Lecture Note 1

1 Info

1.1 General Introduction

1.1.1 Course website

link

1.1.2 Time

Mon 7:00 p.m. to 9:30 p.m.

1.2 Data Science Algorithms

- The story of Cholera outbreak

1.3 Data sources plentiful

- Monitoring of environment
- Cheap sensors
- Traffic lights
- Safety concerns - videos

1.4 Storage / Analysis / Stream processing

1.5 Textbook

Mining of massive Datasets by Leskovec, Rajaraman, Ullman. link

2 Notion of Algorithm running time

There are abundant resources about it on the Internet. Please search "big O notation" on your favourite search engine. Here is a nice tutorial: link.

3 Streaming

3.1 Dynamic Graphs

A dynamic graph is basically a graph that is subject to a sequence of updates. One common case for dynamic graph is that vertices are known before hand, but edges come in a stream.

3.2 Google

- web crawl
- thanksgiving recipes
- indexing web

4 Bonferroni Principle

link

- 1 billion population (10^9)
- Each person on average goes to a hotel 1% of the time (1 day in 100)
- Each hotel holds 100 guests: #hotels is 10^5 (why ?), since each day 10^7 people go to a hotel and there are 10^5 hotels needed to hold them
- Access 3 years of records (1000 days)
- Look for people who on two different days went to the same hotel

4.1 Calculation

Assume random behavior

- Prob of deciding to go to a hotel is $0.01 = 1/100$, picking hotel at random
- A and B both decide to go th a hotel on the same day is $1/10^4$

- Chance of visiting a hotel again is 10^{-18}
- n choose 2 = $n^2/2$
- How many events mistake we flag?
  - pairs of people is $1/2\ (10^9)^2 = 5 * 10^{17}$
  - pairs of days is $5 * 10^5$ ($1/2*1000^2$)
- Suspicious events
  - #pairs * #pairs-of-days * prob
  - $5*10^{17} * 5*10^5 * 10^{-18}$ = 250,000

5 Random Sampling

5.1 Stream

5.2 Windows

- webtraffic
- traffic through a router
- Reservoir sampling of a single item

6 Hashing function?

Intuitively speaking, we can think of a hash function as randomly throwing a dart onto a map, but ensures that the same input will be hashed at the same position

6.1 Estimate the number of distinct elements

For each coming input x, remember the smallest or largest $h(x)$ we have seen. Intuitively speaking, the more darts we have, the more likely we have some large $h(x)$. From this $h(x)$, we can estimate the number of distinct elements.