Internet-Algorithmen

Sommersemester 2014

Prof. Dr. Georg Schnitger
Hannes Seiwert

Arbeitsgruppe Theoretische Informatik, Institut für Informatik

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Übungsblatt 5

Ausgabe: 02.06.2014
Abgabe: 16.06.2014 vor Vorlesungsbeginn

Hinweis: Alle Lösungen sind ordentlich und in lesbarer Schrift zu verfassen. Fasse dich kurz, beschränke deine Erläuterungen und Rechnungen auf die wesentlichen Punkte.

Hinweis: Die Gesamtpunktzahl beträgt 24 Punkte. Darüber hinaus erworbene Punkte werden als Bonuspunkte angerechnet.

Aufgabe 5.1. (8)

Reduktion

Gegeben sei die Knotenmenge V eines ungerichteten Graphen G mit n Knoten. Die Kantenmenge E wird in einer unbekannten Reihenfolge durch einen Datenstrom zur Verfügung gestellt.

Auf dem letzten Blatt haben wir einen Algorithmus entworfen, der die Bipartitheit von G mit $\mathcal{O}(n \log n)$ Bits entscheidet. Wir wollen nun eine untere Schranke für den notwendigen Speicher angeben.

Zeige, dass jeder deterministische Algorithmus, der die Bipartitheit entscheidet, mindestens $\Omega(n)$ Bits an Speicherplatz benötigt. **Reduziere** dazu das Identitätsproblem $x \stackrel{?}{=} y$ auf Bipartitheit.

Erinnerung: Ein Graph ist genau dann bipartit, wenn alle Kreise gerade Länge haben.

Aufgabe 5.2. (8)

doppelseitige Kommunikation

Wir betrachten das Modell der doppelseitigen, randomisierten Kommunikation. Alice und Bob erhalten die dem Partner unbekannten Eingaben $x \in A$ (für Alice) und $y \in B$ (für Bob). Ihr Ziel ist die Berechnung des Funktionswerts $f(x, y)$ für eine den beiden Spielern bekannte Funktion $f : A \times B \to \{0, 1\}$ mit Fehlerwahrscheinlichkeit höchstens δ.

Alice beginnt die Kommunikation und schickt eine Nachricht M_A^1 an Bob, wobei M_A^1 nur von x und einem von Alice angefordertem Zufallsstring abhängt. Bob antwortet mit der Nachricht M_B^1, die von y, M_A^1 und einem von Bob angeforderten Zufallsstring abhängt. Im weiteren Verlauf wechseln sich Alice und Bob ab und verschicken Nachrichten, die jeweils von der eigenen Eingabe, den bisher erhaltenen Nachrichten und einem angeforderten Zufallsstring abhängen dürfen. Irgendwann beendet Alice die Kommunikation und schickt Bob ihre Vermutung über den Funktionswert $f(x, y)$ zu.
Wir fordern, dass die Vermutung von Alice mit Wahrscheinlichkeit mindestens $1 - \delta$ korrekt ist. Die Gesamtanzahl kommunizierter Bits soll – im worst case – möglichst klein sein.

Betrachte nun das Vergleichsproblem $x \stackrel{?}{\leq} y$:
Alice besitzt einen Bitstring $x = (x_1, \ldots, x_n)$ und Bob einen Bitstring $y = (y_1, \ldots, y_n)$. Es ist zu entscheiden, ob $x \leq y$ gilt, wobei $\leq$ die lexikographische Ordnung sei: $x \leq y \Leftrightarrow \sum_i x_i 2^{-i} \leq \sum_i y_i 2^{-i}$.

Entwirf einen randomisierten Algorithmus, der das Vergleichsproblem $x \stackrel{?}{\leq} y$ im doppelseitigen Kommunikationsmodell mit höchstens $\mathcal{O}((\log n)^2)$ Bits Kommunikation und Fehlerwahrscheinlichkeit $\delta \leq 1/4$ entscheidet.

Hinweis: Benutze, dass das Identitätsproblem $x \stackrel{?}{=} y$ für Bitstrings x und y der Länge n mit Fehlerwahrscheinlichkeit $1/n$ gelöst werden kann, wenn $\Theta(\log_2 n)$ Bits ausgetauscht werden.

Aufgabe 5.3. (4+6+4+2+4) *Entropie-Schätzung*

Die Entropie $\mathcal{H}$ einer Wahrscheinlichkeitsverteilung $(p_i)_{i=1,\dots,m}$ ist definiert[1] durch

$$\mathcal{H}(p_i) = -\sum_{i=1}^{m} p_i \log(p_i)$$

Wir wollen die Entropie eines Datenstroms abschätzen. Sei $D = x_1, \dots, x_n$ ein Datenstrom mit Elementen x_k aus einem Universum $U = [m]$. Für $i \in U$ bezeichne a_i die Häufigkeit des Elementes i. (D.h. $a_i = |\{k \in [n] | x_k = i\}|$.) Wir definieren die Wahrscheinlichkeit eines Elementes als seine relative Häufigkeit: $p_i = a_i/n$. Damit ergibt sich die Entropie des Datenstroms als

$$\mathcal{H}(D) = -\sum_{i=1}^{m} \frac{a_i}{n} \log\left(\frac{a_i}{n}\right)$$

Wir wollen die Entropie mit Speicher $o(n)$ approximieren, es ist also nicht möglich, die Häufigkeiten aller Elemente mitzuzählen. Stattdessen wenden wir ein randomisiertes Verfahren an:

(1) Generiere eine auf $[n]$ gleichverteilte Zufallsvariable J und zähle die Häufigkeit R_J des Elementes x_J ab dem Zeitpunkt J.
// Es ist also $R_J = |\{k \,|\, n \geq k \geq J \wedge x_k = x_J\}|$

(2) Berechne $Z = -\left[R_J \log\left(\frac{R_J}{n}\right) - (R_J - 1)\log\left(\frac{R_J - 1}{n}\right)\right]$

(3) Gib die Schätzung Z aus.

a) **Beschreibe** kurz, wie J und R_J speichereffizient erzeugt werden können. Beachte, dass anfangs die Länge n des Datenstroms unbekannt ist.

b) **Zeige:** $E[Z] = \mathcal{H}(D)$. Schreibe dazu $Z = f(R_J) - f(R_J - 1)$ mit $f(y) = -y \log\left(\frac{y}{n}\right)$, dann gilt $\frac{1}{n}\sum_i f(a_i) = \mathcal{H}(D)$ und es bleibt nur noch zu zeigen: $E[Z] = \frac{1}{n}\sum_i f(a_i)$
Hinweis: Berechne zunächst den bedingten Erwartungswert $E[Z|x_J = i]$ für $i \in U$.

c) Sei $\mathcal{H}(D) > 0$ und $\varepsilon \in (0, 1)$. **Zeige** $\text{prob}[\,|\mathcal{H}(D) - Z| \geq \varepsilon \cdot \mathcal{H}(D)] \leq \frac{\log^2(n)}{\varepsilon^2 \cdot \mathcal{H}^2(D)}$.
Hinweis: Zeige $\text{Var}(Z) \leq \log^2(n)$ und benutze dann die Tschebychew-Ungleichung.

d) Wir wollen die Fehlerwahrscheinlichkeit senken, indem wir die Varianz verringern. Dazu führen wir das Verfahren g mal unabhängig von einander aus und erhalten so die Schätzungen $Z_1, \dots, Z_g$. Anschließend geben wir den Mittelwert $\overline{Z} = \frac{1}{g}\sum_{i=1}^{g} Z_i$ der Einzelschätzungen aus. **Bestimme** jeweils eine obere Schranke für $\text{Var}(\overline{Z})$ und $\text{prob}[\,|\mathcal{H}(D) - \overline{Z}| \geq \varepsilon \cdot \mathcal{H}(D)]$.

e) Sei nun $g = \frac{4\log^2(n)}{\varepsilon^2 \cdot \mathcal{H}^2(D)}$ und $\delta \in (0, 1)$ beliebig. Wir führen das Verfahren aus d) $c = \Theta(\ln(1/\delta))$ mal unabhängig von einander aus und erhalten dabei die Mittelwerte $\overline{Z}_1, \dots, \overline{Z}_c$. Anschließend geben wir den Median M der Mittelwerte aus. **Zeige:** $\text{prob}[\,|\mathcal{H}(D) - M| \geq \varepsilon \cdot \mathcal{H}(D)\,] \leq \delta$.
Hinweis: Benutze die Chernoff-Ungleichung.

Fazit: Unser Verfahren liefert also mit Wahrscheinlichkeit $1 - \delta$ eine ε-Approximation der Entropie und benötigt dabei $\mathcal{O}\big(g \cdot c \cdot \text{Speicher}(R_J)\big) = \mathcal{O}\left(\frac{\ln(1/\delta)\cdot\log^2(n)}{\varepsilon^2 \cdot \mathcal{H}^2(D)} \cdot \text{Speicher}(R_J)\right)$ Speicher.

[1] Es gelte die Konvention $0 \cdot \log(0) := \lim_{x \to 0}(x \log(x)) = 0$. Es bezeichne log stets den Logarithmus zur Basis 2 und $[n]$ die Menge $\{1, 2, \dots, n\}$.