Problem Set 5 Solutions

Exercise 5-1.  Do Exercise 24.4-6 on p. 670 in CLRS.

Exercise 5-2.  Do Exercise 25.3-3 on p. 705 in CLRS.

Exercise 5-3.  Do Exercise 25.3-4 on p. 705 in CLRS.
**Problem 5-1. Sampling s-t cuts**

Consider a network of nodes connected by links that can be switched on or off. We can model the network with a connected undirected graph $G = (V, E)$. Two nodes $s, t \in V$ in the network must be kept disconnected from each other at all times by “switching off” a set of links such that there is no path from $s$ to $t$ that consists of only links that remain on. Each link has an upper bound on the rate that it can be kept off, however. Specifically, let $p : E \to [0, 1]$ be an edge-weight function such that $p(e)$ is an upper bound on the frequency with which link $e$ can be turned off. Consequently, different sets of links must take turns switching off so that the overall fraction of times an individual link $e$ spends switched off does not exceed $p(e)$ with high probability.

To solve this problem, we shall design a randomized algorithm that repeatedly samples which sets of links to switch off. That is, it produces a sequence $G_1, G_2, \ldots, G_T$ of graphs in which $s$ is disconnected from $t$, where for $i = 1, 2, \ldots, T$, we have $G_i = (V, E_i)$ where $E_i \subseteq E$.

For two distinguished vertices $s, t \in V$, define an **s-t cut** of $G$ to be a partition $(C, V - C)$ of $V$ such that $s \in C$ and $t \in V - C$. We say that an edge $(u, v) \in E$ **crosses** the cut $(C, V - C)$ if $|\{u, v\} \cap C| = 1$. Define $\chi(C) \subseteq E$ to be the set of edges that cross the cut $C$.

**(a)** For an s-t cut $(C, V - C)$, show that $s$ and $t$ are disconnected in the graph $G = (V, E - \chi(C))$.

**Solution:** Suppose there is a path $Path$ from $s$ to $t$ in $G$. There must exist at least one edge $(u, v) \in Path$ such that $u \in C$ and $v \in V - C$, otherwise all the nodes in $Path$, including $s$ and $t$, would all be in $C$ or all in $V - C$. This edge crosses the cut $(C, V - C)$ and does not belong to $E - \chi(C)$, a contradiction.

We say that a distribution over s-t cuts satisfies the **bounded disconnect (BD)** property if for every $e \in E$, when we sample a cut from the distribution, the probability $e$ crosses the cut is at most $p(e)$.

Let $\delta : V \times V \to \mathbb{R}^+$ be the shortest-path distance function between two vertices in $G$, where each edge $e \in E$ has edge weight $p(e)$. Define $f : V \to [0, 1]$ by $f(v) = \min \{\delta(s, v), 1\}$.

**(b)** Prove the following statements:

1. $|f(u) - f(v)| \leq p(e)$ for every edge $e = (u, v) \in E$.

**Solution:** According to the triangle inequality for the shortest-path distance, we have $\delta(s, u) \leq \delta(s, v) + p(e)$. This follows because the right-hand side is the distance of a path that goes from $s$ to $v$ to $u$, which by definition must be greater than or equal to $\delta(s, u)$, the shortest-path distance from $s$ to $u$. By symmetry, we have $\delta(s, v) \leq \delta(s, u) + p(e)$. Moving the $\delta$’s to one side, we get

$$
\delta(s, u) - \delta(s, v) \leq p(e)
$$

$$
\delta(s, v) - \delta(s, u) \leq p(e)
$$

This is the same as saying $|\delta(s, u) - \delta(s, v)| \leq p(e)$. Next we show that $|f(u) - f(v)| \leq |\delta(s, u) - \delta(s, v)|$. To see this, consider the three cases
(a) If \(\delta(s, u) \geq 1\) and \(\delta(s, v) \geq 1\), then \(|f(u) - f(v)| = 0\).

(b) If \(\delta(s, u) \leq 1\) and \(\delta(s, v) \leq 1\), then \(f(u) - f(v) = \delta(s, u) - \delta(s, v)\).

(c) If \(\delta(s, v) \leq 1 \leq \delta(s, u)\), then \(f(v) = \delta(s, v), f(u) \leq \delta(s, u), \) and \(f(u) \geq f(v)\). So, \(|f(u) - f(v)| = f(u) - f(v) \leq \delta(s, u) - \delta(s, v) = |\delta(s, u) - \delta(s, v)|\).

A similar argument applies when \(\delta(s, u) \leq 1 \leq \delta(s, v)\).

2. \(f(s) = 0\).

**Solution:** \(\delta(s, s) = 0 = f(s)\).

3. If \(f(t) < 1\), then no distribution over \(s-t\) cuts in \(G\) satisfies the BD property.

**Solution:** Suppose there exists a distribution over \(s-t\) cuts in \(G\) that satisfies the BD property. Let \(\{e_1, e_2, \ldots, e_m\}\) denote the distinct set of edges that make up a shortest path from \(s\) to \(t\). From part (a), we know that there must always be an edge \(e_i\) for some \(1 \leq i \leq m\) that crosses a given \(s-t\) cut. In other words if we sample a cut \((C, V - C)\) from the distribution that satisfies the BD property, then \(Pr\{\exists e_i \text{ such that } e \text{ crosses } C\} = 1\). By the union bound,

\[
1 = Pr\{\exists e_i \text{ such that } e \text{ crosses } C\} \leq Pr\{e_1 \text{ crosses } C\} + \ldots + Pr\{e_m \text{ crosses } C\}.
\]

But we also have,

\[
1 > f(t) = p(e_1) + \ldots + p(e_m) \geq Pr\{e_1 \text{ crosses } C\} + \ldots + Pr\{e_m \text{ crosses } C\}.
\]

Contradiction.

Assume that there is a distribution over \(s-t\) cuts in \(G\) that satisfies the BD property. Consider the following sampling algorithm.

\[
\text{Sample}(G, f)
\]

1. Let \(G = (V, E)\).
2. Pick \(\lambda\) uniformly at random from the half-open interval \([0, 1)\).
3. Set \(C = \{v \in V : f(v) \leq \lambda\}\).
4. **Return** \(C\)

(c) Argue the following:
1. The algorithm always produces an $s$-$t$ cut $C$.

**Solution:** Since there is a distribution that satisfies the BD property, then $f(t) = 1$ from part (b.3). Since $\lambda < 1$, the SAMPLE algorithm can never place $t$ in the set $C$. However, $s$ is always in $C$ because $f(s) = 0$. $s$ and $t$ are separated, resulting in an $s$-$t$ cut.

2. For every $e \in E$, we have $\Pr\{e \text{ crosses } C\} = |f(u) - f(v)|$.

**Solution:** $e = (u, v)$ crosses $C$ if and only if $f(u) \leq \lambda < f(v)$ or $f(v) \leq \lambda < f(u)$. In either case $\lambda$ must be chosen in an interval of length $|f(v) - f(u)|$. Since $\lambda$ is uniformly distributed, the probability of this happening is simply $|f(v) - f(u)|$.

Consider the following algorithm for scheduling disconnections at time units $i = 1, 2, \ldots, T$:

\begin{verbatim}
DISCONNECT(G, f)
1   if no distribution satisfies the BD property
2     fail
3   for $i = 1$ to $T$
4     $C = \text{SAMPLE}(G, f)$
5     $G_i = (V, E_i)$, where $E_i = E - \chi(C)$.
\end{verbatim}

For every edge $e \in E$, let $k_e$ be the number of time units during which $e$ is switched off, that is, $k_e = |\{i \in \{1, 2, \ldots, T\} : e \notin E_i\}|$.

(d) Argue that if the algorithm does not fail, then for any $\varepsilon > 0$, for every edge $e \in E$, we have

$$\Pr\left\{\frac{k_e}{T} \geq p(e) + \varepsilon\right\} \leq e^{-2T\varepsilon^2}.$$ 

**Solution:** We will use the Chernoff bound given in Quiz 1, restated here. Let $X_1, \ldots, X_n$ be $n$ independent Boolean random variables. Suppose that for $i = 1, 2, \ldots, n$, we have $\Pr\{X_i = 1\} = \delta_i$ for $0 \leq \delta_i \leq 1$. Let $X = \sum_{i=1}^{n} X_i$ and $\delta = (1/n) \cdot \sum_{i=1}^{n} \delta_i$. Then, for any $\delta \leq \gamma$,

$$\Pr\{X \geq \gamma n\} \leq e^{-2(\gamma - \delta)^2 n}.$$ 

Note that we’ve removed the constraint $\gamma \leq 1$ as was originally written on Quiz 1. If $\gamma \geq 1$, then $\Pr\{X \geq \gamma n\} = 0$ because $X$ can never sum to be more than $n$. In this case, the Chernoff bound still holds but is just not tight.

Let us examine a particular edge $e = (u, v)$. Define $X_i$ to be an indicator r.v. that is 1 if $e$ is switched off in the $i$-th iteration of DISCONNECT, and 0 otherwise. From part
(c.2), \( \Pr \{ \text{e switched off} \} = \Pr \{ \text{e crosses } C \} = |f(u) - f(v)|. \) Let \( n = T. \) Then, \( X = k_e, \) and \( \delta = |f(u) - f(v)|. \) With \( \gamma = p(e) + \epsilon, \) the Chernoff bound gives

\[
\Pr \{ k_e \geq (p(e) + \epsilon)T \} \leq e^{-2(p(e)+\epsilon-\delta)^2T} \leq e^{-2T\epsilon} e^{-2T(p(e)^2+2p(e)\epsilon)} \leq e^{-2T\epsilon^2}
\]

Dividing by \( T \) on both sides,

\[
\Pr \left\{ \frac{k_e}{T} \geq p(e) + \epsilon \right\} \leq e^{-2T\epsilon^2}
\]

N.B. Other forms of the Chernoff bound could have been used as long as they are applied properly.

(e) Suppose that the algorithm does not fail. Show that for any \( \epsilon > 0, \) if \( T \geq \left( \ln |E| \right)/\epsilon^2, \) we have

\[
\Pr \left\{ \exists e \in E \text{ with } \frac{k_e}{T} \geq p(e) + \epsilon \right\} \leq \frac{1}{|E|}.
\]

Solution: Enumerate the edges in \( E \) as \( e_1, \ldots, e_{|E|}. \) Let \( A_i \) be the Boolean event (True or False) that the inequality \( k_{e_i}/T \geq p(e_i) + \epsilon \) is satisfied for edge \( e_i. \) From part (d) with \( T \geq \left( \ln |E| \right)/\epsilon^2, \) we have \( \Pr \{ A_i \} \leq 1/|E|^2 \) for every \( 1 \leq i \leq |E|. \)

With the \( A_i \) notation, we can rewrite the desired probability as

\[
\Pr \left\{ \exists e \in E \text{ with } \frac{k_e}{T} \geq p(e) + \epsilon \right\} = \Pr \left\{ A_1 \text{ or } \ldots \text{ or } A_{|E|} \right\}.
\]

By union bound,

\[
\Pr \left\{ A_1 \text{ or } \ldots \text{ or } A_{m} \right\} \leq \Pr \left\{ A_1 \right\} + \ldots + \Pr \left\{ A_{|E|} \right\} = \frac{1}{|E|}.
\]

(f) Analyze the running time of the algorithm.

Solution: Function \( f \) will be evaluated over all nodes for line 1 of DISCONNECT and line 3 of SAMPLE. Instead of recomputing it every time, \( f \) can be pre-computed using Dijkstra’s algorithm in \( O(V \log V + E) \) time. Each iteration of the for-loop in lines 3-5 of DISCONNECT takes \( O(V) \) time to run SAMPLE and \( O(E) \) time to
compute $E_i$. Over $T$ iterations, this loops takes $O(T(V + E))$. Thus the total time of DISCONNECT is $O(V \log V + T(V + E))$ time.

While the above analysis is a sufficient solution for the problem, let us examine what happens in the special case that $T \geq (\ln |E|)/\epsilon^2$ so that we can assure the bound in part (e). The runtime becomes $O(V \log V + \ln E/\epsilon^2(V + E))$. Note that $\ln(E) = \Theta(\log V)$, so the runtime is $O((V(1 + 1/\epsilon^2) + E/\epsilon^2) \log V)$ time.

**Problem 5-2. Repeated DNA sequences**

A genome’s DNA sequence can be written as a string taken over the alphabet $\{A, T, C, G\}$ where each letter represents one of four possible nucleotides. This string often contains substrings that are repeated at multiple locations and identifying them is important in providing evolutionary and functional insights into a genome. Formally, let $S = s_1 s_2 \cdots s_n$, where $s_i \in \{A, T, C, G\}$ for $1 \leq i \leq n$, be a string representing a genomic sequence. A substring $R$ of $S$ is a repeated substring with length $|R|$ if it appears in at least two distinct locations $x \neq y$ in $S$ such that $R = s_x \cdots s_{x+|R|-1} = s_y \cdots s_{y+|R|-1}$ where $1 \leq x, y \leq n - |R| + 1$. Note that the two repetitions could overlap; for example, substring ATCA is repeated in $S = ATCATCA$. Design an $O(n)$ time algorithm that identifies the repeated substring $R_{\text{max}}$ in $S$ with maximal length, i.e., if $R$ is a repeated substring in $S$, then $|R| \leq |R_{\text{max}}|$. (Hint: Use suffix trees.)

**Solution:** A repeated substring of $S$ that appears at two distinct locations $x \neq y$ is a common prefix of the suffixes that start at $x$ and $y$. Thus, the repeated substring with maximal length can be found by searching for the longest common prefix between any two suffixes. In a suffix tree, the string, denoted as $\text{string}(v)$, that is determined by the path from the root to an intermediate (non-leaf) node $v$ is the longest common prefix among all of the leaf nodes in $v$’s subtree. So, we simply need to search for the intermediate node $v$ that has the longest $\text{string}(v)$ out of all intermediate nodes in the suffix tree.

To do this, first build a suffix tree in $O(n \log \sigma) = O(n)$ time where $\sigma = 4$ is constant. Run a depth-first search on $T$ starting from the root node. At each step of the search, keep track of the longest $\text{string}(v)$. The final value of $\text{string}(v)$ is the repeated substring with maximal length. The tree construction and depth-first search both take linear time.