Fingerprinting

Suppose we just finished transmitting an enormous file to the Moon. We'd like to verify that our file got there correctly, without any errors. We could of course re-send the file a second time, but bandwidth to the Moon is hard to come by, so this would take a long time. It'd be nice to have a better solution.

So, the problem is this: if we have a n-bit bitstring s on Earth, and a n-bit bitstring s' on the Moon, how can we efficiently verify that $s = s'$? We will measure the efficiency of a solution by the number of bits that need to be sent between the Earth and Moon to finish this task. We will assume that the Earth and Moon have a reliable but low-bandwidth channel that they can use to communicate.

The approach

The gist of our idea is for the Earthlings to compute a short *fingerprint* of their data, and then send that fingerprint to the Moon. The Moon can compute the fingerprint of their own copy of the data and check that it matches the fingerprint received from the Earth. We will denote the fingerprint of data s by $F(s)$, so the Earth computes $F(s)$ and sends $F(s)$ to the Moon; the Moon computes $F(s')$; and then the Moon checks whether $F(s) = F(s')$. If there is a mismatch in the fingerprints ($F(s) \neq F(s')$), then we can conclude that the two endpoints do not have the same data ($s \neq s'$).

If the two fingerprints *do* match ($F(s) = F(s')$), what can we conclude? In this case, there is no guarantee that the two endpoints definitely have the same data. Since we want the fingerprint $F(s)$ to be shorter than the data s, the function $F(\cdot)$ cannot possibly be one-to-one (injective), so it is possible that $s \neq s'$ but $F(s) = F(s')$. This sounds bad, but we will show how to make it very unlikely that we encounter this bad case.

The idea is to use a probabilistic fingerprint, where the fingerprint also depends upon some random value r. Now our fingerprint becomes $F(s,r)$. We will show that, if there is any mismatch between the two files at the Earth and Moon, then with probability very close to 1, the fingerprints will be different. (In other words, if $s \neq s'$, then we have $F(s,r) \neq F(s',r)$ with very high probability—or equivalently, if $s \neq s'$, there is only a very small chance that we choose a random value r such that $F(s,r) = F(s',r)$.) This means that if the Earth and Moon do not have the same data, then this fact will almost certainly be detected by our fingerprinting scheme.

A scheme based upon polynomials

We will pick a prime q, and we will work modulo q. As we shall see, the prime q does not need to be terribly large, but we will assume that $q \geq 2^{100}n$. For instance, q might be the smallest prime larger than $2^{100}n$, so that the number q is approximately $100 + \lg n$ bits in length.

We will view the data s as a polynomial $P(x)$ whose coefficients are numbers modulo q. For example, if $s = \langle s_1, s_2, \ldots, s_n \rangle$, where s_i is the i-th bit of s, then we might use the following polynomial:

$$P(x) = s_n x^{n-1} + \cdots + s_2 x + s_1.$$

The exact details of how we encode our data s as a polynomial $P(x)$ aren't terribly important, and we could have used some other encoding. The important thing is that the degree of $P(x)$ is known to be at most n (in this case, at most $n-1$, in fact).

Now the fingerprint $F(s,r)$ will be given by

$$F(s,r) = P(r) \bmod q.$$

To verify that the Earth and Moon have the same data, the Earth picks a random number r from $\{0,1,2,\ldots,q-1\}$, computes $F(s,r)$, and then sends $F(s,r)$ to the Moon. The Moon receives r, computes $F(s',r)$, and then checks whether $F(s,r) = F(s',r)$.

Note that $0 \leq r < q$ and $0 \leq F(s,r) < q$, so the Earth only has to send $2\lg q$ bits to the Moon, or about $200 + 2\lg n$ bits in total. This is much more efficient than re-sending the entire file s to the Moon (which would require sending n bits; when n is large, $200 + 2\lg n$ bits is much less than n bits).

Analysis

The only way this can go wrong is if the Earth and Moon have two different files $s \neq s'$, but due to bad luck, the Earth happens to pick a random value r such that the fingerprints match ($F(s,r) = F(s',r)$). What are the chances of this?

We can show that this is very unlikely. Fix s, s' such that $s \neq s'$. Call a value r *unlucky* if we have $F(s,r) = F(s',r)$. Let's count how many values of r are unlucky. The condition $F(s,r) = F(s',r)$ is equivalent to the condition

$$P(r) \equiv P'(r) \pmod q,$$

where $P(x)$ is the polynomial corresponding to s and $P'(x)$ is the polynomial corresponding to s'. This condition is in turn equivalent to the condition

$$Q(r) \equiv 0 \pmod q,$$

where $Q(x)$ is the polynomial $Q(x) = P(x) - P'(x)$. We know that the polynomials $P(x), P'(x)$ have degree at most n, so the same is true of $Q(x)$. And we also know that any polynomial of degree at most n has at most n roots modulo any prime. (See Property 1 from Lecture Note 10.) This means that $Q(x)$ has at most n roots modulo q, or equivalently, that there are at most n values of r such that $Q(r) \equiv 0 \pmod q$.

In summary, we have shown that there are at most n unlucky values of r. However, Earth chooses r randomly from among a total of q possible values. This means that the probability of choosing an unlucky value of r is at most n/q. Since $q \geq 2^{100}n$, this means that the chance of choosing an unlucky r is at most $1/2^{100}$.

This is an incredibly small chance. For comparison, the chances of being struck by lightning during any given year is reportedly about $1/700,000$ ($\approx 1/2^{19}$), and the chances of being killed by lightning during any given year is about $1/3,000,000$ ($\approx 1/2^{22}$), in the US. The chances of winning a lottery where you must pick 6 numbers out of 49 is about $1/14,000,000$ ($\approx 1/2^{24}$). So the odds of failing to detect a mismatch in the Earth-Moon data ($1/2^{100}$) are smaller than the chances of buying four lottery tickets in four consecutive weeks and winning the jackpot in every single one ($\approx 1/2^{96}$)! In short, a probability like $1/2^{100}$ is so incredibly small that, for all practical purposes, it is indistinguishable from zero. For most engineering purposes, we can safely ignore the possibility of events with such a small probability.

In conclusion, we have an efficient way of confirming that the Earth and Moon share an identical copy of a large file. If the Earth and Moon do indeed share identical copies of the data, then the fingerprint will

confirm this. And if there is any mismatch between the Earth and Moon's data, the fingerprint is almost guaranteed to detect this—there is only a $1/2^{100}$ chance that the fingerprint fails to detect this.

Example: We have sent 1GB of data to the Moon ($n = 2^{33}$ bits). We want to confirm that the data got there correctly, with no errors. So, we use a 133-bit prime q. We pick a random value r and send the fingerprint $F(s,r)$ to the Moon. In total, we send 266 bits to the Moon—much more efficient than re-sending all 8,000,000,000 bits a second time.