Avoiding collisions
Cryptographic hash functions

Foundations of Cryptography
Computer Science Department
Wellesley College

Fall 2016

Table of contents

Introduction

Davies-Meyer

Hashes in Practice
Collision-resistant hash functions are generally designed in two steps:

1. A compression function is designed to handle fixed-length hashes.

2. The compression function is extended to handle arbitrary input lengths (think Merkle-Damgård).

Today we discuss the first step.

Hash functions from block ciphers

- Collision-resistant compression functions can be built from block ciphers.

- **Davies-Meyer construction:** Given a block cipher $F$ with $n$-bit key length and $\ell$-bit block length, define $h : \{0, 1\}^{n+\ell} \rightarrow \{0, 1\}^{\ell}$ by

  $$h(k, x) \overset{\text{def}}{=} F_k(x) \oplus x.$$

- Bad news: A prove of collision-resistance based on the assumption that $F$ is a strong pseudorandom permutation is not known. However, ...
The ideal cipher model

- The ideal-cipher model posits that all parties have access to an oracle for a random keyed permutation $F : \{0,1\}^n \times \{0,1\}^\ell \rightarrow \{0,1\}^n$ and its inverse $F^{-1}$.
- The only way to compute either $F$ or $F^{-1}$ is to explicitly query the oracle with key $(k, x)$.
- Such a model implies the absence related-key attacks since permutations $F(k, \cdot)$ and $F(k', \cdot)$ must behave independently even if $k$ and $k'$ differ by a single bit.

Theorem 6.5. If $F$ is modeled as an ideal cipher, then the Davies-Meyer construction yields a collision-resistant compression function. Concretely, any attacker making $q < 2^{\ell/2}$ queries to its ideal-cipher oracles finds a collision with probability at most $q^2/(2^{\ell} - 2^{\ell/2})$.

Proof. Here we consider a probabilistic experiment in which for each $k \in \{0,1\}^n$ the function $F(k, \cdot) : \{0,1\}^\ell \rightarrow \{0,1\}^\ell$ is chosen uniformly from the set $\text{Perm}_\ell$ of permutations on $\ell$-bit strings and the attacker is given access to $F$ and $F^{-1}$.

That attacker then tries to find a colliding pair $(k, x), (k', x')$ for which $F(k, x) \oplus x = F(k', x') \oplus x'$.

We assume that: if attacker outputs a colliding pair $(k, x), (k', x')$ then it has previously made the queries to determine $F(k, x)$ and $F(k', x')$; the attacker never make the same query more than once; and once it has learned $y = F(k, x)$ is never queries $F^{-1}(k, y)$ (and vice versa).
**Introduction**

Davies-Meyer Hashes in Practice

The ideal cipher model

Proof continued. The $i$th query $(k_i, x_i)$ to $F$ reveals only hash value

$$h_i \overset{\text{def}}{=} h(k_i, x_i) = F(k_i, x_i) \oplus x_i.$$  

Similarly, The $i$th query to $F^{-1}$ given result $x_i = F^{-1}(k_i, y_i)$ reveals only hash value

$$h_i \overset{\text{def}}{=} h(k_i, x_i) = y_i \oplus F^{-1}(k_i, y_i).$$

The attacker does not obtain a collision unless $h_i = h_j$ for some $i \neq j$.

Fix $i, j$ with $i > j$, and consider the probability that $h_i = h_j$. At the time of the $i$th query, the value of $h_j$ is fixed. Collision occurs only if the attacker queries $(k_i, x_i)$ to $F$ and obtains $F(k_i, x_i) = h_j \oplus x_i$ or queries $(k_i, y_i)$ to $F^{-1}$ and obtains $F^{-1}(k_i, y_i) = h_j \oplus y_i$.

Either event occurs with probability at most $1/(2^\ell - (i - 1))$ and since

$$i \leq q < 2^\ell/2,$$

the probability that $h_i = h_j$ is at most $2/(2^\ell - 2^\ell/2)$.

Taking the union bound over all $\binom{q}{2} < q^2/2$ pairs does the trick.

---

**MD5**

- MD5 is a hash function with 128-bit output length designed in 1991.
- In 2004 a team of Chinese cryptanalysts presented a method for finding collisions.
- Since then there methods have been improved and collisions can be found in under a minute on a desktop PC.
- MD5 should not be used, but is still found in legacy code.
SHA-0, SHA-1, and SHA-2

- The **Secure Hash Algorithm (SHA)** refers to a series of cryptographic hash functions standardized by NIST. The best known, SHA-1, was introduced in 1995.

- So far, no explicit collisions have been found in SHA-1. However, theoretical analysis indicates that they can be found using significantly fewer than $2^{80}$ hash-function evaluations required by the birthday attack.

- Time to move on to SHA-2* which currently does not appear to have these weaknesses.

*Comprised of two related functions: SHA-256 and SHA-512 with 256- and 512-bit outputs lengths.

**Secure Hash Algorithm family**

All hash functions in the SHA family are constructed using the design we have just seen.

- A compression function is defined by applying Davies-Meyer construction to a block cipher.

- This function is then extended to support arbitrary length inputs using Merkle-Damgård.

- The block cipher used here was designed specifically for building the compression function.
SHA-3 (Keccak)

- In 2007 NIST announced a public competition to design a new cryptograph hash function to be called SHA-3. As for AES, competition was completely open and transparent.

- The 51 first-round candidates were narrowed to five finalists in 2010 and subjected to intense scrutiny by the cryptographic community over the next two years.

- NIST announced the selection of Keccak as the winner in October 2012.