Lecture Note 2

Material covered today is from Chapter 1 and chapter 4

1 Bonferroni Principle

1.1 Idea

Get an idea the frequency of events when things are random

- 1 billion = 10^9
- Each person has a 1% chance to stay in a hotel
- Each hotel has 100 single rooms
- 1/100 * 10^9: #people staying in hotels today
- #hotels = 1/100 * 1/100 * 10^9 = 10^5, or 100,000
- Sample a window of 1000 days

1.2 Want

Suspected terrorists meet and stay in the same hotel twice in the 1000 day period. How many such events might we expect? In other words, we want to identify a pair of persons (A, B) who stay at the same hotel on two different days d_1, d_2.

1.2.1 Probability

- $\Pr[A \text{ stays at a hotel on } d_1] = \frac{1}{100}$
- $\Pr[B \text{ stays at a hotel on } d_1] = \frac{1}{100}$
- The probability that they both visit on d_1 is $\frac{1}{100} \cdot \frac{1}{100} = 10^{-4}$
- The probability that they stay at the same hotel: $10^{-4} \cdot 10^{-5}$
- It happens on both d_1 and d_2: $p = (10^{-4} \cdot 10^{-5})^2$

1.2.2 Number of events

- #ways to select a pair of individuals is $\binom{n}{2} = \frac{n(n-1)}{2} \simeq \frac{1}{2} \cdot n^2$
- #choices for $(A, B, d_1, d_2) \simeq \frac{1}{2} \cdot n^2 \times \frac{1}{2} \cdot d^2$
- Final answer $\frac{1}{2} \cdot n^2 \times \frac{1}{2} \cdot d^2 \times p = \frac{1}{4} \times 10^6$

1.2.3 Exercise

What if #days raised to 2000?

2 How do you sample when you do not know the size of population

2.1 Reservoir Sampling

2.2 Algorithm

Go to the first house, and we need to pick it with probability 1 and call it our sample. If there is no more house, we report this house as our answer. If there are still more houses, we continue and pick the second house with probability 1/2, call it our sample, kicking out the first house. If there is no more house, we report our sample. Otherwise, we will keep going. For the k-th house, we pick it with probability $1/k$, and replace the existing choice.

2.3 Analysis

We prove that all elements got chosen with equal probability (which is weaker than what Reservoir Sampling actually provices) by induction. Where there is only one house, we select with probability 100%, so the claim is true when $n = 1$. Suppose the claim is true for n houses, we now prove it also works for $n + 1$. According to our assumption, before seeing the $(n + 1)$-th house, each house from 1 to n will have $\frac{1}{n}$ probability of being our sample. Upon seeing the new house, we will select it with probability $\frac{1}{n+1}$, which means it will be our sample with probability $\frac{1}{n+1}$. On the other hand, with probability $\frac{n}{n+1}$, the original sample will persist. So the probability of a certain house from 1 to n persist to be our sample would be $\frac{1}{n} \times \frac{n}{n+1} = \frac{1}{n}$.

2.4 Reservoir Sampling Multiple items

- Goal: pick p items randomly out of n items.
- No knowledge of n.
- Add new item with probability $\frac{p}{n}$. If we end up not adding this new item, we do nothing. If this item gets chosen, one of our old samples will be kicked out uniformly at random to make space for the new item.

2.5 Property of Reservoir Sampling

When selecting a single item, Reservoir Sampling guarantees that every item has the same chance of being chosen. However, when we pick more than one item, Reservoir Sampling provides more than that. It guarantees that every subset is chosen with same probability

2.5.1 What does it mean?

- We have A, B, C, D. $p = 2, n = 4$, and our algorithm is as follows:
  - Toss a coin, return (A, B) if we see head, and (C, D) otherwise.
- Each letter will be sampled with probability 1/2, but we will never see (A, C).
- So Reservoir Sampling is a bit stronger

3 Sampling from a Stream

We want to store $\frac{1}{10}$-th of the stream for analysis

4 Hash function

4.1 A common hash functin

$h(x) = (ax + b \mod p)$, where p is a prime.

4.1.1 Bad example

$h(x) = x \mod B$, where $B = 10$. Then hash even numbers will always get even numbers

4.2 An obvious approach

Generate a random number 0-9, save the query if it is 0.

4.2.1 Bad example

Fails to answer the following query: among all the different queries, what fraction are duplicate queries?

- Suppose a user has issued s search queries one time and d search queries twice, and no queries more than twice. We will have $s + 2d$ queries in total.
- Then the correct answer would be $\frac{d}{s+d}$, since there are $s + d$ different queries, and d have duplicates.
- If we use the previous algorithm, we would get
  - $\frac{d}{100}$ will appear twice: only queries that have duplicates will ever appear twice. Further more, it will appear twice if and only if both got sampled. The probability that both queries got sampled is $\frac{1}{d}$. With d queries that have duplicates, we would expect $\frac{d}{100}$ such appearances.
  - Of the d queries with duplicates, $\frac{18d}{100}$ will appear exactly once. Either only the first query got sampled, or only the second query got sampled. So the probability that a certain duplicate query appear exactly once is $\frac{1}{10} \times \frac{10-1}{10} + \frac{10-1}{10} \times \frac{1}{10} = \frac{18}{100}$. So the total expected number would be $\frac{18d}{100}$.
  - In all, we have $\frac{d/100}{s/100+19d/100} = \frac{d}{10s+19d}$.

4.3 Use hash to sample 1/10 of the users

Use $h(\text{user})$, and select the set $\{u|h(u) = 0\}$.

4.4 Algorithm

Use a hash function on user id (so that if a user come again, we will know whether he was previously sampled or not). Sample user when hash to 0. By this method, we can guarantee that each user is sampled with probability 1/10.

5 Bloom Filters

Goal is to answer membership queries in S.

- If $x \in S$, always say yes
- If $x \notin S$, might say yes (with small probability of error)

5.1 Example

- 1 billion items, $|S| = m$
- Hash table a of size $n << m$
- y different values to insert
- When some value v come, change $a[h(v)]$ to 1
- When asked if a certain value v is in S, answer yes if and only if $a[h(v)]$ is 1

5.2 Analysis

What is the chance that a certain cell in a is 0?

- Suppose we have y darts and x targets
- Probability a single dart does not touch a specific entry:
  - $\left(\frac{x-1}{x}\right) = \left(1 - \frac{1}{x}\right)$
- Probability that all darts (we have y darts) miss this entry:
  - $\left(1 - \frac{1}{x}\right)^y = \left((1 - \frac{1}{x})^x\right)^{\frac{y}{x}} = e^{-\frac{y}{x}}$ for large x.
  - If we have $x = 8 \times 10^9$, $y = 10^9$, then probability an element is hit is $1 - e^{-1/8} \simeq 0.1175$

5.3 What if we have k hash functions?

Suppose $|S| = m$, hash function is of n cells, and we have k hash functions.

- $y = km$, $n = x$
- On seeing v, we set $a[h_i(v)]$ to 1, for all hash function $h_1, \ldots, h_k$.
- When querying v, we answer yes if and only $a[h_i(v)] = 1$ is true for all k hash functions.
- Chance that a certain cell in a is 0: $e^{-\frac{km}{n}}$
- Chance that we get false postive, i.e. a value v is not in S, but we thought it is in: $\left(1 - e^{-\frac{km}{n}}\right)^k$.
- If $k = 2, n = 10 \cdot m, y = k \cdot m$
  - The probability is: $\left(1 - e^{-\frac{1}{5}}\right)^2 \simeq 0.0329$.

6 Count Min Sketch

Want to count the frequency of number.

6.1 Algorithm

- Initialize an array a with 0
- When a number x come, use 2 hash functions h_1, h_2, and increase $a[h_1(x)]$ and $a[h_2(x)]$
- When asked the frequency of x, answer $\min\{a[h_1(x)], a[h_2(x)]\}$.
- truth $\leq$ estimate $\leq$ $(1+\epsilon)$ truth, with high probability

7 Verify matrix multiplication

7.1 Freivald's Algorithm

To check if $A = B \cdot C$, check if $Ar = B \cdot (C \cdot r)$ for random binary vector r.

8 Closest Pair

Instead of $O(N \log N)$, we can get expected running time $O(N)$ using randomization (answer is always correct, but might take longer) [Rabin, and independently Khuller & Matias link]