A note on the fingerprint of Karp-Rabin

Rossano Venturini

Definition 1 *Let $\pi(u)$ denote the number of primes that are smaller than or equal to an integer u.*

For example let $u = 29$, then $\pi(29)$ is equal to 10 because the primes smaller than or equal to 29 are $2, 3, 5, 7, 11, 13, 17, 19, 23, 29$.

Theorem 1 *For any positive integer u, it holds*

$$\frac{u}{\ln u} \leq \pi(u) \leq 1.26\frac{u}{\ln u}$$

Lemma 1 *If $u \geq 29$, then the product of all the primes that are smaller than or equal to u is greater than 2^u.*

For example let $u = 29$, the primes smaller than or equal to 29 are $2, 3, 5, 7, 11,\ 13, 17, 19, 23, 29$. Their product is equal to $6,469,693,230$ which is greater than $2^{29} = 536,870,912$.

Corollary 1 *If $u \geq 29$ and x is any positive integer smaller than 2^u, then x has fewer than $\pi(u)$ distinct prime divisors.*

Proof: By contradiction assume that x has $k \geq \pi(u)$ distinct prime divisors $q_1, q_2, \ldots, q_k$. It is easy to see that $x \geq q_1 \cdot q_2 \cdot \ldots \cdot q_k$ (notice that the equality holds when each prime occurs with exponent 1 in the factorization of x). On the other hand, $q_1 \cdot q_2 \cdot \ldots \cdot q_k$ is at least as large as the product of the first $\pi(u)$ prime numbers. Thus, by Lemma 1 it should be larger than or equal to 2^u. This implies $x \geq 2^u$ which contradicts the hypothesis $x < 2^u$. ∎

Next Theorem establishes a bound to the probability that a false match occurs during the execution of the algorithm. Recall that we have a false match iff the pattern P and the substring T_r are different but the values $H_q(P)$ and $H_q(T_r)$ are the same. Observe that $H_q(P) = H_q(T_r)$ is equivalent to $H(P) \equiv H(T_r)\ (mod\ q)$, which corresponds to say that q divides $|H(P) - H(T_r)|$.

Theorem 2 *Let $I > nm$ be a positive integer, and q a randomly chosen prime smaller than or equal to I. If $nm \geq 29$, then the probability of a false match between P and T is at most $\frac{\pi(nm)}{\pi(I)}$.*

Proof: Let R be the set of positions in T where P **does not** begin. We have $\prod_{r \in R} |H(P) - H(T_r)| < 2^{mn}$. In fact, notice that each factor $|H(P) - H(T_r)|$ is at most 2^m[1] while the number of factors (i.e., the cardinality of R) is at most $n - m$. If there is a false match at position r, q must divide $|H(P) - H(T_r)|$ (see observation above). Since $r \in R$, q must also divide the product above. By the Corollary 1 this product has at most $\pi(nm)$ distinct prime divisors, thus, a false match implies that q is one of these primes. Since q is chosen randomly out of a set of size $\pi(I)$, the probability of a false match is at most $\frac{\pi(nm)}{\pi(I)}$. ∎

We can fix I properly in order to achieve a small probability for a false match.

Theorem 3 *By setting $I = n^2m$, the probability of a false match is at most $\frac{2.53}{n}$.*

Proof: The probability of a false match is at most $\frac{\pi(nm)}{\pi(n^2m)} \leq 1.26\frac{nm\ln(n^2m)}{n^2m\ln(nm)}$. With some algebra we can obtain the claimed bound. ∎

[1] The maximum value of the function H is 2^m and it is given by the m-long strings containing only 1s. Conversely, the minimum value is 0 and it is given by the string formed by only 0s. Thus, $|H(P) - H(T_r)|$ is maximized when $H(P) = 2^m$ and $H(T_r) = 0$ or vice versa.