CS 127/CSCI E-127: Introduction to Cryptography

Prof. Salil Vadhan — Fall 2013

Lecture Notes 18:

Collision-Resistant Hash Functions

Reading.

- Katz-Lindell (2nd ed) §5.0-5.3.1, 5.4.1, 5.6.1, 8.4.2.

1 Definition

Idea: Sign (or MAC) a long message m by first hashing it. What properties will we want from the hash function h?

- $||h(x)|| \ll ||x||$.
- h easy to evaluate.
- Hard to find *collisions*, i.e. (x, x') s.t. $x \neq x'$ and $h(x) = h(x')$.

Definition 1 (collision-resistant hash functions) $\mathcal{H} = \bigcup_n \{h_i : \{0,1\}^{\ell_{in}(n)} \to \{0,1\}^{\ell_{out}(n)}\}_{i \in \mathcal{I}_n}$ *is a family of* collision-resistant hash functions *if*

- *(generation) There is a PPT* $\mathsf{Gen}(1^n)$ *which outputs* $i \in \mathcal{I}_n$.
- *(hashing)* $\ell_{in}(n) > \ell_{out}(n)$.
- *(easy to evaluate) Given* x, i, *can compute* $h_i(x)$ *in poly-time.*
- *(hard to form collisions) For every PPT* $\mathcal{A}$, *there is a negligible function* ε *such that*

$$\forall n \qquad \Pr\left[\mathcal{A}(1^n, I) = (X, X') \text{ s.t. } X \neq X' \text{ and } h_I(X) = h_I(X')\right] \leq \varepsilon(n)$$

where the probability is taken over $I \stackrel{R}{\leftarrow} \mathsf{Gen}(1^n)$ *and the coin tosses of* $\mathcal{A}$. *Equivalently,* $\mathcal{A}$ *succeeds with negligible probability in the following collision-finding game:*

Typically, we want the range to be much smaller than the domain, we can think of $\{0,1\}^{\ell_{in}(n)} = \{0,1\}^*$, $\{0,1\}^{\ell_{out}(n)} = \{0,1\}^n$.

2 Hash-then-Sign

We present it for signatures, but it also works for MACs. Let $(\mathsf{Gen}, \mathsf{Sign}, \mathsf{Vrfy})$ be a signature scheme for message space $\{0,1\}^{\ell_{out}(n)}$, and let $\mathcal{H}$ be a collection of hash functions with domain $\{0,1\}^{\ell_{in}(n)}$ and range $\{0,1\}^{\ell_{out}(n)}$. Define a new signature scheme $(\mathsf{Gen}', \mathsf{Sign}', \mathsf{Vrfy}')$ for message space $\{0,1\}^{\ell_{in}(n)}$ by setting

- $pk' = (pk, i)$, $sk' = (sk, i)$.
- $\mathsf{Sign}'_{sk'}(m) = \mathsf{Sign}_{sk}(h_i(m))$.
- $\mathsf{Vrfy}'_{pk'}(m, \sigma) = \mathsf{Vrfy}_{pk}(h_i(m), \sigma)$.

Theorem 2 *If* $(\mathsf{Gen}, \mathsf{Sign}, \mathsf{Vrfy})$ *is a secure signature scheme for message space* $\{0,1\}^{\ell_{out}(n)}$ *and* $\mathcal{H}$ *is collision resistant, then* $(\mathsf{Gen}', \mathsf{Sign}', \mathsf{Vrfy}')$ *is a secure one-time signature scheme for message space* $\{0,1\}^{\ell_{in}(n)}$.

Proof:

■

Hash-then-sign also works for general (i.e. many-time) signatures and MACs.

Q: What are potential advantages of this construction over the domain extension for MACs that we saw last time $(\mathsf{Mac}'_k(m_1\|m_2\|\cdots\|m_d) = \mathsf{Mac}_k(m_1\|1\|r\|d)\mathsf{Mac}_k(m_2\|2\|r\|d)\cdots\mathsf{Mac}_k(m_d\|d\|r\|d))$?

3 Domain Extension for CRHFs

The definition of Collision -Resistant Hash Functions only requires shrinking by one bit. To shrink more may apply "Merkle–Damgård" methodology:

- First design a collision-resistant "compression function" $h_i : \{0,1\}^{\ell+n} \to \{0,1\}^n$.
- For a message $m \in \{0,1\}^*$, break into ℓ-bit blocks $m_1 m_2 \cdots m_t$, where m_t *contains the length of* m, and define $H_i(m) = h(m_t \circ h(m_{t-1} \circ h(m_{t-2} \cdots h(m_1 \circ \mathrm{IV}))))$, where IV is a fixed initial vector (e.g. $\mathrm{IV} = 0^n$).

Proposition 3 *If* $\{h_i\}$ *is a collision-resistant family, then so is* $\{H_i\}$.

Proof:

4 Attacks on CRHFs

There are different attacks on collision-resistant hash functions:

- Random guessing: Suppose $\ell_{in} = 2n$. pick m, m' randomly from $\{0,1\}^{2n}$. The probability of success is greater than $\frac{1}{2^n} - \frac{1}{2^{2n}}$.
- Birthday attack: pick random messages to find a collision. We choose t messages randomly from $\{0,1\}^{2n}$ and the expected number of collisions is:

$$\begin{aligned} \mathrm{E}[\#\text{ collisions}] &= \#\text{ pairs} \cdot \Pr[\text{any one pair collide}] \\ &\geq \binom{t}{2} \cdot \left(\frac{1}{2^n} - \frac{1}{2^{2n}}\right) \\ &\sim \frac{t^2}{2^{n+1}} \end{aligned}$$

  If we pick $t = \Theta(2^{n/2})$, the expected number of collisions is large, and in fact it can be shown that a collision will be found with high probability. Quadratic savings over exhaustive search (though still exponential in n). **Q:** Disadvantages over exhaustive search?

5 Constructions

5.1 Number-Theoretic Constructions

Theorem 4 *Collections of collision-resistant hash functions exist under either the Factoring Assumption or the Discrete Log Assumption.*

Proof Sketch: Construction based on Discrete Log: First construct $h_{p,g,y} : \mathbb{Z}_{p-1} \times \{0,1\} \to \mathbb{Z}_p^*$ by $h_{p,g,y}(x,b) = y^b \cdot g^x \bmod p$. A collision for $h_{p,g,y}$ yields the discrete log of y. □

5.2 Hash Functions in Practice

Typical design features:

- Tailor-designed functions $H : \{0,1\}^* \to \{0,1\}^n$, with e.g. $n = 128$ or $n = 160$. (Note that n is larger than for block ciphers to protect against birthday attacks)
- Very fast.
- Designed to be collision-resistant (in strong sense), have "random looking" output.
- Confusion & diffusion
- Not related to any nice complexity assumption.
- Not a "family" — conjecture that collisions are hard to find in this particular function.

Some popular hashfunctions:

- MD4 — Message Digest 4
  - Designed by Ron Rivest (1990), $n = 128$, $\ell = 512$.
  - Collisions have been found (1995). Design is basis for stronger hash functions (MD5, SHA).
  - Follows Merkle–Damgård with compression function $h : \{0,1\}^{512+128} \to \{0,1\}^{128}$.
- MD5 — improvement to MD4 (Rivest, 1992). Collisions have been found (1998).
- SHA-1 — another improvement to MD4 (NIST w/NSA, 1994)
  - hash size $n = 160$, so compression function is $h : \{0,1\}^{512+160} \to \{0,1\}^{160}$.
  - In retrospect, compression function designed from a block cipher $\{f_k\}$ using Davies–Meyer construction $h(k, x) = f_k(x) \oplus x$. (Collision resistance of Davies–Meyer can be proved in "ideal cipher model," where the f_k's are modelled as a publicly computable family of random and independent permutations. This is even more idealized than random-oracle model, and is definitely not an appropriate model for some block ciphers like DES.)
  - Collisions can be found in time 2^{60} (better than "birthday attack") (2005).
- SHA-3
  - Selected by NIST in 2012 based on public competition.
  - Output lengths 256 and 512.
  - Not based on Merkle–Damgård construction.
  - Still being standardized.