CS6720 : Data Mining Homework 1

CS15M015 : Hasit Bhatt

Problem

There has been a terrorist attack by a team of two people. You are a detective and you want to identify the pair of terrorists through data mining. These two terrorists must have stayed at least twice in the same hotel on the same day to plan their attack (You can get this information by collecting hotel logs).

Data Assumptions

- Number of people: 10^9

- Tracked over 1000 days : 10^3

- A person stays in a hotel with a probability of 1% on any given day

- Each hotel hosts 100 people at a time

If there are no terrorists(i.e., everyone behaves randomly), how many pairs of people will be classified as terrorist pairs?

Solution

Inferred Data

- Number of people is 10^9 and with 1% probability a person stays in hotel. Therefore, approximate number of people living in a hotel per day = $10^9 \times 0.01 = 10^7$.

- One hotel hosts 100 people at a time, therefore, approximate number of hotels needed per day is $\frac{10^7}{100} = 10^5$

Calculation

Expected number of pairs classified as terrorists = N_{sp} x N_{pd} x P_p
$\quad N_{sp}$ = maximum possible number of suspected pairs
$\quad N_{pd}$ = number of possible combinations of pair of two days
$\quad P_p$ = probability of a pair staying in the same hotel on any two days

- N_{sp} = combination of possible suspected pairs

  = all the pairs of two person from population of 10^9 people.

  $\Rightarrow N_{sp} = \begin{pmatrix} 10^9 \\ 2 \end{pmatrix} = \frac{10^9 \times (10^9 - 1)}{2} \approx \frac{10^9 \times 10^9}{2} \approx 5 \times 10^{17}$

- N_{pd} = Combination of pair of days = all the pairs from 10^3 days.

  $\Rightarrow N_{pd} = \begin{pmatrix} 10^3 \\ 2 \end{pmatrix} = \frac{10^3 \times (10^3 - 1)}{2} \approx \frac{10^3 \times 10^3}{2} \approx 5 \times 10^5$

- P_p = Probability of pair staying in same hotel on any two days

  So if P_{ab} = Probability of pair ab, containing a and b person being in same hotel on given day,

$$P_p = P_{ab}{}^2.$$

Here, the probability P_p is not $\begin{pmatrix} P_{ab} \\ 2 \end{pmatrix}$ because, terrorist pair can also stay in the same hotel some other day (with replacement).

Let's find the probability of a pair ab staying in same hotel P_{ab}.

$P_{ab} = P_a \times P_b \times$ probability of them choosing the same hotel

where, P_a and P_b are probability of a and b staying in hotel (0.01). And probability of choosing them same hotel will be 10^{-5} (probability of b choosing the same hotel as a)

$P_{ab} = P_a \times P_b \times 10^{-5} = 0.01 \times 0.01 \times 10^{-5} = 10^{-9}$.

Answer

Expected number of pairs classified as terrorists

$\approx 5 \times 10^{17} \times 5 \times 10^5 \times (10^{-9})^2$

$\approx 2.5 \times 10^5 = 250000$