Reading 001: "Verifying Distributed Erasure-Coded Data"

J. Hendricks, G. Ganger, & M. K. Reiter. "Verifying distributed erasure-coded data," in *Proceedings of the 26th Annual ACM Symposium on Principles of Distributed Computing (PODC)*, R. Wattenhofer, Chair, Portland, OR, USA, Aug. 12-15, 2007, pp. 139-146. doi: 10.1145/1281100.1281122

Notes

- "An *m-of-n* erasure code encodes a block of data into n frements, each $1/m^{th}$ the size of the original block, such that any m can be used to reconstruct the original block. Thus, $(n-m)$ of the fragments can be unavailable (e.g., due to corruption or server failure) without loss of access."
  - Concrete example: a 3-of-5 erasure code encodes a block of data into 5 fragments, each 1/3 the size of the original block, such that any 3 can be used to reconstruct the original block. Thus, 2 of the fragments can be unavailable without loss of access to the original data.
  - The difficulty with erasure codes: ***any*** randomly chosen subset of m fragments can be used to reconstruct the original block.
    1. Each fragment is much smaller than the original ($1/m$ the size)
    2. Any m fragments collectively contain all the information needed
    3. Fewer than m fragments are insufficient to reconstruct the data

- "Unfortunately, erasure coding creates a fundamental challenge: determining if a given fragment indeed corresponds to a specific original block. If this is not ensured for each fragment, then reconstructing from different subsets of fragments may result in different blocks, violating any reasonable definition of data consistency."
  - I hadn't even considered this – but it makes sense. If the fragments are stored on different machines, an adversary could tamper with a fragment and introduce malicious data.
  - Solution introduced in this paper: fingerprints. My understanding at this point is that the fingerprint acts as a key of sorts, which can be used to verify the original data block when reconstructing from fragments.
- Let $\mathbb{F}_{q^k}[x]$ denote the set of polynomials with cofficients in $\mathbb{F}_{q^k}$, with "+" and "$\cdot$" defined as in normal polynomial arrithmetic.
  - Translation: $\mathbb{F}_{q^k}[x]$ is the set of all polynomials of shape $c_n \cdot x^n + c_{n-1} \cdot x^{n-1} + \cdots + c_1 \cdot x^1 + c_0$, where $c_i \in \mathbb{F}_{q^k}$ and x represents the variables, and q^k represents the size of the finite field $\mathbb{F}_{q^k}$:
    * q is a prime number
    * k is some positive integer
    * q^k tells you how many elements are in $\mathbb{F}_{q^k}$
- A vector $d \in \mathbb{F}_{q^k}^{\delta}$ of δ elements of $\mathbb{F}_{q^k}$ has a natural representation as a polynomial $d(x) \in \mathbb{F}_{q^k}[x]$ of degree less than δ with coefficients in $\mathbb{F}_{q^k}$...
  - d = A vector of δ elements. If $\delta = 5$, then $d = (d_0, d_1, d_2, d_3, d_4)$, where each $d_i \in \mathbb{F}_{q^k}$
  - $d(x)$ = A polynomial of degree less than δ (largest exponent is $\delta - 1$) with coefficients in $\mathbb{F}_{q^k}$
  - $d(x) = d_0 + d_1 \cdot x + d_2 \cdot x^2 + \cdots + d_{\delta-1} \cdot x^{\delta-1}$

- Let $\mathbb{F}_2$ denote a field of order 2, let $K = \{2, 3, 4, \ldots, 2^\gamma\}$,
  - $\mathbb{F}_2 = \{0, 1\}$ - a binary field

- K = a set of integer labels, where γ is a positive integer.

- Division fingerprinting:
  - $\mathbb{F}_{q^k}$ = a field of finite values of order (size) q^k
  - $|K|$ = the number of monic irreducible polynomials of prime degree γ with coefficients in $\mathbb{F}_{q^k}$
    - "monic" = the leading coefficient is 1
    - "irreducible" = cannot be factored into the product of two non-constant polynomials with coefficients in $\mathbb{F}_{q^k}$
    - "prime degree" = the degree of the polynomial is a prime number (e.g., 2, 3, 5, 7, etc.)
    - Note on γ and max degree: The degree of the polynomial is at most $\gamma - 1$
    - Example: For $\mathbb{F}_2$, $\gamma = 3$: $x^2 + x + 1$
  - $P_{q^k} : K \to \mathbb{F}_{q^k}[x]$ = a deterministic algorith that outputs monic irreducible polynomials of prime degree γ with coefficients in $\mathbb{F}_{q^k}$ chosen uniformly at random from K, with probabilities taken over the choice of input $r \in K$ uniformly at random
    - A function that maps each element of K to a polynomial in $\mathbb{F}_{q^k}[x]$
    - The probability of a coefficient's value is determined by the choice of r uniformly at random from K
    - The function is deterministic, meaning that for a given input r, it will always produce the same output polynomial. However, since r is chosen uniformly at random from K, the output polynomial can be considered random when viewed over the randomness of r.
  - $fp(r, d) : K \times \mathbb{F}_{q^k}^{\delta} \to \mathbb{F}_{q^k}^{\gamma}$
    - r = an element of K (a random seed)
    - d = a vector of δ elements of $\mathbb{F}_{q^k}$, which can be represented as a polynomial $d(x)$

$$\begin{aligned} fp(r, d(x)) : p(x) &\leftarrow P_{q^k}(r); \\ &\text{return } (d(x) \mod p(x)) \end{aligned}$$

    - $d(x)$ = the polynomial representation of the data vector d
    - $p(x)$ = a monic irreducible polynomial of prime degree γ with coefficients in $\mathbb{F}_{q^k}$, determined by the input r through the function P_{q^k}
    - $fp(r, d)$ = the fingerprint of d with respect to the random seed r, computed as the remainder of $d(x)$ divided by $p(x)$

  - Chance of collision:

$$\begin{aligned} \varepsilon &= \frac{\delta}{q^{k\gamma} - q^{\frac{k\gamma}{2}}} \\ &\approx \frac{\delta}{q^{k\gamma}} \text{ for sufficiently large } \gamma \end{aligned}$$

- Evaluation fingerprinting:
  - $\mathbb{E}_{q^{k\gamma}} = \dfrac{\mathbb{F}_{q^k}[x]}{p(x)}$
    - A field of polynomials with coefficients in $\mathbb{F}_{q^k}$ of degree less than γ, with "$\cdot$" defined modulo $p(x)$
    - $p(x)$ = a constant monic degree-γ irreducible polynomical with coefficients in $\mathbb{F}_{q^k}$
  - $K = \{0, \ldots, q^{k\gamma} - 1\}$
    - The set of integer labels, where γ is a positive integer and $q^{k\gamma}$ is the size of the field $\mathbb{E}_{q^{k\gamma}}$

- $S : K \to \mathbb{E}_{q^{k\gamma}}$
  - A deterministic algorhtm that outputs an element belonging to $\mathbb{E}_{q^{k\gamma}}$ chosen uniformly at random, with probabilities taken over the choice of input $r \in K$ uniformly at random
  - A mapping of each element of K to a polynomial in $\mathbb{E}_{q^{k\gamma}}$
- $fp(r, d) : K \times \mathbb{F}^{\delta}_{q^k} \to \mathbb{F}^{\gamma}_{q^k}$
  - δ = the number of elements in the data vector d
  - γ = the number of elements in the fingerprint vector $fp(r, d)$
  - r = an element of K (a random seed)

$$fp(r, d(y, x)) : s(x) \leftarrow S(r); \text{ return } d(s(x), x)$$

  - $d(y, x)$ = the bivariate polynomial representation of the data vector d
  - $s(x)$ = an element of $\mathbb{E}_{q^{k\gamma}}$ determined by the input r through the function S
  - $fp(r, d)$ = the fingerprint of d with respect to the random seed r, computed as the evaluation of the bivariate polynomial $d(y, x)$ at $y = s(x)$
  - $d(s(x), x)$ = the fingerprint; returns a univariate polynomial in x of degree less than γ, which can be represented as a vector of γ elements of $\mathbb{F}_{q^k}$
- Chance of collision:

$$\varepsilon = \frac{\delta/\gamma}{q^{k\gamma}}$$

  - Gets smaller as γ increases

- Both division and evaluation fingerprinting are homomorphic, meaning that the fingerprint of the sum of two data vectors is equal to the sum of their fingerprints:
  - $r \in K$, $d, d' \in \mathbb{F}^{\delta}$, and $b \in \mathbb{F}$
  - Additive homomorphism:
    - $fp(r, d) + fp(r, d') = fp(r, d + d')$
    - "If you fingerprint two pieces of data separately and add the fingerprints together, you get the same result as if you had first added the data together and then fingerprinted the sum."
  - Multiplicative homomorphism:
    - $b \cdot fp(r, d) = fp(r, b \cdot d)$
    - "If you fingerprint data and then multiply the fingerprint by some constant, you get the same result as if you had first multiplied the data by that constant and then fingerprinted it."

- Cross-checksum fingerprinting
  - Cross checksum: An array containing a hash of each fragment
  - Old way:
    - Send fragment, checksum
    - To verify that all fragments came from the same block: reconstruct block; re-encode all n fragments; hash each one and compare cross-checksum
    - Very computationally expensive
  - What this paper proposes:
    - fpcc.cc array of size n - n fragments, n cross-checksums

- fpcc.fp array of size m - m fingerprints, able to reconstruct the original block from m fragments
- Ensuring fragment d_i:
  1. hash(d_i) = fpcc.cc[i]?
  2. fp(r, d_i) = encode(fpcc.fp[1], ..., fpcc.fp[m])?
- Even if a malicious adversary tries really hard (e.g., makes $\mathcal{X}$ queries, tries different fragment combinations), the probability they can create fragments that (1) all pass verification (look valid) but (2) reconstruct to different blocks ($B \neq B'$) is astronomically small.

Questions

1. What is an erasure code?

2. "An *ε-fingerprinting function* $fp : K \times \mathbb{F}^{\delta} \to \mathbb{F}^{\gamma}$ satisfies"

$$\max_{\substack{d,d' \in \mathbb{F}^{\delta} \\ d \neq d'}} \Pr\left[fp(r,d) = fp(r,d') : r \xleftarrow{\mathrm{R}} K\right] \leq \varepsilon$$

   - Can you translate fp into English?
   - What is K?
   - What is $\mathbb{F}^{\delta}$?
   - What is $\mathbb{F}^{\gamma}$?

3. "...a deterministic algorithm that outputs monic irreducible polynomials of prime degree γ..."
   - What is a monic irreducible polynomial?
   - What does γ represent? What is a "prime degree"?

Potential Further Reading

1. Source [26]: Reed-Solomon codes
2. Source [25]: Rabin's information dispersal algorithm

BOOKMARK: p. 4, Section 2.2