times \frac{1}{3} = \frac{1}{4}$
  - The probability that the third item is sampled $= \frac{3}{4} \times \frac{1}{3} = \frac{1}{4}$ The probability that the third item is sampled $= \frac{3}{4} \times \frac{1}{3} = \frac{1}{4}$

- Can you generalize the algorithm to any $i$?

- Upon seeing the $i$th item—select it with probability $\frac{1}{i}$, if selected, discard the element that was previously selected.
  - The probability that the $i$th item is sampled $= \frac{1}{i}$.
  - The probability that the $j$th $j < i$ item is sampled $= \frac{i-1}{i} \times \frac{1}{i-1} = \frac{1}{i}$
Reservoir Sampling

Upon seeing the **fourth** element—select it with probability \( \frac{1}{4} \), if selected then discard the element that was previously selected.

- The probability that the fourth item is sampled = \( \frac{1}{4} \).
- The probability that the third item is sampled = \( \frac{3}{4} \times \frac{1}{3} = \frac{1}{4} \).
- The probability that the third item is sampled = \( \frac{3}{4} \times \frac{1}{3} = \frac{1}{4} \). The probability that the third item is sampled = \( \frac{3}{4} \times \frac{1}{3} = \frac{1}{4} \).

**Can you generalize the algorithm to any \( i \)?**

Upon seeing the \( i \)th item—select it with probability \( \frac{1}{i} \), if selected, discard the element that was previously selected.

- The probability that the \( i \)th item is sampled = \( \frac{1}{i} \).
- The probability that the \( j \)th \( j < i \) item is sampled = \( \frac{i-1}{i} \times \frac{1}{i-1} = \frac{1}{i} \).

What happens when the reservoir can store “s” elements?
Reservoir Sampling!
Sampling

- A very useful method to obtain appropriate summary of data
- Will learn more in the coming classes
- But needs to be done with care
- Link to video
  https://www.youtube.com/watch?v=xmhVdsOTh1E
Mini Exercise-1

• Implement reservoir sampling when reservoir has size 1. Let the items from 1 to 100 appear one by one.
  – Report the item sampled in one run of the algorithm.
  – Repeat the algorithm for 1000 times and plot the number of times each element is selected.
  – Repeat the algorithm for 10000 times and plot the number of times each element is selected.
  – Repeat the algorithm for 100000 times and plot the number of times each element is selected.

2. Suppose $n$ is the total number of items that arrived. Show that the probability of selecting a particular set of $s$ items in the reservoir sampling algorithm is $\frac{1}{\binom{n}{s}}$.

  – DUE: Tuesday, 30th.
Next Few Classes

• Probability review before we enter into the more interesting regime!