CSC3631 Cryptography

Cryptographic Hash Functions

Changyu Dong

Newcastle University

Hash Functions

- A hash function H is a deterministic function that maps a message M of an arbitrary length to a fixed-length (e.g. 128-bit, 160-bit etc.) output $H(M)$.
- Hash is not encryption
  - Keys are not required
  - No way to decrypt
  - Output is fixed-length
- The output is called a hash value or a message digest
- Hash functions used in cryptography is not exactly the same as used in Hash-table.
  - A good cryptographic hash functions must have some security properties

Security properties of Hash Functions

- Collision resistant: Hard to find find M, M' such that H(M)=H(M')
- Weak Collision Resistant (second pre-image resistant): Given M, hard to find M' (different from M) such that $H(M) = H(M')$
- One-way (pre-image resistant): Given the hash value $H(M)$, it is hard to find M' such that $H(M) = H(M')$
- Collision resistant $>$ weak collision resistant $>$ one-way
  - A collision resistant hash function is also weak collision resistant.
  - A weak collision resistant hash function is also one-way
  - but not the other way round
- A hash function is never collision-free since input space is larger than finite output space
- It is just computationally too hard to find such collisions.

Find a Collision: Birthday Paradox

- how many people must I gather in a room in order to have a probability > 0.5 that one of them share the same birthday as me?
  - 253
- How many people must I gather in a room in order to have a probability > 0.5 that two of them share the same birthday?
  - 23
- Why?

Find a Collision: Birthday Paradox

- Think in this way: there are 366 bins, and you randomly throw balls into them one after another.
- Whenever one ball ends up in a bin that has a ball in it, you find a collision
  - $Pr[\text{1st ball falls in an empty bin}] = \frac{366}{366}$
  - $Pr[\text{2nd ball falls in an empty bin}] = \frac{366-1}{366}$
  - $Pr[\text{3nd ball falls in an empty bin}] = \frac{366-2}{366}$
  - so on so forth
- After throw m balls, the probability of none bins has more than 2 balls is

$$p = \prod_{i=1}^{m} Pr[\text{ith ball falls in an empty bin}] = \frac{366 * \cdots * (366 - m + 1)}{366^m}$$

- Then the probability of at least one bin has more than 1 ball (collision) is $1 - p$
- For N balls and M bins, $Pr[collision] \approx \frac{N^2}{2M}$

Find a Collision: Birthday Paradox

- To merely witness a collision is much easier than find a collision to a specific value.
- If a hash function produces n-bit output, then there are 2^n different outputs.
- $Pr[collision] \approx \frac{N^2}{2 \cdot 2^n}$
- However, an attacker only needs $N = 2^{\frac{n}{2}}$ different inputs in order to find a collision (with a probability $\approx 1/2$).
- The output size must be large enough to withstand the birthday attack.
  - Current standard: 256-bit at least
- The attacker should not be able to find a collision better than birthday attack.

Example Application 1: Password Storage

- In Unix-like systems, user passwords are not stored in clear. Instead, the hash values of the password are stored
- When a user tries to login, the hash value of his password is reproduced and compared with the stored value.
- It is hard for an attacker to recover the passwords from the hash values.
  - Make use of the property of one-way: hard for an attacker to inverse the hash values back to the password

Example Application 2: File Integrity

- Many file download sites also provide a hash value of the softwares on the download pages
- After you download a software, you can recompute the hash value
- If this matches the one provided by the website, the file is not corrupted during transmission.

Question

Which property of hash functions is used here?

A Collision resistant: Hard to find M, M' such that H(M)=H(M')

B Weak Collision Resistant: Given M, hard to find M' (different from M) such that $H(M) = H(M')$

C One-way: Given the hash value $H(M)$, it is hard to find the input M

Example Application 3: File Identities

- Hash values are used to identify files on peer-to-peer filesharing networks.
- The files with the same hash values are considered to be copies of the same file, even if they have different names.
- More seeds can be identified.

Question

Which property of hash functions is used here?

A Collision resistant: Hard to find find M, M' such that H(M)=H(M')

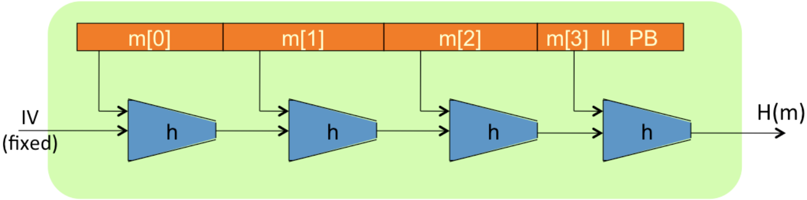
B Weak Collision Resistant: Given M, hard to find M' (different from M) such that $H(M) = H(M')$

C One-way: Given the hash value $H(M)$, it is hard to find the input M

Rainbow Table

- A rainbow table is a precomputed (value, hash) table
- Usually used in attacking password systems which store passwords as hash values
- If an attacker knows the hash value of a password, he can look up in the table to find the matching password.
- Trade space for time
- Can be easily defeated by adding “salt” which is a large random value when hashing passwords
- The attacker has to build a different rainbow table for each salt
- `https://hashtoolkit.com/`
- `https://crackstation.net/`

The Merkle-Damgård Construction



- h: compression function
- Each iteration compute $s_i = h(m[i]||s_{i-1})$, where $s_0 = IV$
- PB: the padding block
- If the compression is collision-resistant, then the hash function is collision resistant

MD5

- 128-bit hash function
- MD5 splits message into 512-bit blocks
- Supposed to require 2^{64} operations to find a collision (birthday attack)
- But more efficient way has been found which requires only 2^{24} operations
- Collisions can be found within seconds
- Should no longer be used

Attacks on MD5

- 2004: Collision can be found by modifying specific bits within two related 512 bit input blocks to create two slightly different messages that have the same hash value.
- Chosen prefix attack: For any prefix, can find colliding messages have this prefix and differ up to 716 random-looking bytes
- More meaningful collision: researchers were able to create pairs of PostScript document with the same hash value, and forge digital certificate.

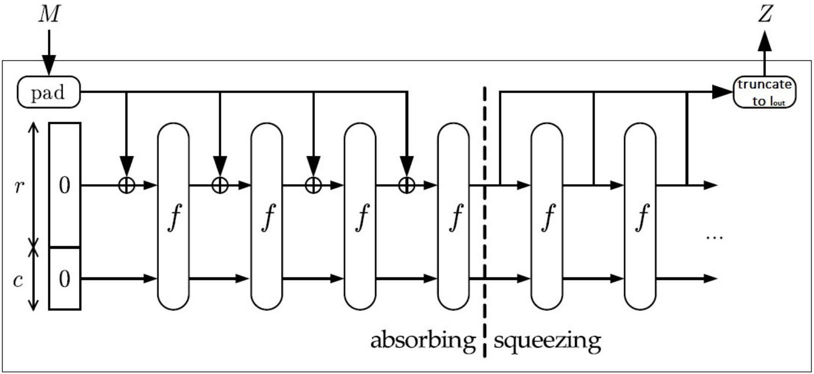
SHA-1

- US standard
- 160-bit
- Also less secure than supposed to be (2^{80} vs 2^{51})
- No longer safe to trust

SHA-224,SHA-256,SHA-384,SHA-512

- SHA-2 family
- The numbers represent the output length
- SHA-224 and SHA-384 are simply truncated versions of SHA-256 and SHA-512, computed with different initial values.
- Slower but no known efficient attack

The Sponge Construction



- New construction used by SHA-3
- *f* is a fixed permutation, message is broken into blocks of *r*-bit
- After absorbing the whole message, squeezing to get your output (*r* bits per round until you get enough bits)

SHA-3

- NIST is having an ongoing competition for SHA-3, the next generation of standard hash algorithms
- 5 candidates finalised
- Result revealed recently: Keccak as the winner
- `http://keccak.noekeon.org/`

Ramdon Oracle

- An idealisation of hash function
  - When using a hash function to construct a cryptographic protocol/scheme, sometimes it is difficult/impossible to prove security even if assuming it is collision resistant.
  - We need a stronger assumption to prove security.
- Therefore we replace hash function with a random oracle in security proofs
  - Random oracle exists only in theory (ideal world), but not practice
  - But we often think hash functions behave close enough to a Random oracle.
- Some people love it, some people hate it
  - Random oracle model vs standard model

Reading

- Cryptography made simple §14.1, 14.2, 14.3,14.4, 14.6, 14.8
- Cryptography Theory and practice: §4.1, 4.2, 4.3
- Applied cryptography §18