**ALG 4.1**

Randomized Pattern Matching:

Reading Selection:
CLR: Chapter 12
Handout: R. Karp and M. Rabin, "Efficient Randomized Pattern-Matching"

**Generalized Pattern Matching**

**Input**
- index set $R$
- for each $r \in R$ strings $X(r)$, $Y(r)$ $\in \{0,1\}^m$

**Problem**
- find $r \in R$ s.t. $X(r) = Y(r)$

**Examples:**

1. **ID String Pattern Matching**
   
   - **Input**: pattern $X = X_1 \ldots X_m \in \{0,1\}^m$
   - **Text**: $Y = Y_1 \ldots Y_n \in \{0,1\}^n$
   - **Index set**: $R = \{1,2,\ldots, n-m+1\}$
   - $\forall r \in R$
     - $X(r) = X$
     - $Y(r) = Y_r \ Y_r+1 \ldots \ Y_{r+(m-1)}$
(2) 2D Array Matching

**Input pattern**

$s \times s$ binary array $X = (X_{ij}), m = s^2$

text $b \times b$ binary array $Y = (Y_{ij}), n = b^2$

**Index set**

$R = \{< i,j > | s \leq i,j \leq b\}$

$X(< i,j >) =$ string of rows of $X$

$Y(< i,j >) =$ string of rows in $s \times s$ block of $Y$

with $(i,j)$ in lower right position.

*(note Karp & Pratt reverse n,m)*

**Pattern Matching by Fingerprinting**

*S is a finite set*

$\forall p \in S \quad \Phi_p (\cdot) : \{0,1\}^m \rightarrow \text{small range } D_p$

*$\Phi_p (X)$ is "fingerprint" for string $X$

**Idea**

compare $X(r) = Y(r)$ only if fingerprints agree: $\Phi_p(X(r)) = \Phi_p(Y(r))$

**Algorithm**

$p \leftarrow$ random element of $S$

for each $r \in R$ in order do

begin

compute $a_p(r) = \Phi_p(X(r))$

compute $b_p(r) = \Phi_p(Y(r))$

if $a_p(r) = b_p(r)$ then

if $X(r) = Y(r)$ then output "Match at r"

end
Requirements

(1) small domain \(D_p\)

(2) small probability of false match
\(\Phi_p(X(r)) = \Phi_p(Y(r)) \text{ but } X(r) \neq Y(r)\)

(3) fingerprints \(\Phi_p(X(r)), \Phi_p(Y(r))\) are easily updatable from previous \(r\)

Examples of fingerprints:

(A) integer modular functions.

(B) unimodular matrices

(C) irreducible polynomial modular functions.

represent binary string \(X = X_1 \ldots X_m\)

by integer \(H(x) = \sum_{i=1}^{m} X_i 2^{m-i}\)

modular fingerprint \(\Phi_p(x) = \text{res } (H(x), p)\)

modular fingerprint \(\Phi_p(X) = \text{res } (H(X), p)\)

\[= p \cdot \frac{H(X)}{p} - H(X)\]

note \(\Phi_p(x) \equiv H(X) \mod p\)
Define \( S = \{ p \mid p \text{ is prime and } p \leq M \} \)

where \( M \) is a (suf. large) integer

**idea**

choose random \( p \in S \)

\( \Rightarrow \) must prove \( \Phi_p(X) = \text{res}(H(X), p) \) is

good fingerprint

---

**Facts about Prime Numbers**

let \( \Pi(k) = \text{number of primes } \leq k \)

**FACT 1**

If \( k \geq 29 \) and \( a \leq 2^k \), then

\( a \) has \( \leq \Pi(k) \) distinct prime divisors

**proof**

follows from Rosser & Schoenfeld bound:

\[
\prod_{\text{prime } p \leq k} p > e^{1 - \frac{1}{2} \ln \frac{k}{\ln k}}
\]

for all \( k \geq 17 \)

\[
\frac{k}{\ln k} \leq \Pi(k) \leq 1.25506 \frac{k}{\ln k}
\]

(prime number theorem)
Suppose Randomized Pattern Match

Algorithm is executed, with
fingerprint \( \Phi_p(X) = \text{res}(H(x), p) \)
with set \( S = \{ p \mid p \text{ prime } \leq M \} \)
where \( M = mn^2 \) and \( mn \geq 29 \)

**Theorem.**

for each \( r \in R \), probability of
false match \( \Phi_p(X(r)) = \Phi_p(Y(r)) \) but \( X(r) \neq Y(r) \)
is \( \leq \frac{2.511}{n} \)

**proof**

A false match occurs only if \( \exists r \in R \)
\( X(r) \neq Y(r) \) and \( p \mid (H(x(r)) - H(Y(r))) \)
iff \( p \mid L \) where \( L = \prod_{X(r) \neq Y(r)} | H(X(r)) - H(Y(r)) | \)

But \( L \leq 2^{mn} \) so by Fact 1,
\( L \) has at most \( \Pi(mn) \) prime divisors

Since \( p \) is chosen at \text{random} from \( \Pi(M) \) primes,
and only \( \Pi(mn) \) give false matches,

\[
\text{Prob(false match)} \leq \frac{\Pi(mn)}{\Pi(M)} \leq \frac{2.511}{n}
\]

by Fact 2.
Updating Modular Fingerprints

**Pattern**

\[ X = X_1 \ldots X_m \]

**Text**

\[ Y = Y_1 \ldots Y_n \]

\[ X(r) = X, \quad Y(r) = Y_r Y_{r+1} \ldots Y_{r+m-1} \]

Since

\[ Y(r+1) = (Y(r) - 2^{m-1} Y_r) \cdot 2 + Y_{r+m} \]

**Update Formula:**

\[ a_p(r+1) = (a_p(r) + a_p(r) + \xi Y_r + Y_{r+m}) \mod p \]

where \( \xi = -2^{m} \mod p \)

**Proof**

Expected Time is:

\[ O(n) + nm \text{ Prob(false match) } \leq O(n) + nm \left( O\left( \frac{1}{n} \right) \right) \]

\[ \leq O(n) \]

**Theorem**

Total Exp. Time for finding a match is \( O(n) \)
2D Randomized Pattern Matching

**input**  
pattern \( X = (X_{ij}) \) is \( s \times s \) boul. array

**text**  
\( Y = (Y_{ij}) \) is \( b \times b \) boul. array

**text window**  
\( Y(< i,j >) \) = concatenation of rows of \( s \times s \) subarray of \( Y \) with \( < i,j > \) in lower right corner

---

**Index i Update**

\[
\text{fingerprint } \quad a_p(< i+1,j >) = a_p(< i,j >) + \left( a_p(< i,j >) - \lambda \cdot Y_{i-s+1,j} + Y_{i+1,j} \right) \mod p
\]

where \( \lambda = 2^s \mod p \)

\[
\text{Index j Update } \quad a_p(< i,j+1 >) = a_p(< i,j >) \cdot 2^{-s} + \left[ Y_{i,j+1} \cdot 2 + Y_{2,j+1} \cdot 2^2 + \ldots + Y_{s,j+1} \cdot 2^s \right] \mod p
\]

where \( \theta = 2^{s(s-1)} \mod p \)
**Unimodular Matrices as Fingerprints**

**Definition**

homomorphism \( k \) from \( \{0,1\}^* \) into unimodular matrices with

\[
k(\varepsilon) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad k(0) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad k(1) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}
\]

\( k(X \ast Y) = k(X) \cdot k(Y) \)

concatenation \( \uparrow \) \quad matrix multiplication

**Fact 1'** If \( X \in \{0,1\}^m \), then each entry of \( k(X) \) is \( \leq F_m = m \text{th Fibonacci} \)

(where \( F_0 = F_1 = 1 \), and \( F_m = F_{m-1} + F_{m-2} \) for \( m \geq 2 \))

**Fact 2'** \( \log F_m \sim 0.694m \)

Suppose the unimodular fingerprint \( \Phi_p \) is used with random \( p \in S = \{p | p \leq M \text{ is prime}\} \) where \( M = mn^2 \).

**Theorem**

The Random Pattern Matching Algorithm has probability \( \leq \frac{6.971}{n} \) of false match.
proof

A false match occurs if \( \exists i,j \in \{1,2\} \)
\[
\Phi_p(X(r))_{i,j} = \Phi_p(Y(r))_{i,j} \text{ but } k(X(r))_{i,j} \neq k(Y(r))_{i,j}
\]
iff \( p \mid L' \) where \( L' = \prod_{r \in R, \{1,2\}} |k(X(r))_{i,j} - k(Y(r))_{i,j}| \)

But \( L' \leq (F_m)^{4n} \leq 2^{4n \log F_m} \)

By fact 1, the number of primes that divide \( L' \) is at most
\[\prod \left( 2^{4n \log F_m} \right)\]

The probability of a false match is
\[\prod \left( 2^{4n \log F_m} \right) \leq \frac{6.971}{n} \text{ since by Fact 2'}
\]
\[
\log(F_m^{\frac{1}{n}}) = .694m
\]

Updating Using Unimodular Matrices

ID string matching:
\[
a_p(r) = \Phi_p(Y(r))
\]
\[
a_p(r+1) = K_p(Y_r)^{-1} a_p(r) H(Y_{r+m}) \mod p
\]

where \( K_p(0)^{-1} = \begin{pmatrix} 1 & 0 \\ p-1 & 1 \end{pmatrix} \)

and \( K_p(1)^{-1} = \begin{pmatrix} 1 & p-1 \\ 0 & 1 \end{pmatrix} \)

(can also extend to 2D string matching)
Simplifications of Modular Fingerprinting

\[ S = \{ p \mid p \leq M \text{ and } p \text{ prime or pseudoprime} \} \]

- **p is pseudoprime** if \(2^{p-1} \equiv 1 \mod p\) but \(p\) is not prime.
- \# pseudoprimes \( \leq Me^{-c(\log M \log \log M)} \).

A false match occurs when \(p|L\) where
\[
L = \prod (H(X(r)) - H(Y(r)))
\]

Let \(N\) be the number of M-fat integers dividing \(L\)

\[
N \leq \frac{\ln 2}{2} \frac{n^4}{(\log n)^2}
\]

If \( M = \frac{n^4}{(\ln n)^2} \), then \( \frac{N}{|S|} < 0.5 \).

- **Fact**
  \[ |S| \sim M(\ln 2 + o(1)) \]

- **Bound** \( L < 2^{nm} \)

- **False Match Probability**
  \[
  \text{prob of false match} = \frac{N}{|S|} < 0.5
  \]
(3) \( S = \{ p | p \leq M \} \) with some \( M \)

**idea** use new \( p \) when get false match

**expected time**

\[ cn \left( .5 + (.5)^2 + (.5)^3 + \ldots \right) \leq O(n) \]

---

**Fingerprinting by Random Polynomials**

**Galois Field**

\[ GF(2^k) = \{ b_1 \ldots b_k | b_1, \ldots, b_k \in \{0,1\} \} \]

\[ Z_2[t] = \text{polynomials of form} \]

\[ \frac{p(t)}{p(t)} = t^k + a_{k-1} t^{k-1} + \ldots + a_0 \quad \text{where} \]

\[ a_{k-1}, a_{k-2}, \ldots, a_0 \in \{0,1\} \]

\( p(x) \) **irreducible** if can’t be factored

**Lemma**

If \( k \) is prime, the number of irreducible polynomials of degree \( k \) in \( Z_2[t] \) is

\[ \frac{2^k - 2}{k} \]

**Fingerprint fn**

\( \Phi_p(x) = x_1 t^{m-1} + \ldots + x_m \mod p(t) \)

(residue comp can be done efficiently)
Theorem

If use random Fingerprint fn
with degree $k > \log(nm\epsilon^{-1})$, then prob of
false match is $\leq \epsilon$.

proof

use usual argument and above Lemma

Open Problems

(1) Are there deterministic methods
for Fingerprinting?

(2) What are optimal trade offs for
prob of error and size of $S$
for randomized Fingerprinting?
for fixed random $p$,

**idea** store $\Phi_p(F_1), \ldots, \Phi_p(F_k)$

fingerprints of files $F_1, \ldots, F_k$

**Security:**

only operator knows $p$, so if

any file $F_i$ modified, to $F'_i$

then with high likelihood

$$\Phi_p(F_i) \neq \Phi_p(F'_i)$$

$\implies$ can build a secure operating system

from this idea!