CS648 Assignment - 2

Team Name:

Ja1nXR7AArKBa En7AoK

Soham Samaddar
200990

Aditya Tanwar
200057

March 2023

Notes

- We use RV to abbreviate the term “Random Variable".
- We might write phrases like “E with HP” or “E with LP” in our solutions. We use them as a shorthand for “Event E takes place with High Probability” and “Event E takes place with Low Probability” respectively.
- The base of logarithm everywhere is e unless specified otherwise/ explicitly mentioned.

Q1 Internalizing the proof of Chernoff Bound

Consider a collection $X_1, \dots, X_n$ of n independent geometrically distributed random variables with expected value 2. Let $X = \sum_{i=1}^{i=n} X_i$ and $\delta > 0$.

1. Derive a bound on $\mathbb{P}\left[X \geq (1+\delta)(2n)\right]$ by applying the Chernoff bound to a sequence of $(1+\delta)(2n)$ fair coin tosses.

Solution: Note that X_i can also be interpreted as the expected number of fair coins flipped till we get a heads. For a fair coin, the probability of getting a heads is 1/2, and hence the expected value of a geometric variable based on this success is $\frac{1}{1/2} = 2$, as desired. So, $X = \sum_{i=1}^{i=n} X_i$ can now be viewed as the number of fair coins flipped till we get n heads. Now, the event $X \geq (1+\delta)(2n)$ can mean one of two things:

1. The n^{th} head was flipped at exactly the $(1+\delta)(2n)^{\text{th}}$ coin flip.
2. Strictly less than n heads were flipped after $(1+\delta)(2n)$ coins were flipped.

Now, consider the following experiment and RV. We flip $(1+\delta)(2n)$ fair coins and count the number of heads obtained. **Let Y denote the number of heads obtained in the first $(1+\delta)(2n) - 1$ coin flips, and let Z denote the number of heads obtained in the last coin flip**. Now, we partition the event space in two, one subspace where the last coin flipped was heads and other where the last coin flipped was tails. This has been done to deal with the two cases as defined above.

$$\mathbb{P}\left[X \geq (1+\delta)(2n)\right] = \mathbb{P}\left[Y < n \mid Z = 0\right] \cdot \mathbb{P}\left[Z = 0\right] + \mathbb{P}\left[Y < n \mid Z = 1\right] \cdot \mathbb{P}\left[Z = 1\right]$$

$$
\begin{aligned}
&= \mathbb{P}\left[Y < n\right] \cdot \mathbb{P}\left[Z = 0\right] + \mathbb{P}\left[Y < n\right] \cdot \mathbb{P}\left[Z = 1\right] && \text{(Independent events)} \\
&= \mathbb{P}\left[Y < n\right]
\end{aligned}
$$

Essentially, the first term ($\mathbb{P}\left[Y < n \mid Z = 0\right] \cdot \mathbb{P}\left[Z = 0\right]$) deals with the case that strictly less than n heads were flipped throughout the experiment and the second term ($\mathbb{P}\left[Y < n \mid Z = 1\right] \cdot \mathbb{P}\left[Z = 1\right]$) deals with the case that at most n heads were flipped but the n^{th} head is guaranteed to be the last coin flip. Now, we can finally apply Chernoff's bound for Y (we get less than or equal to $n - 1$ heads in $2n(1+\delta) - 1$ coin flips):

$$
\begin{aligned}
\mathbb{P}\left[X \geq (1+\delta)(2n)\right] &= \mathbb{P}\left[Y \leq n - 1\right] \\
&= \mathbb{P}\left[Y \leq \left(1 - \frac{2n\delta + 1}{2n(1+\delta) - 1}\right) \frac{2n(1+\delta) - 1}{2}\right] \\
&\leq e^{-\frac{\left(\frac{2n(1+\delta)-1}{2}\right) \cdot \left(\frac{2n\delta+1}{2n(1+\delta)-1}\right)^2}{2}} \\
&= \boxed{e^{-\frac{(2n\delta+1)^2}{4(2n(1+\delta)-1)}}} \qquad (1.1)
\end{aligned}
$$

2. Directly derive a Chernoff like bound on $\mathbb{P}\left[X \geq (1+\delta)(2n)\right]$ from scratch.

Solution: Before doing the main calculation, we calculate $\mathbb{E}\left[e^{tX_i}\right]$ as a preliminary, where X_i is a geometrically distributed RV with success probability $1/2$, and t is any constant.

$$
\begin{aligned}
\mathbb{E}\left[e^{tX_i}\right] &= \frac{1}{2} \cdot e^t + \left(1 - \frac{1}{2}\right)^1 \cdot \frac{1}{2} \cdot e^{2t} + \left(1 - \frac{1}{2}\right)^2 \cdot \frac{1}{2} \cdot e^{3t} + \ldots \\
&= \sum_{k=1}^{k\to\infty} \left(1 - \frac{1}{2}\right)^{k-1} \cdot \frac{1}{2} e^{tk} \\
&= \sum_{k=1}^{k\to\infty} \left(\frac{1}{2}\right)^k e^{tk} \\
&= \sum_{k=1}^{k\to\infty} \left(\frac{e^t}{2}\right)^k \\
&= \frac{e^t}{2 - e^t} && \left(\text{Series converges iff } e^t/2 < 1 \Leftrightarrow \boxed{t < \ln 2}\ \right)
\end{aligned}
$$

We now bound the probability in a way similar to the one in lectures-

$$
\begin{aligned}
\mathbb{P}\left[X \geq (1+\delta)(2n)\right] &= \mathbb{P}\left[e^{tX} \geq e^{(1+\delta)(2n)}\right] && \text{(For any } 0 < t < \ln 2\text{)} \\
& && \text{(Follows from Th. 1 of Chernoff's Bound Lecture)} \\
&\leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{t(1+\delta)(2n)}} && \text{(From Markov's Inequality)} \\
&= \frac{\mathbb{E}\left[e^{t\sum_{i=1}^{i=n} X_i}\right]}{e^{t(1+\delta)(2n)}} && \left(X = \sum_{i=1}^{i=n} X_i\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\mathbb{E}\left[\prod_{i=1}^{i=n} e^{tX_i}\right]}{e^{t(1+\delta)(2n)}} \\
&= \frac{\prod_{i=1}^{i=n} \mathbb{E}\left[e^{tX_i}\right]}{e^{t(1+\delta)(2n)}} && \text{(From Independence of } X_i\text{)} \\
&= \frac{\prod_{i=1}^{i=n} e^t/(2-e^t)}{e^{t(1+\delta)(2n)}} \\
&= \frac{(e^t/(2-e^t))^n}{e^{t(1+\delta)(2n)}} \\
&= \frac{1}{e^{tn(1+2\delta)}(2-e^t)^n}
\end{aligned}
$$

This inequality holds for all values of $0 < t < \ln 2$, so we would like to find such a t which minimizes the expression on the right hand side, in order to allow us to get the tightest bound from this approach. We realise the following equivalences which follow from reciprocating and taking logarithm:

$$
\min\left(\frac{1}{e^{tn(1+2\delta)}(2-e^t)^n}\right) \Leftrightarrow \max\left(e^{tn(1+2\delta)}(2-e^t)^n\right) \Leftrightarrow \max\left(tn(1+2\delta) + n\ln\left(2-e^t\right)\right)
$$

We now differentiate the last expression and equate it to 0 in order to find the value of t:

$$
\begin{aligned}
\left(t(1+2\delta)n + n\ln\left(2-e^t\right)\right)' &= 0 \\
n(1+2\delta) + n\cdot\frac{-e^{t^*}}{2-e^{t^*}} &= 0 \\
1+2\delta &= \frac{e^{t^*}}{2-e^{t^*}} \\
2(1+2\delta) &= e^{t^*}(2+2\delta) \\
e^{t^*} &= \frac{1+2\delta}{1+\delta} \\
t^* &= \ln\left(\frac{1+2\delta}{1+\delta}\right) && < \ln 2 \;\forall\; \delta > 0
\end{aligned}
$$

Since this value of $t = t^*$ lies in the interval $(0, \ln 2)$, we plug this value of t into the inequality and simplify:

$$
\begin{aligned}
\mathbb{P}\left[X \geq (1+\delta)(2n)\right] &\leq \left(\frac{1+\delta}{1+2\delta}\right)^{n(1+2\delta)} \cdot \left(\frac{1}{2-(1+2\delta)/(1+\delta)}\right)^n \\
&= \left(\frac{1+\delta}{1+2\delta}\right)^{n(1+2\delta)} \cdot (1+\delta)^n \\
&= \boxed{\left(\frac{(1+\delta)^{2(1+\delta)}}{(1+2\delta)^{(1+2\delta)}}\right)^n} \qquad (1.2)
\end{aligned}
$$

3. Which bound is better?

Solution: We conducted an empirical analysis on the two bounds for different values of $n(=1, 10, 100, \ldots)$, and later tried to justify them analytically.

For every value of n, there seems to exist a $\delta^*(n)$, where the inequality between the bounds flips. For $\delta < \delta^*(n)$, the first bound is better, while for $\delta > \delta^*(n)$, the second bound provides a tighter result. $\delta^*(n)$ seems to be a decreasing function in n. We have tried to study how fast it decreases with respect to n.
We have, $\delta^*(10) \approx 0.15$, so even for $n > 10$, the second bound is better for typical values (of δ) of interest like $\delta = 1, 2, 3 \ldots$

Putting $\delta = 1/n$, and allowing $n \to \infty$, we have that the first bound is $e^{\frac{-9}{4(2n+1)}}$, while the second bound provides a trivial bound of 1. This lets us know analytically[1] that $1/n = \mathcal{O}(\delta^*(n)) \Leftrightarrow \delta^*(n) = \Omega(1/n)$.
Putting $\delta = 1/\sqrt{n}$, and again allowing $n \to \infty$, we have that the first bound is $e^{-1/(2\sqrt{n})}$, while the second bound is e^{-2}. Obviously in this case, the second bound is better. We conclude[1] that $\delta^*(n) \leq 1/\sqrt{n} \Rightarrow \delta^*(n) = \mathcal{O}(1/\sqrt{n})$.

Q2 Estimating all-pair distances exactly

Consider an undirected unweighted graph G on n vertices. For simplicity, assume that G is connected. We are also given a partial distance matrix M : For a pair of vertices i, j the entry $M[i, j]$ stores exact distance if i and j are separated by distance $\leq n/100$, otherwise M stores a symbol # indicating that distance between vertex i and vertex j is greater than $n/100$. Unfortunately, there are $\Theta(n^2)$ # entries in M, i.e., for $\Theta(n^2)$ pairs of vertices, the distance is not known. Design a Monte Carlo algorithm to compute exact distance matrix for G in $\mathcal{O}(n^2 \log n)$ time. All entries of the distance matrix have to be correct with probability exceeding $1 - 1/n^2$.

Algorithm

The algorithm takes in three parameters which have been elaborated below. Additionally, it creates a set (of vertices) S; BFS (Breadth-first search) is executed once for each vertex (possibly duplicate entries) in the set S. It also creates and returns a distance matrix $D_{n\times n}$ which stores the distance between any two vertices u and v to the best of its knowledge, i.e., with correctness probability $> 1 - 1/n^2$.

- $G = (V, E)$: The undirected unweighted graph with $|V| = n$ and $|E| = m$.
- $M_{n\times n}$: The partial distance matrix with the exact same definition as in the question's statement.
- P : The number of vertices (possibly duplicate) put into the set S, and Dijkstra's algorithm executed from.

The steps of the algorithm are as follows:

[1] Assuming that the point of intersection between the two bounds is unique for $\delta > 0$

```
Data: $G, M_{n\times n}, P$
$D[i,j] \leftarrow \infty,\ \forall\ i,j \in [1,n]$
$S \leftarrow \varnothing$
for $i = 1$ to $P$:
    Pick any[2] $v \in V$
    $S \leftarrow S \cup \{v\}$
for $s$ in $S$:
    Execute BFS on $G$ from $s$
    Update $D(s,v) = D(v,s) = d\ \forall\ v \in V$, where $d$ is from the BFS
for $v$ in $V$:
    for $u$ in $V$:
        if ( $M[v,u] = M[u,v] \neq \#$):     // Correct distance is in $M$
            $D[u,v] \leftarrow M[u,v]$
            $D[v,u] \leftarrow M[v,u]$
        else:
            for $s$ in $S$:
                $D[v,u] \leftarrow \min\{D[v,u], M[v,w] + M[w,u]\}$
                $D[u,v] \leftarrow \min\{D[u,v], M[u,w] + M[w,v]\}$
Return  $D$
```

Observe how the algorithm can be easily modified to return distance between a specific pair (or even a set of pairs); we have returned a distance matrix for all pairs[3] here.

The above algorithm shall be executed with input parameters ($G = G, M = M, P = 400 \log n$), to obtain a correctness probability of $1 - 1/n^4$ for each vertex pair.

Complexity Analysis

We use $p = P/|V| = (400 \log n/n)$ in our whole analysis for shorthand. The value of P is taken to be $(400 \log n)$ since that is the value with which we execute the algorithm.

The algorithm uses two auxiliary data structures S and D. The size of D is $\Theta(n^2)$, and the size of S is at most $P = 400 \log n$, or in other words, it is $\mathcal{O}(\log n)$. In any case, the algorithm uses $\Theta(n^2)$ auxiliary space.
This usage of auxiliary space can be removed altogether by modifying the algorithm's steps slightly, and using the matrix M to store distances in place of D. However, we have not included these modifications in the spirit of simplicity and being concise.

As for the time complexity, we analyze the five "major" steps of the algorithm:

- Initializing D takes $\mathcal{O}(n^2)$ time.
- The set S can be simulated by simply using a vector, so runtime of this loop is, at most, $\mathcal{O}(P) = \mathcal{O}(\log n)$.

[2] Randomly, Uniformly, and Independently from the previous iterations

[3] upto an error probability

- The number of times BFS is executed is $P = \mathcal{O}(\log n)$, since there are P vertices (possibly duplicate) in the set S, and each iteration corresponding to a vertex $v \in S$ runs BFS from it.
  Further, each run of BFS takes $\mathcal{O}(m+n)$ time using appropriate data structures. The time complexity due to the overall execution of BFS (over all iterations) is thus,

  $$\mathcal{O}(P \cdot (m+n)) = \mathcal{O}(m \log n + n \log n) = \mathcal{O}(n^2 \log n)$$

  The values of the matrix D are updated inside BFS, so the runtime of updating the entries of D is contained within the executions of BFS.

  To summarise, this step takes at most $\mathcal{O}(n^2 \log n)$ time.

- There are $\Theta(n^2)$ pairs, and updation of each pair's cell in the matrix D takes at most $|S|$ time. Since the size of S is $P = \mathcal{O}(\log n)$, the final updation of D, summed over all pairs, takes time at most $\mathcal{O}(n^2 \cdot P) = \mathcal{O}(n^2 \log n)$.

- Returning the matrix is trivial, it can take at most $\mathcal{O}(n^2)$ time.

Adding the runtime of all these five "major" steps, we conclude that the runtime of the algorithm is $\boxed{\mathcal{O}(n^2 \log n)}$.

Error Probability/ Correctness

Finally, we analyse the provided algorithm for its correctness. For any pair of vertices (u, v), we split the analysis into the following cases:

- $M[u, v] \neq \#$: In this case, since the matrix M stores the exact distance between the two vertices, we are assured that the cell $D[u, v]$ stores the correct entry as well.

- Let there be a vertex $w \in S$, such that w lies on *some* shortest path from u to v. In such a case, we have the following claims:

Claim 1: $D[w, v]$ stores the shortest distance from vertex w to vertex v.

Proof: Since $w \in S$, BFS must have been executed from w, thus the matrix D stores the exact distance from w to any other vertex $z \in V$.

Claim 2: $D[u, w]$ stores the shortest distance from vertex w to vertex u.

Proof: $D[u, w] = D[w, u]$ from the undirected nature of graph G. The rest of the argument is the same as the previous claim.

Claim 3: $D[u, v] = D[u, w] + D[w, v]$ and it is the smallest distance from vertex u to vertex v.

Proof: We have already shown that $D[w, u]$ and $D[w, v]$ store the shortest paths from vertex w to vertices u and v respectively.
We know from the optimal substructure property of shortest paths that $D[u, w] + D[w, v]$ stores the shortest distance path from u to v *containing* the vertex w.

Lastly, as $w \in S$, it will appear at least once when S is iterated corresponding to the vertex pair (u, v), and since, w lies on some shortest path from u to v, we obtain that $D[u, v] = D[u, w] + D[w, v]$

We have asked for w to lie on *some* shortest path from u to v, since there might be multiple paths of shortest lengths. However, we need only one such path to have a vertex from the set S. This fact will be crucial in the third (and final) case.

- Since the pair (u, v) did not fall under the 2nd case, we know that *no vertex* in the shortest path from u to v was selected in the set S.
  From the connected-ness of the graph G, we are prescribed to have at least one shortest path from u to v. Further, since the pair (u, v) did not fall into the 1st case, we know that there are at least $n/100$ vertices between them (on their shortest path). Select these vertices (along <u>any</u> shortest path) along with the vertices u and v.

  We have that from a <u>fixed</u> specific[4] set (of size <u>at least</u> $n/100$ vertices), *no vertex* was selected. Let this event be E. The probability of E taking place in a run of our algorithm, is bounded above by:

$$\begin{aligned}
\mathbb{P}[E] &\leq \left(1 - \frac{n/100}{n}\right)^{P} && \text{(Loop of ``picking a vertex'' runs } P \text{ times)} \\
& && \text{(There are at least } n/100 \text{ vertices in our set of interest)} \\
&= \left(1 - \frac{1}{100}\right)^{400 \log n} \\
&= \left(\frac{99}{100}\right)^{400 \log n} \\
&\leq e^{400 \log n \cdot \log(99/100)} \leq e^{-4.02 \log n} \leq n^{-4.02} \\
&\leq n^{-4}
\end{aligned}$$

  Let E_{uv} be the event that the distance matrix does not store the correct distance for the vertex pair (u, v), then we have $\mathbb{P}[E_{uv}] \leq \mathbb{P}[E] \leq 1/n^4$.
  This is because E only corresponds to <u>one</u> of the shortest paths from u to v. For the entry $D[u, v]$ to store incorrect distance, no vertex must have been picked from **any** of the shortest paths from u to v, i.e., $E_{uv} \subseteq E \Rightarrow \mathbb{P}[E_{uv}] \leq \mathbb{P}[E]$.

  Since this error probability is bounded above by n^{-4} for every pair of vertices, and there are at most n^2 pairs of vertices, we obtain by union-bound theorem that the probability the distance matrix stores an erroneous distance is at most

$$\sum_{u \neq v} \mathbb{P}[E_{uv}] \leq \sum_{u \neq v} \mathbb{P}[E] \leq \sum_{u \neq v} \frac{1}{n^4} \leq \frac{1}{n^4} \cdot n^2 = \frac{1}{n^2}$$

  Finally, we arrive at the result that the Monte Carlo Algorithm runs in $\boxed{\mathcal{O}(n^2 \log n)}$ time and stores correct distances for all pairs with probability exceeding $\boxed{1 - \frac{1}{n^{-2}}}$

[4] this specificity comes from the existence of "a" shortest path from u to v in G, *and* arbitrarily picking vertices from any shortest path among them

Q3 Rumour Spreading

There are n people in a city. On day 0, a person comes to know about a rumour. Starting from day 1, each person knowing the rumour picks the phone number of a randomly selected person and calls them to communicate the rumour.

Show that everyone in a town with population n will know the rumor in $\mathcal{O}(\log n)$ days with probability at least $1 - 1/n$.

Conventions and Assumptions

- Let $\mathbb{K}_i$ denote the set of people that know about the rumour on the i^{th} day. Also, at the start of the i^{th} day, let m denote the cardinality of $\mathbb{K}_i$, i.e., $m := |\mathbb{K}_i|$
- Define $\mathbb{D}_i$ in a similar manner as $\mathbb{K}_i$ respectively, but to denote the set of people that *do not know* about the rumour (yet).
- Let f_i be the fraction of people that know about the rumour at the *beginning* of the i^{th} day, i.e., we have $f_i = |\mathbb{K}_i|/n$.
- Let r_i be the fraction of people that do not know about the rumour at the *beginning* of the i^{th} day, i.e., we have $r_i = |\mathbb{D}_i|/n$. We also have $f_i + r_i = 1 \; \forall \; i$
- We will use "people" and "bins" interchangeably. It is expected that it will be understood with the help of surrounding context. Similarly, for "people" (from the set $\mathbb{K}_i$) and "balls".

Preliminaries

Let us consider the start of a general i^{th} day. We interpret the set $\mathbb{K}_i$ as *balls* (there is one ball corresponding to each person that knows about the rumour), and the set $\mathbb{K}_i \cup \mathbb{D}_i$ as *bins* (there is always a bin corresponding to each person). There is one added constraint: "if there is a ball $b \in \mathbb{K}_i$, then it cannot go to the b^{th} bin".

At the end of the day, the non-empty bins (corresponding to the people) in the set $\mathbb{D}_i$ come to know about the rumour and the bins (corresponding to the people) in the set $\mathbb{K}_i$ continue to know about the rumour. The rounds are repeated till every bin has received at least one ball (i.e., every person comes to know about the rumour at least once).

This reduction to a variant of "balls into bins" exactly captures the rumour spreading problem.

We use this reduction to utilize the tools from the "balls into bins" problem. Let α be the label of a person in the set $\mathbb{K}_i \cup \mathbb{D}_i$. Define a Bernoulli RV X_α corresponding to the person α as follows:

$$X_\alpha = \begin{cases} 0, & \text{if person } \alpha \text{ does not hear the rumour from anyone} \\ 1, & \text{if person } \alpha \text{ does hear the rumour from someone else} \end{cases}$$

It is easy to see that all of these random variables are identically distributed for the people of set $\mathbb{D}_i$.

These Bernoulli variables shall be reused again and again each day, so we omit notation such as X^i_α([5]) since it is always clear from the context which day the Bernoulli variables correspond to.

Let the size of $\mathbb{K}_i$ be m. We would like to know the probability distribution of X_α for a person $\alpha \in \mathbb{D}_i$. Specifically, we would like to know the value $\mathbb{P}\left[X^i_\alpha = 0 \mid |\mathbb{K}_i| = m\right]$ (to be written as $\mathbb{P}\left[X_\alpha = 0 \mid |\mathbb{K}_i| = m\right]$ from now on). This value is simply $(1 - \frac{1}{n-1})^m \approx (1 - \frac{1}{n})^m$([6]).
We use the balls and bins analogy to argue for the same, as there are m balls ($|\mathbb{K}_i| = m$), there are $n-1$ bins at disposal for each ball (due to the added constraint) and the balls select a bin available to them randomly uniformly (phone number picked randomly uniformly). Thus, we have,

$$\begin{aligned}
\mathbb{P}\left[X_\alpha = 0 \mid |\mathbb{K}_i| = m\right] &= \left(1 - \frac{1}{n}\right)^m \\
\Leftrightarrow \mathbb{P}\left[X_\alpha = 1 \mid |\mathbb{K}_i| = m\right] &= 1 - \left(1 - \frac{1}{n}\right)^m \\
\Leftrightarrow \mathbb{E}\left[X_\alpha = 1 \mid |\mathbb{K}_i| = m\right] &= 1 \cdot \mathbb{P}\left[X_\alpha = 1 \mid |\mathbb{K}_i| = m\right] = 1 - \left(1 - \frac{1}{n}\right)^m
\end{aligned}$$

Finally, we would like to know the expected size of $\mathbb{K}_{i+1}$ given that the size of $\mathbb{K}_i$ is m. By the definition of the problem, $\mathbb{K}_i \subseteq \mathbb{K}_{i+1}$. The only new additions to the set $\mathbb{K}_{i+1}$ come from the set $\mathbb{D}_i$, essentially the people that come to know about the rumour for the first time on the i^{th} day. Thus, we have the following relation(s):

$$\begin{aligned}
\mathbb{K}_{i+1} &= \mathbb{K}_i \cup \{\alpha \mid \alpha \in \mathbb{D}_i, X^i_\alpha = 1\} \\
\Leftrightarrow |\mathbb{K}_{i+1}| &= |\mathbb{K}_i| + \sum_{\alpha \in \mathbb{D}_i} X_\alpha && (\text{As } \mathbb{K}_i \cap \mathbb{D}_i = \varnothing) \\
&= m + \sum_{\alpha \in \mathbb{D}_i} X_\alpha \\
\mathbb{E}\left[\mathbb{K}_{i+1} \mid |\mathbb{K}_i| = m\right] &= \mathbb{E}\left[m + \sum_{\alpha \in \mathbb{D}_i} X_\alpha\right] \\
&= m + \sum_{\alpha \in \mathbb{D}_i} \mathbb{E}\left[X_\alpha\right] && (\text{Linearity of Expectation}) \\
&\approx m + \sum_{\alpha \in \mathbb{D}_i} \left[1 - \left(1 - \frac{1}{n}\right)^m\right] && (\text{Identically distributed}) \\
&= m + \left(1 - \left(1 - \frac{1}{n}\right)^m\right) \cdot |\mathbb{D}_i|
\end{aligned}$$

[5] This denotes a Bernoulli RV for the person α, whose value can be determined at the *end* of the i^{th} day

[6] The solution assumes n to be large enough in value

$$= m + (n-m) \cdot \left(1 - \left(1 - \frac{1}{n}\right)^m\right) \quad (|\mathbb{K}_i \cup \mathbb{D}_i| = n \text{ and } \mathbb{K}_i \cap \mathbb{D}_i = \varnothing)$$

Remark: Since, $\mathbb{K}_i \subseteq \mathbb{K}_{i+1}$, we have $|\mathbb{K}_i| \leq |\mathbb{K}_{i+1}|$. Also, as a person can call at most one person in a single day, i.e., one person can tell the rumour to at most one new person, we have $|\mathbb{K}_{i+1}| \leq 2|\mathbb{K}_i|$.

Defining the Stages

The whole "experiment" has been split into two stages, based on the number of people that know about the rumour currently, i.e., m. The stages are:

Stage 1. $1 \leq m \leq n/2$

Stage 2. $n/2 \leq m \leq n$

Since $m = 1$ initially, we start from the first stage.

Stage 1 $\quad 1 \leq m \leq n/2$

In this stage, we would like to study how (fast) the fraction of people knowing about the rumour, i.e. f_i, grows.

We simplify the relation obtained in the preliminaries section:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{K}_{i+1} \mid |\mathbb{K}_i| = m\right] &= m + (n-m) \cdot \left(1 - \left(1 - \frac{1}{n}\right)^m\right) \\
&= n - (n-m) \cdot \left(1 - \frac{1}{n}\right)^m \\
&\geq n - (n-m) \cdot \left[1 - \binom{m}{1}\frac{1}{n} + \binom{m}{2}\frac{1}{n^2}\right] && \text{(Proof provided here)} \\
\frac{\mathbb{E}\left[\mathbb{K}_{i+1} \mid |\mathbb{K}_i| = m\right]}{n} &\geq \frac{n}{n} - \frac{n-m}{n} \cdot \left[1 - \binom{m}{1}\frac{1}{n} + \binom{m}{2}\frac{1}{n^2}\right] \\
\mathbb{E}\left[\frac{\mathbb{K}_{i+1}}{n} \mid |\mathbb{K}_i| = m\right] &\geq 1 - (1 - f_i) \cdot \left[1 - f_i + \frac{f_i^2}{2}\right] \\
& && (\because \mathbb{E}\left[c \cdot X\right] = c \cdot \mathbb{E}\left[X\right], \text{ for constant } c) \\
\mathbb{E}\left[f_{i+1} \mid f_i = \frac{m}{n}\right] &\geq f_i \cdot \left(2 - \frac{3f_i}{2} + \frac{f_i^2}{2}\right) \\
&\geq f_i \cdot \left(2 - \frac{3}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2^2}\right) && \text{(Proof provided here)} \\
&= f_i \cdot \left(1 + \frac{3}{8}\right) && (3.0.1)
\end{aligned}$$

Now, we call a day **good** in Stage 1 if the the fraction of people knowing the rumour increases by a factor of at least $(1 + 1/4)$, i.e., if $f_{i+1} \geq f_i \cdot (1 + 1/4)$. Naturally, the "experiment" can only enjoy at most

$$\log_{5/4}(n/2) = (\log_{5/4} e) \cdot \log_e(n/2) = \mathcal{O}(\log n)$$

good days before it shifts to Stage 2. We would thus like to know the probability (or at least a lower bound) of a day in Stage 1 being good. Let that probability be p, i.e., $p \coloneqq \mathbb{P}\left[f_{i+1} \geq f_i \cdot 5/4 \mid f_i = m/n\right]$:

$$\begin{aligned}
f_i \cdot \left(1 + \frac{3}{8}\right) &\leq \mathbb{E}\left[f_{i+1} \mid f_i = m/n\right] && \text{(From 3.0.1)} \\
&= p \cdot \mathbb{E}\left[f_{i+1} \mid f_{i+1} \geq f_i \cdot 5/4,\ f_i = m/n\right] \\
&+ (1-p) \cdot \mathbb{E}\left[f_{i+1} \mid f_{i+1} < f_i \cdot 5/4,\ f_i = m/n\right] \\
&\leq p \cdot \mathbb{E}\left[f_{i+1} \mid f_{i+1} \geq f_i \cdot (1 + 1/4), f_i = m/n\right] + (1-p) \cdot f_i \cdot \frac{5}{4} \\
&\leq p \cdot 2f_i + (1-p) \cdot f_i \cdot \frac{5}{4} && (\because |\mathbb{K}_{i+1}| \leq 2|\mathbb{K}_i|,\ \therefore f_{i+1} \leq 2f_i) \\
\Rightarrow \left(1 + \frac{3}{8}\right) &\leq 2p + (1-p) \cdot \frac{5}{4} \\
\frac{1}{6} &\leq p
\end{aligned}$$

So, a day is **good** with probability at least 1/6, and we need at most $\log_{5/4}(n/2) \leq (5 \log n)$ of them to proceed to Stage 2.

Let the number of days spent in Stage 1 be denoted by $\mathcal{D}_1$. Finally then, we would like to bound $\mathcal{D}_1$ by $(60 \log n)$ with high probability. For the same, we come up with the experiment: "Toss a coin (with Heads' probability $= p_0 = 1/6$) until $(5 \log n)$ Heads have come up". Let the total number of tosses be $\mathcal{T}_1$ in this coin-toss experiment.

We claim that $\mathbb{P}\left[\mathcal{D}_1 > 60 \log n\right] \leq \mathbb{P}\left[\mathcal{T}_1 > 60 \log n\right]$. The argument goes the same as the one provided in lectures, that it is possible for the rumor to spread to at least $n/2$ people before having $\log_{5/4}(n/2)$ good days (and thus $(5 \log n)$ good days). Also, $\mathbb{P}\left[\mathcal{T}_1 > 60 \log n\right]$ is the same as getting less than $(5 \log n)$ heads in $(60 \log n)$ tosses since we would have stopped tossing the coin if we obtained $(5 \log n)$ heads before $(60 \log n)$ coins tosses. Therefore, we have,

$$\begin{aligned}
\mathbb{P}\left[\mathcal{D}_1 > 60 \log n\right] &\leq \mathbb{P}\left[\mathcal{T}_1 > 60 \log n\right] \\
&= \mathbb{P}\left[\text{Less than } (5 \log n) \text{ Heads in } 60 \log n \text{ tosses with } p_0 = p(H) = 1/6\right] \\
&= \mathbb{P}\left[\mathcal{H}_1 < 5 \log n\right] \\
&\leq \mathbb{P}\left[\mathcal{H}_1 \leq 5 \log n\right] \\
&= \mathbb{P}\left[\mathcal{H}_1 \leq \mu/2\right] && (\mu \coloneqq \mathbb{E}\left[\mathcal{H}_1\right] = p_0 \cdot 60 \log n = 10 \log n) \\
&= \mathbb{P}\left[\mathcal{H}_1 \leq (1-\delta)\mu\right] && (\delta = 1/2) \\
&\leq e^{-\delta^2 \mu / 2} && \text{(Chernoff's Bound)} \\
&= e^{-\frac{10 \log n}{8}} = n^{-5/4} \\
&= \frac{1}{n^{5/4}}
\end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow \mathbb{P}\left[\mathcal{D}_1 \leq 60\log n\right] \geq 1 - \frac{1}{n^{5/4}} && (3.1)\\
&\Rightarrow \mathbb{P}\left[\mathcal{D}_1 = \mathcal{O}(\log n)\right] \geq 1 - \frac{1}{n^{5/4}}
\end{aligned}$$

We shall use the second last inequality in the conclusion.

Stage 2 $n/2 \leq m \leq n$

At this point of time, there are too many people who know the rumour. Hence the rumour spreading slows down since people mostly spread the rumour to people who already know the same. The rate at which new people know the rumour slows down. We now shift our focus on the fraction of people who **do not know the rumor**, that is r_i. First, we provide an upper bound on

$$\mathbb{E}\left[|\mathbb{K}_{i+1}|\,|\,|\mathbb{K}_i| = m\right]$$

using the result derived in the preliminaries section:

$$\begin{aligned}
\mathbb{E}\left[|\mathbb{K}_{i+1}|\,|\,|\mathbb{K}_i| = m\right] &= m + (n-m)\left(1 - \left(1 - \frac{1}{n}\right)^m\right)\\
&= n - (n-m)\left(1 - \frac{1}{n}\right)^m\\
&\geq n - (n-m)\left(1 - \frac{1}{n}\right)^{\frac{n}{2}} && (\because m \geq \tfrac{n}{2})\\
&\geq n - \frac{(n-m)}{\sqrt{e}} && (\text{Using } 1 - x \leq e^{-x})\\
\mathbb{E}\left[\frac{|\mathbb{K}_{i+1}|}{n}\middle|\frac{|\mathbb{K}_i|}{n} = f_i\right] &\geq 1 - \frac{(1-f_i)}{\sqrt{e}}\\
1 - \mathbb{E}\left[\frac{|\mathbb{K}_{i+1}|}{n}\middle|\frac{|\mathbb{K}_i|}{n} = f_i\right] &\leq \frac{(1-f_i)}{\sqrt{e}} && (\text{Shifting the 1 to the other side})\\
\mathbb{E}\left[1 - \frac{|\mathbb{K}_{i+1}|}{n}\middle|\frac{|\mathbb{K}_i|}{n} = f_i\right] &\leq \frac{(1-f_i)}{\sqrt{e}} && (\mathbb{E}[X+c] = \mathbb{E}[X] + c, \text{ for constant } c \text{ and r.v } X)\\
\mathbb{E}\left[r_{i+1}\middle|\frac{|\mathbb{D}_i|}{n} = r_i\right] &\leq \frac{r_i}{\sqrt{e}} \approx \frac{r_i}{1.648} && (f_i + r_i = 1)
\end{aligned}$$

Define a day as **bad** if the number of people unknown to the rumour at the end of the day is greater than $\frac{1.5}{\sqrt{e}}$ of the number of people unknown to the rumour at the start of the day. Equivalently, the i^{th} day is bad if $r_{i+1} \geq \frac{1.5 \cdot r_i}{\sqrt{e}}$. A day is **good** otherwise. From Markov's inequality:

$$\mathbb{P}\left[r_{i+1} \geq \frac{1.5 \cdot r_i}{\sqrt{e}}\middle|\frac{|D_i|}{n} = r_i\right] \leq \frac{\mathbb{E}\left[r_{i+1}\middle|\frac{|\mathbb{D}_i|}{n} = r_i\right]}{\frac{1.5 \cdot r_i}{\sqrt{e}}} \leq \frac{2}{3}$$

So, the probability of getting a good day is at least $\frac{1}{3}$. The maximum number of good days possible is $\log_{\left(\frac{\sqrt{e}}{1.5}\right)}\left(\frac{n}{2}\right) \leq 12\log n = \mathcal{O}(\log n)$. Finally, we use Chernoff's bound to get a high probability guarantee on the number of days it takes for everyone to know about the rumour. We proceed in a way similar to Stage 1, by coming up with a coin-toss experiment. However, the notation and constants have been changed. We now use $\mathcal{H}_2, \mathcal{D}_2, \mathcal{T}_2, p_0 = \frac{1}{3}$ and toss the coin until we get $(12\log n)$ Heads, and we argue that the number of tosses, $\mathcal{T}_2$, can be bounded by $(72\log n)$ with high probability.
We omit providing the explanation here to avoid being verbose, since it is the same as that in Stage 1 (except for the specific constants used in the argument).
In $(72\log n)$ tosses, the expected number of heads, i.e., $\mathbb{E}\left[\mathcal{H}_2\right]$ is $p_0 \cdot 72\log n = 24\log n$. The final calculation has been provided below:

$$\begin{aligned}\mathbb{P}\left[H < 12\log n = \left(1 - \frac{1}{2}\right) 24\log n\right] &\leq e^{-\frac{(1/2)^2\cdot(24\log n)}{2}} \\ &= e^{-3\log n} = n^{-3} \\ &= \frac{1}{n^3}\end{aligned}$$

Hence $\mathbb{P}\left[H \geq 12\log n\right] \geq 1 - \frac{1}{n^3}$, that is, we get at least $(12\log n)$ **good days** with high probability, ensuring that everyone comes to know about the rumour in stage 2 with high probability in $(72\log n) = \mathcal{O}(\log n)$ days.

Wrapping Up

Define the following events:

- $\mathcal{E}$: Everyone comes to know about the rumour in $\mathcal{O}(\log n)$ days.
- $\mathcal{E}_1$: At most $60\log n$ days are spent in Stage 1. Obviously then, $\overline{\mathcal{E}_1}$ is the event that (strictly) more than $60\log n$ days are spent in Stage 1.
- $\mathcal{E}_2$: At most $72\log n$ days are spent in Stage 2. Obviously then, $\overline{\mathcal{E}_2}$ is the event that (strictly) more than $72\log n$ days are spent in Stage 2.
- $\mathcal{E}_0$: Everyone comes to know about the rumour in at most $60\log n + 72\log n = 132\log n$ days. Obviously then, $\overline{\mathcal{E}_0}$ is the event that (strictly) more than $132\log n$ days are taken for everybody to know about the rumour. We have the following two relations:

$$\begin{aligned}\mathcal{E}_0 &\subseteq \mathcal{E} \\ \overline{\mathcal{E}_0} &\subseteq \overline{\mathcal{E}_1} \cup \overline{\mathcal{E}_2}\end{aligned}$$

We showed in previous two sections that:

$$\mathbb{P}\left[\mathcal{E}_1\right] \geq 1 - \frac{1}{n^{5/4}} \Leftrightarrow \mathbb{P}\left[\overline{\mathcal{E}_1}\right] \leq \frac{1}{n^{5/4}}$$

$$\mathbb{P}\left[\mathcal{E}_2\right] \geq 1 - \frac{1}{n^3} \Leftrightarrow \mathbb{P}\left[\overline{\mathcal{E}_2}\right] \leq \frac{1}{n^3}$$

Now,

$$\begin{aligned}
\overline{\mathcal{E}_0} &\subseteq \overline{\mathcal{E}_1} \cup \overline{\mathcal{E}_2} \\
\mathbb{P}\left[\overline{\mathcal{E}_0}\right] &\leq \mathbb{P}\left[\overline{\mathcal{E}_1}\right] + \mathbb{P}\left[\overline{\mathcal{E}_2}\right] \\
&\leq \frac{1}{n^{5/4}} + \frac{1}{n^3} \leq \frac{2}{n^{5/4}} \\
&\leq \frac{1}{n} && \text{(Asymptotic analysis in } n\text{)} \\
\mathbb{P}\left[\overline{\mathcal{E}_0}\right] &\leq \frac{1}{n} \\
\Leftrightarrow \mathbb{P}\left[\mathcal{E}_0\right] &\geq 1 - \frac{1}{n}
\end{aligned}$$

Also, since, $\mathcal{E}_0 \subseteq \mathcal{E}$, we have,

$$\begin{aligned}
\mathbb{P}\left[\mathcal{E}\right] &\geq \mathbb{P}\left[\mathcal{E}_0\right] \geq 1 - \frac{1}{n} \\
\Rightarrow &\boxed{\mathbb{P}\left[\mathcal{E}\right] \geq 1 - \frac{1}{n}}
\end{aligned}$$

Thus, we arrive at the required result that everyone in the town comes to know about the rumour in $\mathcal{O}(\log n)$ days with probability exceeding $\boxed{1 - 1/n}$

Auxiliary Proofs

Theorem: $(1 - 1/n)^m \leq [1 - m/n + m^2/(2n^2)]$ in Stage 1

Proof: We first show that $\binom{m}{i}\frac{1}{n^i} \geq \binom{m}{i+1}\frac{1}{n^{i+1}}$ by taking their ratio.
Note that the relation trivially holds for all $i \geq m$ since $j > m \Rightarrow \binom{m}{j} = 0$;
A proof is needed only when both the terms are non-zero, i.e., $0 \leq i < m$:

$$\begin{aligned}
\frac{\binom{m}{i} \cdot \frac{1}{n^i}}{\binom{m}{i+1} \cdot \frac{1}{n^{i+1}}} &= \frac{\binom{m}{i} \cdot n}{\binom{m}{i+1}} \\
&= \frac{n \cdot (i+1)}{(m-i)} \\
&= \frac{(m-i) - (m-i) + n(i+1)}{m-i} \\
&= 1 + \frac{i \cdot (n+1) + (n-m)}{m-i} \\
&\geq 1 && (m \leq n/2 \leq n)
\end{aligned}$$

$$\Leftrightarrow \binom{m}{i} \cdot \frac{1}{n^i} \geq \binom{m}{i+1} \cdot \frac{1}{n^{i+1}}$$

$$\Leftrightarrow \delta_i \coloneqq \binom{m}{i} \cdot \frac{1}{n^i} - \binom{m}{i+1} \cdot \frac{1}{n^{i+1}} \geq 0$$

Now, we expand $(1 - 1/n)^m$ using binomial expansion:

$$\begin{aligned} \left(1 - \frac{1}{n}\right)^m &= \sum_{i=0}^{i=m} (-1)^i \cdot \binom{m}{i} \cdot \frac{1}{n^i} \\ &= 1 - \binom{m}{1}\frac{1}{n} + \binom{m}{2}\frac{1}{n^2} - \sum_{j\geq 1} \delta_{2j+1} && (^7) \\ &\leq 1 - \binom{m}{1}\frac{1}{n} + \binom{m}{2}\frac{1}{n^2} && (\because \delta_i \geq 0 \; \forall \; i) \end{aligned}$$

Theorem: $2 - 3f_i/2 + f_i^2/2 \geq 11/8$ for i in Stage 1

Proof: Since the day is in Stage 1, we have $1 \leq m \leq n/2 \Rightarrow 0 \leq f_i \leq 1/2$.

We first show that the function $f(x) = x^2/2 - 3x/2 + 2$ is decreasing in the interval $x \in [0, 1/2]$ (both ends inclusive). We differentiate $f(x)$ with respect to x for the same to obtain $f'(x) = x - 3/2 + 0 < 0 \; \forall \; x \in [0, 1/2]$. Thus, we have $f(x) \geq f(1/2) \; \forall \; x \in [0, 1/2]$. Since $f(1/2) = 11/8$, we get,

$$\boxed{f(x) \geq (1 + 3/8)} = 11/8 = f(1/2)$$

[7] As the series converges absolutely; towards $(1 + 1/n)^m \leq (1 + 1/n)^n < \infty$