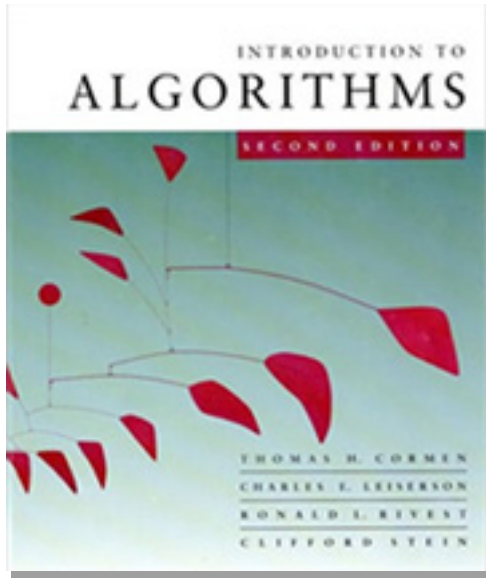


Algoritmalara Giriş

6.046J/18.401J

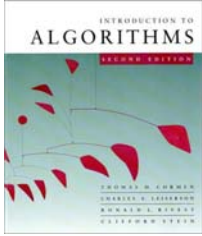


DERS 8

Kıyım Fonksiyonu (Hashing II)

- Evrensel kıyım fonksiyonu
- Evrensellik teoremi
- Evrensel kıyım fonksiyonları kümesini yapılandırmak
- Mükemmel kıyım fonksiyonu

Prof. Charles E. Leiserson



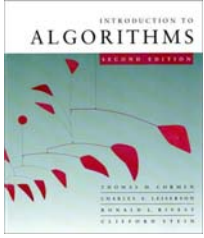
Kıyım fonksiyonunun zaafı

Problem: Her kıyım fonksiyonu h için, kıyım tablosuna ortalama erişim süresini çok büyük ölçüde arttıracak bir anahtar kümesi vardır.

- Rakibiniz bir i yuvası için tüm anahtarları $\{k \in U : h(k) = i\}$ 'den elde edebilir.

FİKİR: Kıyım fonksiyonunu tüm anahtarlardan bağımsız olacak şekilde rastgele seçin..

- Rakibiniz kodunuzu görüyor olsa bile, hangi kıyım fonksiyonunun seçileceğini kesinlikle bilmediğinden, kötü bir anahtar kümesi bulamayacaktır.

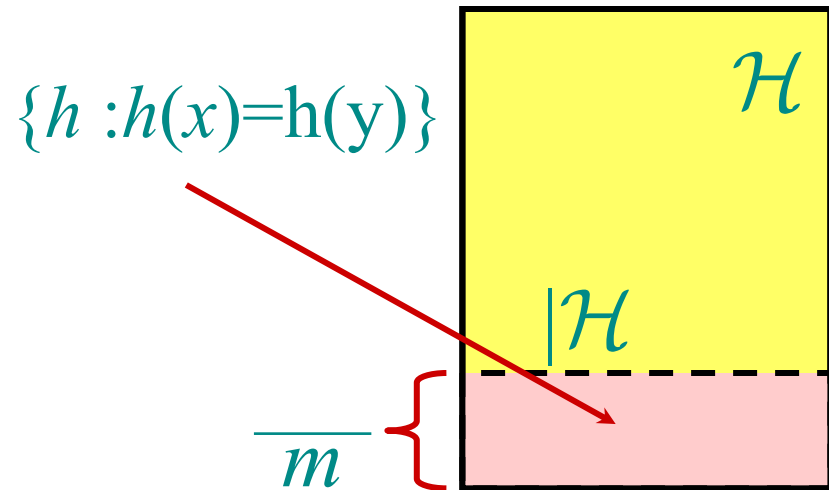


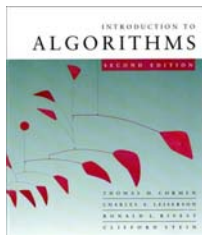
Evrensel kıyım fonksiyonu

Tanım. U bir anahtarlar evreni ve \mathcal{H} de sınırlı sayıdaki kıyım fonksiyonunun kümesi olsun; herbiri U 'yu $\{0, 1, \dots, m-1\}$ 'e eşlemlesin.

\mathcal{H} 'nin *gxt gpn* olması için: $x, y \in U$, ve $x \neq y$, ile $|\{h \in \mathcal{H} : h(x) = h(y)\}| = |\mathcal{H}|/m$ olması gerekir.

Yani, z ile $\{$ arasında bir çarpışma olasılığı $1/m$ 'dir; koşul: h 'nin \mathcal{H} 'den rastgele seçimi .



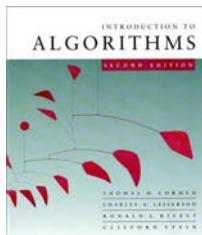


Evrensellik iyidir

Teorem. h (tekbiçimli olarak) rastgele seçilmiş bir kıyım fonksiyonu olsun; seçim evrensel bir \mathcal{H} kıyım fonksiyonları setinden yapılmış olsun. h 'nin n rastgele anahtarı T tablosundaki m yuvaya kıyımladığını farzedin.

Bu durumda belirli bir x anahtarı için:

$$E[x \text{ ile } \text{\char"00A7} \text{ sayısı}] < n/m.$$

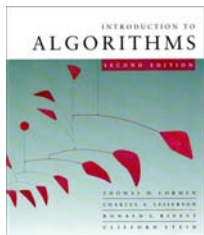


Teoremin kanıtı

Kanıt. C_x , T 'nin içindeki anahtarlarla z 'in toplam çarpışma sayısını gösteren rastgele değişken olsun; ve

$$c_{xy} = \begin{cases} 1 & \text{eğer } h(x) = h(y), \\ 0 & \text{(diğer durumlarda)} \end{cases} \quad \text{olsun.}$$

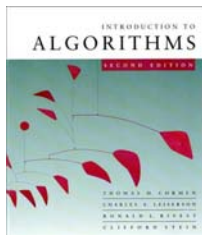
Not: $E[c_{xy}] = 1/m$ ve $C_x = \sum_{y \in T - \{x\}} c_{xy} \cdot$



Kanıt(devamı)

$$E[C_x] = E \left[\sum_{y \in T - \{x\}} c_{xy} \right]$$

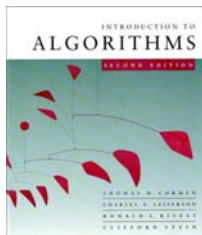
- İki tarafın da umulanını bulun.



Kanıt (devamı)

$$\begin{aligned} E[C_x] &= E \left[\sum_{y \in T - \{x\}} c_{xy} \right] \\ &= \sum_{y \in T - \{x\}} E[c_{xy}] \end{aligned}$$

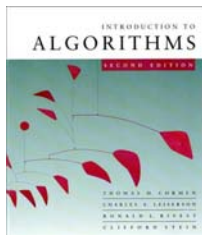
- İki tarafın da umulanını bulun.
- Umulanın doğrusallığı (expectation).



Kanıt (devamı)

$$\begin{aligned} E[C_x] &= E \left[\sum_{y \in T - \{x\}} c_{xy} \right] \\ &= \sum_{y \in T - \{x\}} E[c_{xy}] \\ &= \sum_{y \in T - \{x\}} 1/m \end{aligned}$$

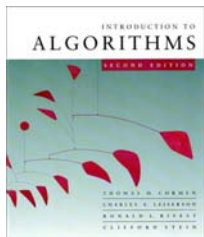
- İki tarafın da umulanını bulun.
- Umulanın doğrusallığı (expectation).
- $E[c_{xy}] = 1/m$.



Proof (continued)

$$\begin{aligned} E[C_x] &= E \left[\sum_{y \in T - \{x\}} c_{xy} \right] \\ &= \sum_{y \in T - \{x\}} E[c_{xy}] \\ &= \sum_{y \in T - \{x\}} 1/m \\ &= \frac{n-1}{m} . \quad \square \end{aligned}$$

- Take expectation of both sides.
- Linearity of expectation.
- $E[c_{xy}] = 1/m$.
- Algebra.



Constructing a set of universal hash functions

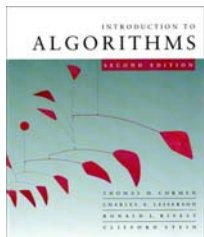
Let m be prime. Decompose key k into $r + 1$ digits, each with value in the set $\{0, 1, \dots, m-1\}$. That is, let $k = \langle k_0, k_1, \dots, k_r \rangle$, where $0 \leq k_i < m$.

Randomized strategy:

Pick $a = \langle a_0, a_1, \dots, a_r \rangle$ where each a_i is chosen randomly from $\{0, 1, \dots, m-1\}$.

Define $h_a(k) = \sum_{i=0}^r a_i k_i \bmod m$. *Dot product, modulo m*

How big is $\mathcal{H} = \{h_a\}$? $|\mathcal{H}| = m^{r+1}$. ← **REMEMBER THIS!**



Universality of dot-product hash functions

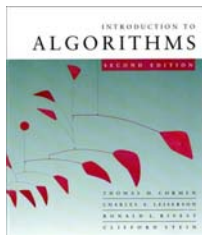
Theorem. The set $\mathcal{H} = \{h_a\}$ is universal.

Proof. Suppose that $x = \langle x_0, x_1, \dots, x_r \rangle$ and $y = \langle y_0, y_1, \dots, y_r \rangle$ be distinct keys. Thus, they differ in at least one digit position, wlog position 0 .

For how many $h_a \in \mathcal{H}$ do x and y collide?

We must have $h_a(x) = h_a(y)$, which implies that

$$\sum_{i=0}^r a_i x_i \equiv \sum_{i=0}^r a_i y_i \pmod{m}.$$



Proof (continued)

Equivalently, we have

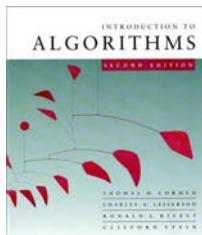
$$\sum_{i=0}^r a_i (x_i - y_i) \equiv 0 \pmod{m}$$

or

$$a_0 (x_0 - y_0) + \sum_{i=1}^r a_i (x_i - y_i) \equiv 0 \pmod{m},$$

which implies that

$$a_0 (x_0 - y_0) \equiv -\sum_{i=1}^r a_i (x_i - y_i) \pmod{m}.$$



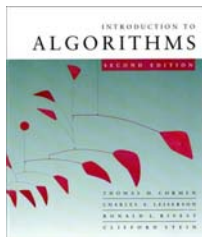
Fact from number theory

Theorem. Let m be prime. For any $z \in \mathbb{Z}_m$ such that $z \neq 0$, there exists a unique $z^{-1} \in \mathbb{Z}_m$ such that

$$z \cdot z^{-1} \equiv 1 \pmod{m}.$$

Example: $m = 7$.

z	1	2	3	4	5	6
z^{-1}	1	4	5	2	3	6



Back to the proof

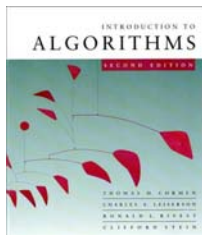
We have

$$a_0(x_0 - y_0) \equiv -\sum_{i=1}^r a_i(x_i - y_i) \pmod{m},$$

and since $x_0 \neq y_0$, an inverse $(x_0 - y_0)^{-1}$ must exist, which implies that

$$a_0 \equiv \left(-\sum_{i=1}^r a_i(x_i - y_i) \right) \cdot (x_0 - y_0)^{-1} \pmod{m}.$$

Thus, for any choices of a_1, a_2, \dots, a_r , exactly one choice of a_0 causes x and y to collide.



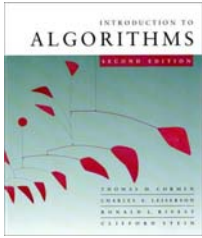
Proof (completed)

Q. How many h_a 's cause x and y to collide?

A. There are m choices for each of a_1, a_2, \dots, a_r , but once these are chosen, exactly one choice for a_0 causes x and y to collide, namely

$$a_0 = \left(\left(- \sum_{i=1}^r a_i (x_i - y_i) \right) \cdot (x_0 - y_0)^{-1} \right) \bmod m.$$

Thus, the number of h 's that cause x and y to collide is $m^r \cdot 1 = m^a = |\mathcal{H}|/m$. \square

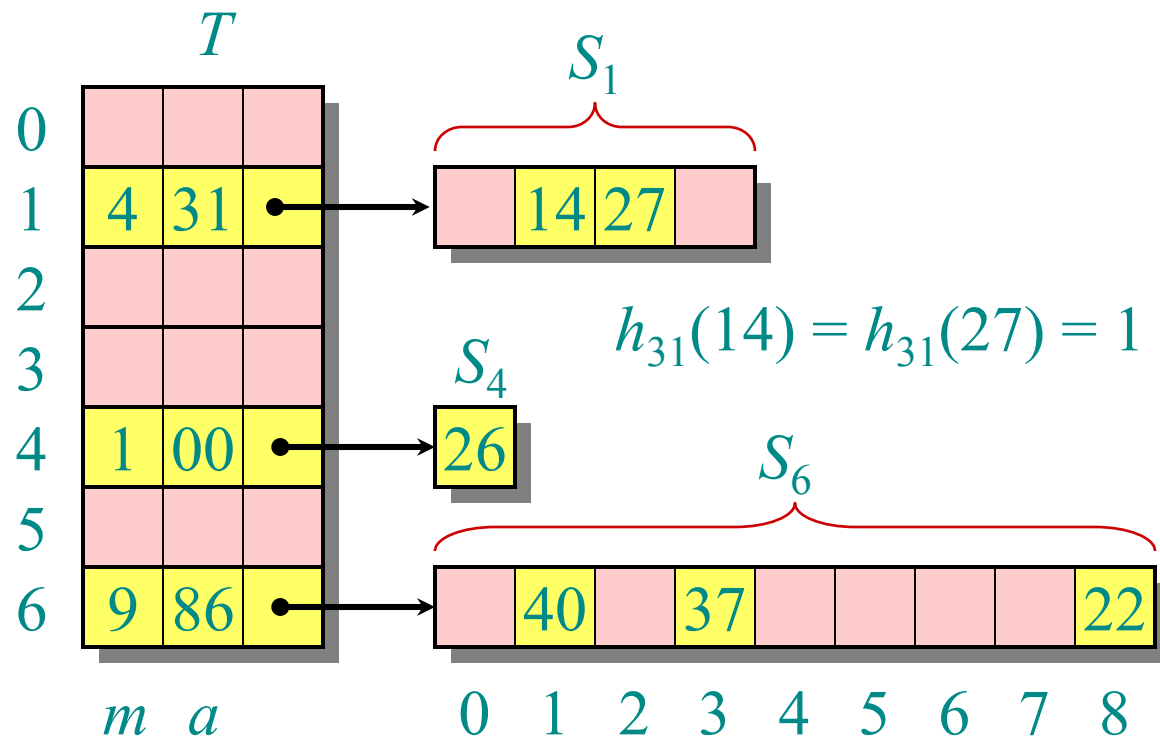


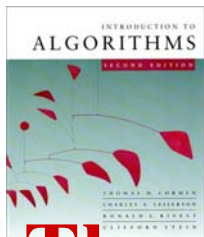
Perfect hashing

Given a set of n keys, construct a static hash table of size $m = O(n)$ such that **SEARCH** takes $\Theta(1)$ time in the *worst case*.

IDEA: Two-level scheme with universal hashing at both levels.

No collisions at level 2!



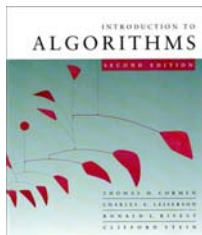


Collisions at level 2

Theorem. Let \mathcal{H} be a class of universal hash functions for a table of size $m = n^2$. Then, if we use a random $h \in \mathcal{H}$ to hash n keys into the table, the expected number of collisions is at most $1/2$.

Proof. By the definition of universality, the probability that 2 given keys in the table collide under h is $1/m = 1/n^2$. Since there are $\binom{n}{2}$ pairs of keys that can possibly collide, the expected number of collisions is

$$\binom{n}{2} \cdot \frac{1}{n^2} = \frac{n(n-1)}{2} \cdot \frac{1}{n^2} < \frac{1}{2} \quad \square$$



No collisions at level 2

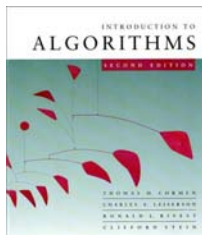
Corollary. The probability of no collisions is at least $1/2$.

Proof. Markov's inequality says that for any nonnegative random variable X , we have

$$\Pr\{X \geq t\} \leq E[X]/t.$$

Applying this inequality with $t = 1$, we find that the probability of 1 or more collisions is at most $1/2$. ■

Thus, just by testing random hash functions in \mathcal{H} , we'll quickly find one that works.



Analysis of storage

For the level-1 hash table T , choose $m = n$, and let n_i be random variable for the number of keys that hash to slot i in T . By using n_i^2 slots for the level-2 hash table S_i , the expected total storage required for the two-level scheme is therefore

$$E \left[\sum_{i=0}^{m-1} \Theta(n_i^2) \right] = \Theta(n),$$

since the analysis is identical to the analysis from recitation of the expected running time of bucket sort. (For a probability bound, apply Markov.)