CSC4120 Spring 2024 - Written Homework 8

Yohandi 120040025
Andrew Nathanael 120040007

April 10, 2024

Problem 1.

Let $G = (V, E)$ be an undirected graph and assume that all its edge weights are distinct. Prove that G has a unique minimum spanning tree.

Assuming the minimum spanning tree of G is not unique, then there exist two different minimum spanning trees of G: T_1 and T_2. Let e_1 be one of the edges that exist in T_1 but not in T_2 (there always exists such an edge as T_1 and T_2 are not identical); consequently, T_2 must include at least one edge e_2 that is not in T_1. Let's assume without loss of generality that the weight of e_1 is less than the weight of e_2. Note that their weights can't be equal due to all weights being distinct.

Adding e_1 to T_2 creates a cycle because T_2 is a spanning tree and spans all vertices in G. This cycle must contain e_2 since removing e_2 from this cycle (after adding e_1) still spans all vertices in G, creating another spanning tree. Let's denote this new tree as T_2'. By our assumption, e_1 has a smaller weight than e_2. Therefore, the total weight of T_2' is less than the total weight of T_2, contradicting the assumption that T_2 is a minimum spanning tree of G. Hence, if all edge weights in an undirected graph G are distinct, G has a unique minimum spanning tree.

Problem 2.

You are given a weighted graph $G = (V, E)$ where all edge weights are positive and distinct, and a starting node s. Bob claims that it is possible for a tree of shortest paths from s and a minimum spanning tree in G not to share any edges. If it is true, give an example. If not, give a proof.

Suppose in the minimum spanning tree of G, s is connected to t (there always exists such t as the degree of every nodes in a tree is at least one), then for the edge e that connects s and t, i.e., $e = (s, t)$, e must be passed through by s to visit t in its shortest path, implying that both tree of shortest paths from s and the minimum spanning tree in G share at least an edge, which is e.

If s does not pass e in its shortest path to t, then there exists a sequence of edges $e_1, \ldots, e_k$ such that $e_1 = (s, x_1), e_2 = (x_1, x_2), \ldots, e_k = (x_{k-1}, t)$ and $\sum_{i=1}^{k}$ weight of $e_i \leq$ weight of e, implying that the MST of G will have an alternative to connect s and t with cheaper cost without using e (due to all weights being positive), contradicting to the supposition that s and t is connected in the minimum spanning tree of G.

Problem 3.

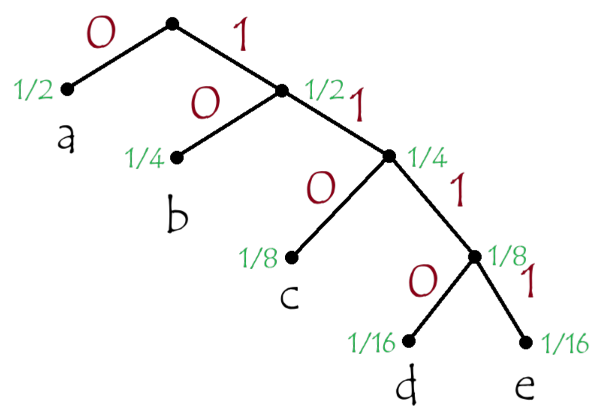
Suppose the symbols a, b, c, d, e occur with frequencies $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{16}$, respectively.

(a) What is the Huffman encoding of the alphabet?

(b) If this encoding is applied to a file consisting of 1,000,000 characters with the given frequencies, what is the length of the encoded file in bits?

(a) The Huffman encoding of the alphabet is constructed as follows:

- Combine d and e into a new node with frequency $\frac{1}{16} + \frac{1}{16} = \frac{1}{8}$.
- Combine c and de into a new node with frequency $\frac{1}{8} + \frac{1}{8} = \frac{1}{4}$.
- Combine b and cde into a new node with frequency $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$.
- Combine a and $bcde$ to form the root.

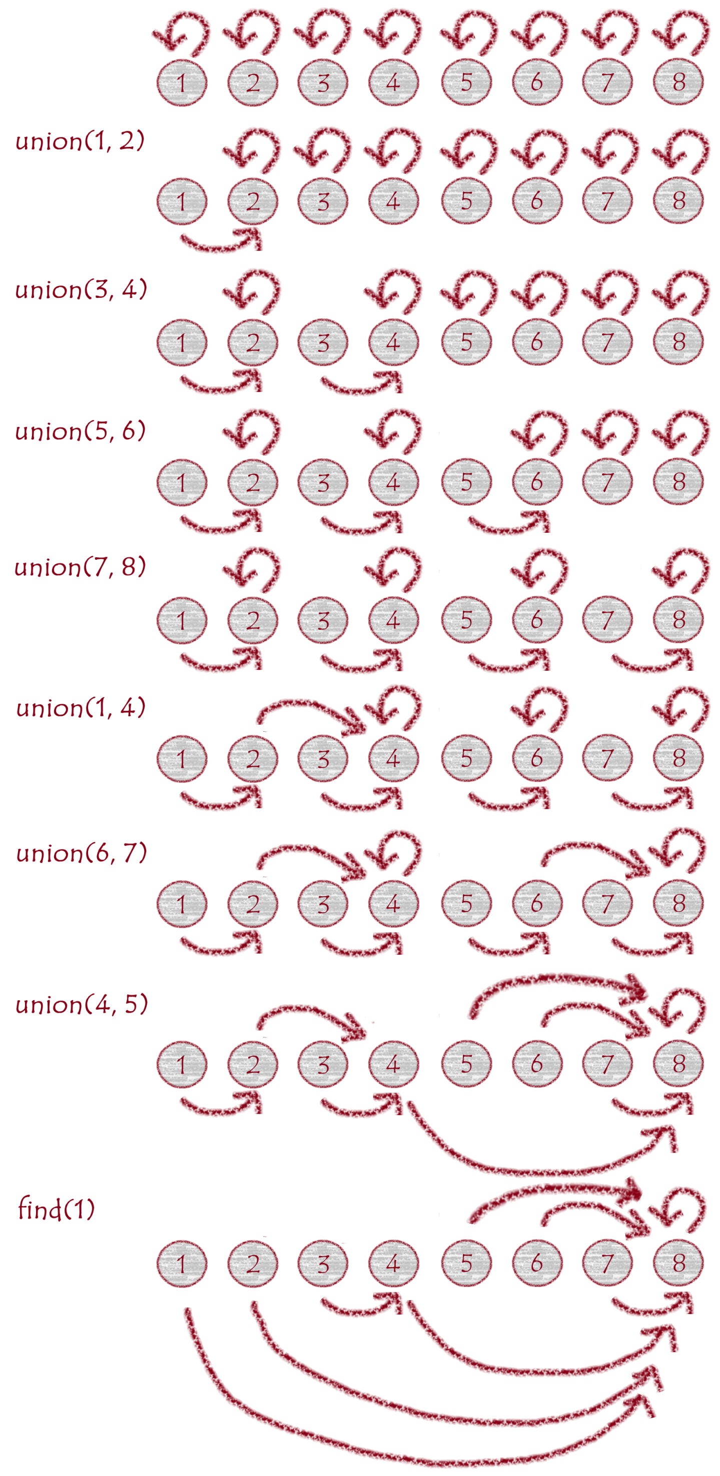


(b)
- The character a occurs $\frac{1}{2} \times 1\,000\,000 = 500\,000$ times with 1 bit encoding.
- The character b occurs $\frac{1}{4} \times 1\,000\,000 = 250\,000$ times with 2 bits encoding.
- The character c occurs $\frac{1}{8} \times 1\,000\,000 = 125\,000$ times with 3 bits encoding.
- The character d occurs $\frac{1}{16} \times 1\,000\,000 = 62\,500$ times with 4 bits encoding.
- The character e occurs $\frac{1}{16} \times 1\,000\,000 = 62\,500$ times with 4 bits encoding.

Then, the length of the encoded file is $500\,000 \times 1 + 250\,000 \times 2 + 125\,000 \times 3 + 62\,500 \times 4 + 62\,500 \times 4 = 1\,875\,000$ bits.

Problem 4.

Give the state of the disjoint-sets data structure after the following sequence of operations, starting from singleton sets $\{1\}, \ldots, \{8\}$. Use path compression. In case of ties, always make the lower numbered root point to the higher numbered one.
union$(1, 2)$, union$(3, 4)$, union$(5, 6)$, union$(7, 8)$, union$(1, 4)$, union$(6, 7)$, union$(4, 5)$, find(1)



Problem 5.

Consider a distribution over n possible outcomes, with probabilities $p_1, p_2, \ldots, p_n$.

(a) Just for this part of the problem, assume that each p_i is a power of 2 (that is, of the form $1/2^k$). Suppose a long sequence of m samples is drawn from the distribution and that for all $1 \le i \le n$, the i-th outcome occurs exactly mp_i times in the sequence. Show that if Huffman encoding is applied to this sequence, the resulting encoding will have length

$$\sum_{i=1}^{n} mp_i \log(\frac{1}{p_i})$$

(b) Now consider arbitrary distributions – that is, the probabilities p_i are not restricted to powers of 2. The most commonly used measure of the *amount of randomness* in the distribution is the *entropy*

$$\sum_{i=1}^{n} p_i \log(\frac{1}{p_i})$$

For what distribution (over n outcomes) is the entropy the largest possible? The smallest possible?

(a) We have $\sum_{i=1}^{n} p_i = 1$ due to there are exactly n possible outcomes. As $p_i = \frac{1}{2^{k_i}}$ and k_i is non-negative ,$\forall i$, then by binary fraction manner, there exists $i \neq j$ and $i, j \in \{1, \dots, n\}$ such that $p_i = p_j \le p_k, k \in \{1, \dots, n\}$ and $k \neq i, j$. Huffman encoding will greedily combine p_i and p_j to a new node $2 * p_i$. We introduce (i, j) as a new outcome and remove both i and j outcomes from our original distribution, i.e., the distribution now has $n-1$ possible outcomes, with probabilities:

$$p_1, \dots, p_{\min(i,j)-1}, p_{\min(i,j)+1}, \dots, p_{\max(i,j)-1}, p_{\max(i,j)+1}, \dots, p_n, p_{ij}$$

Since p_{ij} is also a power of 2, the same procedure can also be made recursively until $n = 1$ (induction), where $p_{\text{all combined}} = 1$.

Then, in the application of Huffman encoding to the original sequence, an outcome i with probability p_i will be stored using $\log_2(\frac{1}{p_i})$ bits as p_i is combined to $\underbrace{2p_i, 4p_i, \dots, \frac{1}{2}, 1}_{\log_2(\frac{1}{p_i}) \text{ times}}$.

The resulting encoding will have length:

$$\sum_{i=1}^{n} \text{frequency}_i \times \text{encoding length}_i = \sum_{i=1}^{n} mp_i \log_2(\frac{1}{p_i})$$

of bits.

(b) Our objective is given as follows:

$$\begin{aligned}
\max \quad & \sum_{i=1}^{n} p_i \log_2(\frac{1}{p_i}) = -\sum_{i=1}^{n} p_i \log_2(p_i) \\
\text{s.t.} \quad & \sum_{i=1}^{n} p_i = 1
\end{aligned}$$

Let λ be a Lagrange multiplier and $\mathcal{L}$ be a Lagrange function, then:

$$\mathcal{L}(\mathbf{p}, \lambda) = -\sum_{i=1}^{n} p_i \log_2(p_i) + \lambda \underbrace{\left(\sum_{i=1}^{n} p_i - 1\right)}_{\text{constraint}}$$

- $\dfrac{\partial \mathcal{L}}{\partial p_i} = -\dfrac{\ln(p_i) + 1}{\ln(2)} + \lambda = 0 \Rightarrow p_i = 2^{\lambda - \frac{1}{\ln(2)}}$
- $\dfrac{\partial \mathcal{L}}{\partial \lambda} = \sum_{i=1}^{n} p_i - 1 = 0 \Rightarrow \sum_{i=1}^{n} p_i = 1$

As $p_i = p_j, \forall i \neq j$, then $p_i = \frac{1}{n}$ (uniform distribution). Moreover, $\frac{\partial^2 \mathcal{L}}{(\partial p_i)^2} = -\frac{1}{p_i \ln(2)} < 0$, which further implying that the entropy function is concave; hence, uniform distribution is the largest possible.

Consider an event with 1 outcome, i.e., $\mathbf{p} = [1, \underbrace{0, 0, \ldots, 0}_{n-1}]$. Then,

$$\begin{aligned}
\sum_{i=1}^{n} p_i \log_2(\frac{1}{p_i}) &= -\sum_{i=1}^{n} p_i \log_2(p_i) \\
&= -(1 \log_2(1) + \underbrace{0 \log_2(0) + \ldots + 0 \log_2(0)}_{n-1}) \\
&= 0 \quad (\text{due to } \lim_{x->0} x \log_2(x) = 0)
\end{aligned}$$

Then, the distribution with only 1 possible outcome is the smallest possible. Note that since $0 \leq p_i \leq 1, \forall i$, then $\log(\frac{1}{p_i}) \geq 0$, which further implies that $\sum_{i=1}^{n} p_i \log(\frac{1}{p_i}) \geq 0$. This means, it is not possible to achieve negative entropy value; hence, 0 is the smallest possible.