CS203: Probability in Computer Science
Assignment 3 Solutions

Yatharth Goswami
Roll No. 191178

April 25, 2021

I Problem 1 Solution

In this problem we were asked to show two things. One was to show that the least number of days to spread the fake news is $log_2(n)$ and other was to find the expected number of days for the rumour to spread under some heuristic.

(a) Notice that the least amount of taken to spread the rumour will be taken when each day every informed person calls an uninformed person. If this happens to be the case, the number of informed people $I(d)$ at any day d will follow the recursive equation

$$I(d+1) = 2 \cdot I(d)$$

Solving the above equation and putting $I(0) = 1$, we get

$$I(d) = 2^d$$

Therefore the time $I(d) = n$ is when $2^d = n$ or after $d = \lceil log_2 n \rceil$.

(b) In this part, we will calculate the expected number of days to spread the rumour. Since we were asked to provide a heuristic answer so most of the arguments might not be totally concrete but I will try to provide the logic behind everything I will be doing. We will try to find the expected number of days to complete broadcasting by finding the expected number of people knowing the information after each day and then trying to find the day when they become close to n.

First of all notice the fact that for the first initial few days, the number of informed people will double with high probability, hence upto a certain ratio of informed people say r, we can safely model the problem as number of informed people being doubled the next day. Therefore, we now divide the problem into essentially two phases. Phase 1 is from the ratio of informed people increasing from $r_0 = 1/n$ to r and the second phase from r to r_f where r_f is a ratio sufficiently close to 1.

For phase 1 the expected time is the time taken for the number of informed people to reach to nr starting with 1 with the number doubling every day. Therefore this time(say t_1) comes out to be $t_1 = log_2(nr)$. Now, after phase one ends we need to find an expression for how the number of informed people will change. For this consider the following modelling. Let X be the random variable denoting the number of new persons being called on a certain day.

$$X_i \rightarrow \begin{cases} 1 & \text{if } i^{th} \text{informed node calls an uninformed person} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, we can see that

$$X = \sum_{i=1}^{I(d)} X_i \quad \text{where } I(d) \text{ denotes the number of informed people on day } d$$

By linearity of expectation, we have

$$E[X] = \sum_{i=1}^{I(d)} E[X_i]$$

Now $E[X_i] = P(X_i = 1)$. Now, this is a difficult expression to find out since this probability will depend on what are the values of other X_is. The exact expression can be found using inclusion-exclusion principle. But for estimating answer using a heuristic, we will take this value to be simply $1 - I(d)/n$ by ignoring the higher level terms. Therefore

$$E[X] = I(d) \cdot (1 - I(d)/n)$$

Now, let the ratio of informed population be denoted by α, therefore using the above equation we get

$$\Delta\alpha = \alpha(1-\alpha)$$

There is another interesting way to come up with same expression using a slightly less rigorous modelling of the problem. The number of calls that will be made by people informed will be αn, but since α fraction already knows the information the new people who will know the information will be $\alpha n(1-\alpha)$ (heuristically). Hence the next day change in ratio of people knowing the info will be i.e $\Delta\alpha = \alpha(1-\alpha)$.

Considering a day to be one time step we can write the above equation as

$$\Delta\alpha = \alpha(1-\alpha)\Delta t$$
$$\frac{\Delta\alpha}{\Delta t} = \alpha(1-\alpha)$$

We can convert the above from discrete time model to continous time model with one day modelled as dt and therefore we get

$$
\begin{aligned}
&\frac{d\alpha}{dt} = \alpha(1-\alpha) \\
&\frac{d\alpha}{\alpha(1-\alpha)} = dt \\
&\int_{r}^{r_f} \frac{d\alpha}{\alpha(1-\alpha)} = \int_{t_1}^{t_f} dt \\
&\int_{r}^{r_f} \frac{d\alpha}{\alpha} + \int_{r}^{r_f} \frac{d\alpha}{1-\alpha} = \int_{t_1}^{t_f} dt \\
&\ln(r_f) - \ln(r) + \ln(1-r) - \ln(1-r_f) = t_f - t_1
\end{aligned}
$$

We want to find the time when the number of people knowing the rumour will be greater than $n-1$. Therefore we need $r_f > 1 - 1/n$. For the purpose of finding the first time this happens we will take $r_f = 1 - 1/n$ and solve the final equation. For large n, we have that $1 - 1/n$ tends to 1 and hence $\ln(r_f)$ can be neglected all together. Therefore, we get

$$
\begin{aligned}
t_f &= t_1 - \ln(r) + \ln(1-r) - \ln\left(\frac{1}{n}\right) \\
t_f &= \log_2(nr) - \ln(r) + \ln(1-r) + \ln(n)
\end{aligned}
$$

This implies

$$
\boxed{t_f = \log_2(n) + \ln(n) + (\text{some constant terms})}
$$

Therefore we end up completing broadcasting in logarithmic time overall.

II Problem 2 Solution

The problem demands us to find two things. First is the probability that the drunkard starting from i^{th} milestone reaches home and other the expected number of steps in which the walk will stop starting from i^{th} milestone.

(a) Let $p(i)$ denote the required probability of reaching home starting at i^{th} milestone. It is trivial to note that

$$p(0) = 1$$
$$p(n) = 0$$

Now, notice that a person being at i^{th} ($i \neq 0$ or n) location currently can move to either $i-1$ or $i+1$ with equal probability. Hence, we obtain the recursive equation for $p(i)$.

$$p(i) = \frac{1}{2} \cdot p(i-1) + \frac{1}{2} \cdot p(i+1) \ \forall \ i \in [1, n-1]$$

Now for any range $[l, m]$ which is a subset of $[1, n-1]$ we have

$$\begin{aligned} \sum_{i=l}^{m} p(i) &= \sum_{i=l}^{m} \frac{1}{2} \cdot p(i-1) + \sum_{i=l}^{m} \frac{1}{2} \cdot p(i+1) \\ p(l) + p(m) + \sum_{i=l+1}^{m-1} p(i) &= \frac{1}{2} \cdot \big(p(l-1) + p(l) + p(m) + p(m+1)\big) + \sum_{i=l+1}^{m-1} p(i) \\ \implies p(l) + p(m) &= p(l-1) + p(m+1) \end{aligned}$$

Plugging $m = n-1$ in previous equation, we get

$$\begin{aligned} p(l) + p(n-1) &= p(l-1) + p(n) \\ \implies p(l-1) - p(l) &= p(n-1) \ \forall \ l \in [1, n-1] \end{aligned}$$

Using the above equation we get that the sequence $p(i)$ is in arithmetic progression with common difference as $-p(n-1)$. Hence, we obtain

$$\begin{aligned} p(i) &= (i+1)^{th} \text{term of the sequence} \\ &= p(0) + (i) \cdot (-p(n-1)) \\ &= 1 - i \cdot p(n-1) \ \forall \ i \in [1, n-1] \end{aligned}$$

Plugging $i = n-1$ in above equation yields

$$p(n-1) = 1/n$$

Hence, we get

$$\boxed{p(i) = 1 - \frac{i}{n} \;\forall\; i \in [1, n-1]}$$

Hence, for $n = 4$ we have the following values

$$\begin{aligned}
p(0) &= 1 \\
p(1) &= 3/4 \\
p(2) &= 1/2 \\
p(3) &= 1/4 \\
p(4) &= 0
\end{aligned}$$

(b) Let X_i denote the number of steps taken to terminate the walk starting from i^{th} milestone. Let $e(i)$ denote the expected value of random variable X_i. It is trivial to see that

$$\begin{aligned}
e(0) &= 0 \\
e(n) &= 0
\end{aligned}$$

Now, notice that a person being at i^{th} ($i \neq 0$ or n) location currently can move to either $i-1$ or $i+1$ with equal probability. Hence, we obtain the recursive equation for $P(X_i = k)$ (where $P(\cdot)$ denote the probability of an event).

$$P(X_i = k) = \frac{1}{2} \cdot P(X_{i-1} = k-1) + \frac{1}{2} \cdot P(X_{i+1} = k-1)$$

Multiplying with k and taking sum for all ks we can find the expectation of X_i.

$$\begin{aligned}
\sum_{k=0}^{\infty} P(X_i = k) \cdot k &= \sum_{k=0}^{\infty} \frac{1}{2} \cdot \big(P(X_{i-1} = k-1) \cdot k + P(X_{i+1} = k-1) \cdot k\big) \\
e(i) &= \Big(\sum_{k=0}^{\infty} \frac{1}{2} \cdot \big(P(X_{i-1} = k-1) \cdot (k-1) + P(X_{i+1} = k-1) \cdot (k-1)\Big) + \\
&\quad \Big(\sum_{k=0}^{\infty} \frac{1}{2} \cdot \big(P(X_{i-1} = k-1) \cdot 1 + P(X_{i+1} = k-1) \cdot 1\big)\Big) \\
e(i) &= \frac{1}{2} \cdot (e(i-1) + e(i+1)) + \Big(\frac{1}{2} + \frac{1}{2}\Big) \\
e(i) &= \frac{1}{2} \cdot (e(i-1) + e(i+1)) + 1 \;\forall\; i \in [1, n-1]
\end{aligned}$$

Now for any range $[l, m]$ which is a subset of $[1, n-1]$ we have

$$\sum_{i=l}^{m} e(i) = \Big(\sum_{i=l}^{m} \frac{1}{2} \cdot e(i-1) + \sum_{i=l}^{m} \frac{1}{2} \cdot e(i+1)\Big) + (m-l+1)$$

$$e(l) + e(m) + \sum_{i=l+1}^{m-1} e(i) = \Big(\frac{1}{2} \cdot \big(e(l-1) + e(l) + e(m) + e(m+1)\big) + \sum_{i=l+1}^{m-1} e(i)\Big) + (m-l+1)$$

$$\implies e(l) + e(m) = e(l-1) + e(m+1) + 2(m-l+1)$$

Plugging $m = n-1$ in above equation

$$e(l) = e(l-1) - e(n-1) + 2(n-l) \ \forall \ l \in [1, n-1]$$

Summing up the above equation from $l = 1$ to $l = i$ gives

$$e(i) = e(0) + \sum_{l=1}^{i} 2(n-l) - i \cdot e(n-1)$$

$$e(i) = 2\Big(i \cdot n - \frac{i(i+1)}{2}\Big) - i \cdot e(n-1)$$

$$e(i) = 2\Big(i \cdot n - \frac{i(i+1)}{2}\Big) - i \cdot e(n-1)$$

Plugging $i = n-1$ in above equation gives

$$e(n-1) = 2\Big((n-1) \cdot n - \frac{(n-1)(n)}{2}\Big) - (n-1) \cdot e(n-1)$$

$$e(n-1) = n-1$$

Putting this in original equation gives

$$e(i) = 2\Big(i \cdot n - \frac{i(i+1)}{2}\Big) - i \cdot (n-1)$$

$$\boxed{e(i) = i(n-i)}$$

Hence, for $n = 4$ we have the following values

$$e(0) = 0$$
$$e(1) = 3$$
$$e(2) = 4$$
$$e(3) = 3$$
$$e(4) = 0$$

III Problem 3 Solution

In this problem we were asked to determine the surprise function. I will list the properties of the function first.

- $S(1) = 0$,
- $p > q \implies S(p) < S(q)$,
- S is a continous function defined in the domain $[0,1]$,
- $S(pq) = S(p) + S(q)$.

The last axiom about the surprise function can be justified by taking two independent events x and y and now thinking about the surprise of seeing x followed by y. Intuitively, this should be sum of individual surprises. You can convince yourself by thinking about two totally unrelated events for eg "The sun is a hot star" and "He owns a cat as a pet". Because they are unrelated so the surprise obtained by hearing them one after the other should simply be the sum of individual surprises.

Now, turning to the case of dependent events(say A and B). Now, we need to find a nice interpretation of pq in the case of dependent events (Conditional probability!). So, if we choose two events such that $P(A|B) = p$ and $P(B) = q$. Then $P(A \cap B) = pq$. And therefore pq has a nice interpretation now. Note that in order to justify the equation intuitively, you need to find a nice interpretation and apparently choosing dependent events with $P(A) = p$ and $P(B) = q$, doesn't lead you to find a nice interpretation for pq. But the conditional one does. Rest is mostly the same. Since, the event A and B represents the probability of A and B occuring one after the other. It's surprise can be broken into surprise we got by B and surprise we got by A after seeing B, which is precisely $S(A|B) + S(B) = S(p) + S(q)$

For the next part we will solve the functional equation and find the surprise function.

$$S(pq) = S(p) + S(q)$$

Let us define another function $f(p) = S(2^p)$ for $p \in (-\infty, 0]$. Then

$$\begin{aligned} f(p+q) &= S(2^{p+q}) \\ &= S(2^p 2^q) \\ &= S(2^p) + S(2^q) \\ &= f(p) + f(q) \end{aligned}$$

Claim: If S is continous, so is f.

Proof. From the above result, we have

$$f(p+h) = f(p) + f(h)$$
$$f(p+h) - f(p) = f(h)$$

Therefore, we need to show that limit as $h \to 0^-$, $f(h)/h$ exists, or $S(2^h)/h$ exits.

$$\lim_{h\to 0^-} \frac{S(2^h)}{h} = \lim_{h\to 0^-} \frac{S(1-h(\ln 2))}{h}$$
$$= \lim_{x\to 0^-} \ln 2 \cdot \frac{S(1-x)-S(1)}{x} \quad \text{where } x = h\ln 2$$

Now since S is continous at 1 the above limit exists and hence f is continous. □

Now, it is a standard problem to solve for the functional equation $f(p+q) = f(p)+f(q)$ given f is continous. It is called the *Cauchy's equation*[1] and it's solution is given by $f(x) = cx$ when f is continous.

Therefore plugging in $x = \log_2(p)$ in the above equation gives

$$f(\log_2(p)) = c\log_2(p)$$
$$S(2^{\log_2(p)}) = c\log_2(p)$$
$$S(p) = c\log_2(p)$$

Plugging in $p = 1/2$ in above equation, we get

$$S(1/2) = c\log_2(1/2)$$
$$c = -S(1/2)$$

Therefore we get

$$\boxed{S(p) = S(1/2)\cdot\log_2(1/p)}$$

Now using the second property, since $S(1) = 0$ and $S(p) > S(1) = 0\ \forall\ p \in [0,1)$ so we have

$$S(p) \geq 0\ \forall\ p \in [0,1]$$
$$S(1/2).\log_2(1/p) \geq 0\ \forall\ p \in [0,1]$$

IV Problem 4 solution

In this problem, we were asked to prove that if every vertex of a bipartite graph is given coloring from $> \log_2(n)$ colors, there will exist a scheme such that no adjacent vertices will receive the same color.

Since the given graph is bipartite in nature, we can divide the vertices into two sets, say R_0 and R_1. Due to properties of bipartiteness there are no edges inside the two regions. Let $C(V)$ denote the set containing the list of colors available to vertex V. Let U be the union of all these sets, i.e. $U = \bigcup_{i=1}^{n} C(V)$. Let us consider random colorings on the graph given by $f : U \to \{0,1\}$ such that $f^{-1}(1)$ and $f^{-1}(0)$ denote the list of colors available for regions R_1 and R_0 respectively.

For showing the claim it suffices to show that for all vertices $v \in R_i$

$$C(v) \cap f^{-1}(i) \neq \phi \;\forall\; i \in \{0,1\}$$

The above statement is true since f is a function and hence $f^{-1}(0) \cap f^{-1}(1) = \phi$, so if we give the vertex $v \in R_i$ one of the color from the set $S(v) \cap f^{-1}(i)$ and since the edges are only present between R_0 and R_1, so this condition along with $f^{-1}(0) \cap f^{-1}(1) = \phi$ implies that such a coloring will not have an edge with two vertices of same color being connected to each other.

Now, let's try to find out the probability of finding such a coloring f if we chose it randomly. Since f is chosen randomly so we have for any $c \in U$

$$P(f(c) = 0) = \frac{1}{2}$$

We can find the required probability using the following equation

$$P(\forall v \in R_i,\ C(v) \cap f^{-1}(i) \neq \phi) \;=\; 1 - P(\exists v \in R_i,\ C(v) \cap f^{-1}(i) = \phi)$$

Now we bound the $P(\exists v \in R_i, C(v) \cap f^{-1}(i) = \phi)$ using union bound on all vertices of the graph. Hence, we get

$$\begin{aligned}
P(\exists v \in R_i,\ C(v) \cap f^{-1}(i) = \phi) &\leq \sum_i \sum_{v \in V_i} P(C(v) \cap f^{-1}(i) = \phi) \\
&\leq \sum_i \sum_{v \in V_i} \left(\frac{1}{2}\right)^{|C(v)|} \\
&\leq \sum_i \sum_{v \in V_i} \frac{1}{n} \\
&< 1
\end{aligned}$$

The value of $P(C(v) \cap f^{-1}(i) = \phi)$ can be calculated by using the fact that for the colors in $C(v)$, f should give the value complement of the region in which v is present and is free to give any value to other colors not in $C(v)$. This fixed value of f for $|C(v)|$ values and hence the required probability.

Using the above analysis we obtain

$$\begin{aligned} P(\forall v \in R_i,\ C(v) \cap f^{-1}(i) \neq \phi) &= 1 - P(\exists v \in R_i,\ C(v) \cap f^{-1}(i) = \phi) \\ &> 1 - 1 \\ &> 0 \end{aligned}$$

This means that there exists a finite probability of existence of a good coloring function f that will lead to no two adjaccent vertices getting the same color if $|C(v)| > \log_2(n)$ for any bipartite graph.

References

[1] Cauchy's functional equation. https://en.wikipedia.org/wiki/Cauchy%27s_functional_equation.