April 9, 2000

Hash

A definition

One dictionary states:

hash (noun)
1. a dish of cooked meat cut into small pieces and recooked.
2. a. a mixture; a jumble.
   b. a mess.
3. re-used or recycled material.

More informally ...

From N. Fefferman:

A **hash** is a way of representing information so that each different message has its own unique hash representation, but that representation does not reveal anything informative about the message itself.

A good analogy: human fingerprints. Each person has his own fingerprint. A fingerprint is generally unique to the person (ignoring the idea of identical twins). It is easy to check whether, given a fingerprint and a person, that person has that fingerprint. It is also very easy to figure out given a person, what that person's fingerprint is (just take a print). However, it is very difficult to take a person from a fingerprint (i.e. given a specific fingerprint, it takes a lot of work to find the person it corresponds to).

In computer science

The web page `http://www.rsasecurity.com/rsalabs/faq/` answers the question, **What is a hash function?** with the following:

- A hash function H is a transformation that takes an input m and returns a fixed-size string, which is called the hash value h (that is, $h = H(m)$). ...
- A hash function H is said to be one-way if it is hard to invert, where "hard to invert" means that given a hash value h, it is computationally infeasible to find some input x such that $H(x) = h$.
- If, given a message x, it is computationally infeasible to find a message y not equal to x such that $H(x) = H(y)$ then H is said to be a weakly collision-free hash function.
- A strongly collision-free hash function H is one for which it is computationally infeasible to find any two messages x and y such that $H(x) = H(y)$.

In business

`www.surety.com` advertises "the Internet's premier provider of digital record notarization services," and has an exceptionally long legal page. The company provides digital notary and timestamp services. The following quotes are from various Surety web pages:

> Digital Records are the heart of your business. From e-commerce transactions with customers and suppliers to the accumulated intellectual property of your R&D department, you depend on the integrity of your corporate data. Can

you establish beyond reproach that vital digital records have not been tampered with or backdated? Can you prove it in a court of law?

. . .

Digital Fingerprints are one-way identifiers; no part of the original data record or file can be reconstructed from its fingerprint. Every time you notarize a digital object, Digital Notary Service issues you a unique Notary Record. With that record and the original data, you can validate a notarization at any time in the future.

Your document's fingerprint, like your own, is unique and impossible to forge, by anyone. And the system has no keys, so it cannot be publicly compromised.

. . .

Digital Notary Service provides a secure, irrefutable audit trail for all your digital records. Whether they are business transactions, lab data, video/audio recordings, or data used for government regulatory compliance, Surety's service seals the contents of anything in digital form at a precise moment in time. It is impossible for anyone, even Surety, to alter even a single bit of the data or change its timestamp in any way without detection. Surety accomplishes this without exposing important records to anyone outside your company.

. . .

Using the Digital Notary Record Authentication software on the Customer's computer, Customer hashes an electronic record ("document") into a unique 288-bit message digest ("hash digest") uniquely representing that document.

Problems?

The point is to have a hash function which can be computed quickly and which also avoids collisions. One current candidate for such a function is MD5. MD5 abbreviates "**M**essage **D**igest 5", which makes one wonder what happened to MD4, MD3, etc.* In fact, the web page `http://www.math.ohio-state.edu/~fiedorow/PGP/MD5_discussion` contains the following paragraph:

> In February 1996 my paper "Cryptanalysis of MD4" appeared . . . In this paper, as an example two versions of a contract are given with the same MD4 hash value. Alf sells his house to Ann, in the first version the price is $176,495 and in the second it is $276,495. The contracts have been prepared by Alf. Now if Ann signs the first version with $176,495 then Alf can alter the price to $276.495 ... In principle this risk occurs, if you use a hash function for which (senseful) collisions can be found, whenever you allow another person to have influence on the contents of a document you are signing. Certainly this does not happen very often in practical applications. But sometimes you *must* have an agreement about a text (contract) which is then signed by two or more parties. And these are often just the most important applications!

* The company mentioned previously uses MD5 together with another hashing algorithm recommended by NIST, the major U.S. government standards organization.