A MATHEMATICAL PROLOGUE TO THE STUDY OF ELECTRONIC ENCRYPTION

NTACC 2010

Eric Bach
Computer Sciences Department
University of Wisconsin
Madison, WI 53706

Over a period of 30 years (1945-1975), encryption underwent a *technological shift*:

> Before: electromechanical devices (e.g. Enigma)
>
> After: all-electronic, microprocessor based (e.g. DES).

What happened in between? To study this properly, we need some mathematical background, including facts about

> Shannon's Information Theory
>
> Pseudo-Random Number Generation
>
> Nonlinear Dynamical Systems
>
> Correlation Analysis

These will be the topic of the present lecture. The next lecture will cover how these ideas were realized.

Shannon's Theory

Goal: Embed cryptography and cryptanalysis into information theory

Three random variables:

P is the message

K is the key

C is the cryptogram

We'll assume:

i) P, K are independent;

ii) C is determined by P and K;

iii) P is determined by C and K.

The system has *perfect secrecy* if P and C are independent.

Idea: the analyst can learn nothing about the message by observing a cryptogram.

We can still define this in more complicated models (in which i)–iii) don't hold), but then we replace independence by relative entropy.

Perfect secrecy is a very strong condition, it implies:

> # of possible keys $\geq$ # possible plaintexts
>
> For a "minimal" realization ($|K| = |C|$), the system must be a *Latin Square*:
>
> > Example: $P = C = \{a, b, c\}$. Encrypt by rotating mod k, where k is chosen at random.
> >
> > The complete system is
> >
> > | | 0 | 1 | 2 |
> > |---|---|---|---|
> > | a | a | b | c |
> > | b | b | c | a |
> > | c | c | a | b |
> >
> > The inner 3×3 array is a table for addition mod 3. Arrays of this type were well known by the 1500's (Vigenère).
>
> P can have any distribution.
>
> It is an open problem to enumerate Latin squares, but there are plenty to go around:
>
> $$L_{10} \approx 7.580 \times 10^{24}$$
>
> (after normalizing first row/column).

One-Time Pads

Plaintext:

$$x_1 x_2 x_3 \ldots x_n$$

Random key stream:

$$k_1 k_2 k_3 \ldots k_n$$

Encrypted message:

$$y_1 y_2 y_3 \ldots y_n$$

where $y_i = x_i + k_i \bmod N$.

Receiver recovers PT by subtracting off the key stream

Nowadays $N = 2$, but other bases have been used.

The Latin Square here is the addition table for the abelian group

$$\mathbf{Z}/(N) \oplus \cdots \oplus \mathbf{Z}/(N).$$

Achieves perfect secrecy if the key is random (Maubourgne – 1920's, before Shannon!).

One-Time Pad (cont'd.)

Secure, if no key is re-used.

Your key material must be long enough to cover all messages (until you get back home).

This was apparently done, at great cost. Examples:

US-UK phone link (1945): added/subtracted analog sound signals, key material was on a vinyl phonograph record (!).

Numerical pads for Soviet field agents during the Cold War.

More efficient version:

Use a pseudo-random key

Must only transmit a short *seed*

Proposed by Vernam (before Maubourgne): two key tapes of length $N, N-1$, cycle through both and XOR to get key of length $\sim N^2$.

Vernam achieved $|\text{seed}| \approx \sqrt{|\text{message}|}$.

Pseudo-Random Numbers (in General)

Choose a finite set X.

Choose a function $f : X \to X$.

Make the sequence

$$x, f(x), f(f(x)), \ldots, f^{(i)}(x), \ldots$$

Because $|X| < \infty$, this is ultimately periodic.

It is common to choose the *seed* x to be already in a cycle.

We can output either the iterates themselves, or transform them first:

$$g(x), g(f(x)), g(f^{(2)}(x)), \ldots, g(f^{(i)}(x)), \ldots$$

Affine-Linear Iterations

X will be a free $\mathbf{Z}/(n)$-module of rank ν.

$$X = \mathbf{Z}/(n) \oplus \mathbf{Z}/(n) \oplus \cdots \oplus \mathbf{Z}/(n).$$

If we write elements of X as column vectors, the iteration function is

$$f(x) = Ax + b.$$

The map is reversible (1-1) iff

$$\det A \in \mathbf{Z}/(n)^*.$$

One can show the probability of this is

$$\Theta(\varphi(n)/n) = \Omega((\log\log n)^{-1}).$$

Proof: Use Chinese Remainder theorem and Landsberg's formula to count matrices that are invertible mod p.

Two Classic Examples

Multiplicative Congruential Generator (Lehmer)

Take $\nu = 1$, so

$$f(x) = \alpha x + \beta, \qquad \alpha \in \mathbf{Z}/(n)^*.$$

Example: $f(x) = 2x + 1$ (double and add):

$$1, 3, 7, 15, 31, 63, 127, 255, \ldots$$

Mod 9 this becomes

$$1, 3, 7, 6, 4, 0, 1, 3, 7, \ldots$$

Shift Register Sequences (Tausworthe)

Take $n = 2$ and $b = 0$.

Choose a linear recurrence relation

$$z_i = c_{\nu-1} z_{i-1} + \ldots + c_0 z_{i-\nu},$$

where $c_i \in \mathbf{Z}/(2)$.

This induces a linear map on X, the consecutive ν-bit blocks.

Example: $z_i = z_{i-1} + z_{i-2}$ (Fibonacci mod 2):

$$1, 1, 0, 1, 1, 0, \ldots$$

Successive blocks are

$$11, 10, 01, 11, 10, 01, \ldots$$

Criteria for Maximal Periods

Lehmer Sequence $f(x) = \alpha x + \beta \bmod n$

This is a polynomial, so we can factor the modulus

$$n = \prod_i p_i^{e_i}$$

and work with the local maps mod $p_i^{e_i}$.

Period mod n = LCM of local periods.

Theorem: If $\beta \in \mathbf{Z}/(p)^*$ and α is close to 1 p-adically, the local period will be $p_i^{e_i}$.

Close to 1 means that

$$\alpha \equiv \begin{cases} 1 \bmod 4, & \text{if } p_i = 2 \text{ and } e_i \geq 2; \\ 1 \bmod p, & \text{otherwise.} \end{cases}$$

Criteria for Maximal Periods (Continued)

Tausworthe (LFSR) Sequence

The linear recurrence

$$z_i = c_{\nu-1} z_{i-1} + \ldots + c_0 z_{i-\nu}$$

has the characteristic polynomial

$$f(Z) = Z^\nu + c_{\nu-1} Z^{\nu-1} + \ldots + c_1 Z + c_0.$$

The zero block 00...00 is a fixed point.

If Z generates $\mathbf{F}_2[Z]/(f)^*$, the nonzero blocks form a cycle of length $2^\nu - 1$.

Proof of the Maximal Period Criterion (and More)

Take a polynomial, e.g.

$$f(Z) = Z^3 + Z + 1$$

and form its companion matrix M_f.

Multiply by M_f on one side to get a shift register:

$$(a \quad b \quad c) \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = (b \quad c \quad a+b)$$

Multiply by M_f on the other side to compute $Z(\alpha Z^2 + \beta Z + \gamma)$ by mod f:

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \gamma \\ \beta \\ \alpha \end{pmatrix} = \begin{pmatrix} \alpha \\ \alpha+\gamma \\ \beta \end{pmatrix}$$

Since M_f and M_f^T are similar, the dynamics will be the same.

Bonus: the "translation" (conjugation) between the two systems is linear.

General Linear and Affine Recurrences

Mod-2 case studied systematically in the 1960s, by engineers.

Main Tools:

Direct sum decomposition of state space:

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$$

On each V_i, the operator A can be realized as a companion matrix:

$$A = \begin{pmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ & & \vdots & 0 & \\ 0 & 0 & \cdots & 1 & -c_k \end{pmatrix}$$

This is the action of a Fibonacci-style linear feedback shift register.

We can treat an affine recurrence $Ax + b$ as a linear one ($b = 0$) by adding one more state variable.

The Distance Problem (nowadays called Discrete Logarithm)

We know the starting point x, and can observe a "downstream" point $x^{(t)} = f^{(t)}(x)$. What is t?

For Lehmer sequences

$$f(x) = \alpha x + \beta \bmod n$$

we can evaluate

$$x^{(t)} = x\alpha^t + \beta[\alpha^{t-1} + \cdots + \alpha + 1]$$

in $O(\log t)$ iterations.

Proof: use projective coordinates, e.g.

$$\begin{pmatrix} \alpha x + \beta \\ 1 \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}$$

and matrix exponentiation.

For maximum-period sequences, it is enough to find $\mathrm{dist}(0, x^{(t)})$

The Distance Problem (cont'd.)

Here is a collision algorithm that will beat exhaustive search:

To find $\text{dist}(0, x^{(t)})$:

Choose $m \approx \sqrt{\text{period}}$ and make m "giant steps":

$$0, 0^{(m)}, 0^{(2m)}, 0^{(3m)}, \ldots$$

And then m "baby steps"

$$x^{(t)}, x^{(t+1)}, x^{(t+2)}, \ldots$$

Sort the two lists and look for a match.

When can we use Baby-Step Giant-Step?

It works any time we have "random access:" a quick method for

$$x \mapsto x^{(t)}$$

Examples:

> Fibonacci-style LFSR sequences have same dynamics as multiplying by Z in $R = \mathbf{F}_2[Z]/(f)^*$. So solve $a^t = b$ in R.
>
> Shift registers for polynomial multiplication (= Galois configuration).
>
> Class numbers of imaginary quadratic fields (Shanks, 1969).

When was it invented?

> The term "giant step" dates back to 1964 (Weiss).

Two Early Discrete Log Applications

Discrete Logs for Planetary Ranging (Golomb)

In the late 1950's, radar became powerful enough to bounce off planets.

How far away is Venus?

We know the speed of light, so it is enough to measure forward + return transit time.

To measure time, design a signal that is modulated with a pseudorandom sequence, with an *easy* distance problem.

Golomb (1961) used a linear combination of several LFSR sequences.

Fast Counters (Clark)

We want to add a "counter" to our computer to detect how many times an instruction has executed.

Use a maximum-length LFSR for which $2^\nu - 1$ is highly composite.

The number of clock ticks is a discrete log, which can be computed using the Chinese Remainder Theorem and baby-step giant-step.

Additive Stream Ciphers

Suppose $x_0, x_1, x_2, \ldots$ is a pseudo-random sequence.

We can use this as a running key stream:

$$\begin{aligned} \text{Plain Text} &\quad p_0\ p_1\ p_2\ p_3\ \cdots \\ + \text{ Key} &\quad x_0\ x_1\ x_2\ x_3\ \cdots \\ = \text{Cipher Text} &\quad c_0\ c_1\ c_2\ c_3\ \cdots \end{aligned}$$

The receiver generates x_i and subtracts to get the plain text.

For Security, Long Periods are Necessary But Not Sufficient.

Messages often have stylized parts (e.g. e-mail headers).

The cryptanalyst can guess an initial piece of plaintext, subtract it to get a piece of the key stream, and try to predict the rest.

Prediction is an NP problem (guess the seed), and we want it to be hard.

Predicting Affine-Linear Sequences

We are given

$$x_0, x_1, \ldots, x_m$$

where $x_i = \alpha x_{i-1} + \beta \bmod n$, but α, β, n are unknown.

Boyar's Poly Time Predictor

Work with the differences

$$y_i = x_i - x_{i-1}$$

to eliminate β. Then

$$y_i \equiv \alpha y_{i-1}.$$

Easy case: when y_i are "random" we can solve

$$\sum_{i=1}^{t} u_i y_i = 1$$

for integers u_i, then shift up:

$$\sum_{i=1}^{t} u_i y_{i+1} = \alpha.$$

We expect $n = \gcd\{y_i - \alpha y_{i-1}\}$.

General case: get a sequence of coefficients and moduli that converges to "correct" values.

Predicting Tausworthe (Shift Register) Sequences

Naive algorithm:

Each pair

$$x' = Ax$$

is a linear equation for the entries of A.

A has ν^2 entries. We can solve for these using $O(\nu^6)$ bops.

We actually have a sparse, structured linear system, so $O(\nu^3)$ is enough. (Peterson)

Sophisticated algorithm:

If the x_i satisfy a linear recurrence relation, then

$$r(Z) = x_0 + x_1 Z^{-1} + x_2 Z^{-2} + \cdots$$

is in $\mathbf{F}_2(Z)$ (rational functions).

Use the continued fraction algorithm or Berlekamp-Massey to find the denominator of r.

Coefficients of the denominator give the recurrence relation.

Cost: $O(\nu^2)$ bops.

Since linear and affine recurrences are predictable, we should study general function iterations.

The Associated Graph

Vertices are elements of X.

There is a directed edge $x \to y$ whenever $y = f(x)$.

$$\begin{array}{ccccccc} & & & & e & & \\ & & & & \downarrow & & \\ a & \to & b & \to & c & \to & \cdots \\ & & & & \uparrow & & \\ & & & & d & & \end{array}$$

Features we can study include:

Number of fixed points

Number of components

Period lengths

In-degrees (number of predecessors)

Random Mappings

Let $|X| = n$. Choose one of the n^n functions on X uniformly at random.

Distribution of various graph parameters studied systematically by Harris (1960) and successors.

Fixed Points

Values of $f(x)$ are independent, so

number of fixed points $\sim$ Binomial$(n, 1/n)$.

On average, one point is fixed.

Poisson limit:

$$\Pr[\ k \text{ fixed points }] = \frac{e^{-1}}{k!}.$$

Predecessors

Model by throwing n balls into n bins.

The chance a given bin is empty is

$$\left(1 - \frac{1}{n}\right)^n \sim e^{-1}.$$

So on average, n/e elements are orphans (no predecessor).

Random Mappings With Constraints

If f is 1-1 (or onto) we have a random permutation.

The properties of components are determined by the (discrete) random splitting process.

All components are cycles.

Mean number of cycles $\sim \log n$.

Mean time until repetition is $n/2$.

Mean number of fixed points is 1.

Quick proof: Expectation is linear, so

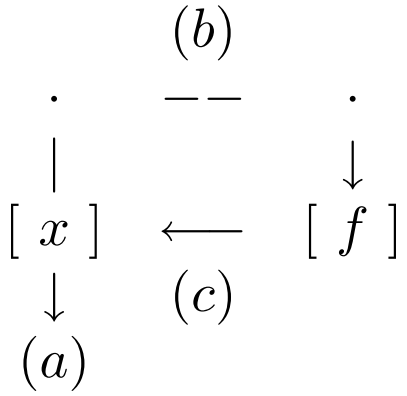
$$E[\text{ \# of fixed points }] = \sum_i \Pr[\ i \text{ is fixed }]$$

$$= n \cdot \frac{1}{n} = 1.$$

This is the combinatorial "hat check" problem.

Nonlinear Maps as One-Cell Shift Registers

Suppose $f : X \to X$. If we put this in a feedback loop, holding x in a memory cell, we have:



Suppose X is an abelian group. We can add plaintext at three places:

a) Ordinary stream cipher

b) Chaining

c) Cipher feedback

These are today's "operation modes" for block ciphers.

Longer Shift Registers

Equivalent to

$$z_n = f(z_{n-1}, \ldots, z_{n-\nu})$$

Example: Wolfram's Rule 30

$$z_n = z_{n-1}z_{n-2} + z_{n-1} + z_{n-2} + z_{n-3}$$

has

5 cycle	111 011 001 100 110
2 cycle	101 010
fixed point	000

Good and Bad Features:

Easy to build

Cycle length difficult to predict

For Nonlinear FSR's, Cycle Length Prediction is NP-hard

We want to test ϕ for satisfiability.

Assume $\phi(0, \ldots, 0) = 0$ (if not, we're done).

Choose a primitive polynomial

$$f(Z) = Z^{\nu} + c_{\nu-1}Z^{\nu-1} + \ldots + c_1 Z + c_0.$$

Consider the nonlinear recurrence

$$\begin{aligned} z_i = & [c_{\nu-1}z_{i-1} + \ldots + c_0 z_{i-\nu}] \\ & \text{AND} \\ & [\sim \phi(z_{i-1}, \ldots, z_{i-\nu+1})] \end{aligned}$$

$\phi \notin$ SAT: LFSR with max cycle length $2^{\nu} - 1$.

$\phi \in$ SAT: second factor is 0 whenever $\phi(z) = 1$. So

$$z0, z1$$

have same successor. From any nonzero block, length of the ultimate cycle is $< 2^{\nu} - 1$.

When is a General Shift Register Reversible?

We'd like reversibility because average cycle length becomes longer ($N/2$ vs. $\sqrt{N}$).

Consider the recurrence

$$z_n = f(\underbrace{z_{n-1}, \ldots, z_{n-\nu+1}}_{\text{call this } w}, \underbrace{z_{n-\nu}}_{x})$$

For each w, the map

$$x \mapsto f(wx)$$

must be onto.

Since $|X| < \infty$, it is a permutation.

We get a map

$$g : X^{\nu-1} \to \Sigma_{|X|}.$$

Special case: Golomb's "eldest bit" rule

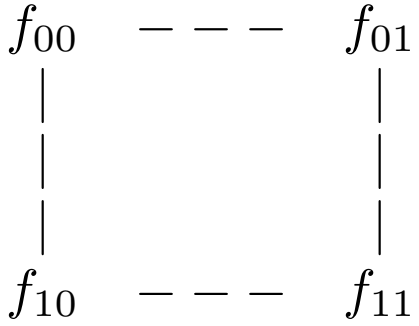
When $X = \{0, 1\}$ there are only two permutations (add 0 or add 1 mod 2).

Theorem: a nonlinear bit SR is reversible iff it has the form

$$z_n = g(z_{n-1}, \ldots, z_{n-\nu+1}) + z_{n-\nu}$$

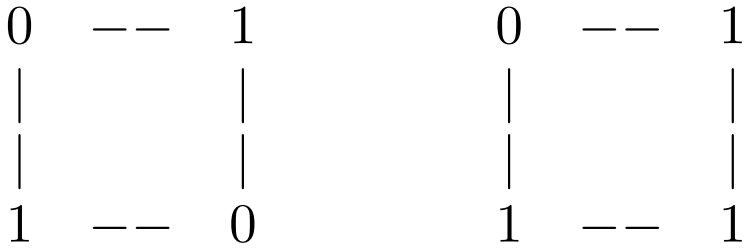
Correlation

A Boolean function $f : \{0,1\}^n \to \{0,1\}$ is equivalent to a placement of 0's and 1's on vertices of the hypercube:



We call f (first-order) *correlation immune* if the center of mass is at the center of the hypercube.

Example: XOR. Non-example: OR.



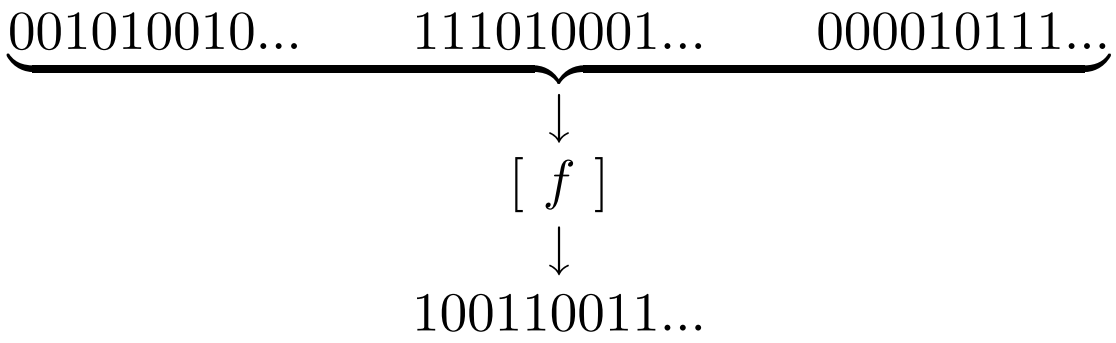
Alternative definition: f is correlation immune when every input is uncorrelated with the output.

Combining Pseudo-Random Streams

Choose n pseudo-random generators with m-bit seeds. Extract one bit from each, and use

$$f(x_1, \ldots, x_n)$$

as a key stream (add mod 2 to plaintext).



The cryptanalyst

- Knows the combiner f, and design of generators
- Can observe keystream
- Seeks initial states of the n generators

Combining Pseudo-Random Streams (Cont'd.)

Brute force solution:

Try all n-tuples of m-bit seeds

Cost: about 2^{mn} trials.

Solution using correlation:

For $i = 1..n$
Replace all but i-th generator
by a random stream
Test seeds for the i-th generator until
its output correlates with keystream

Cost: If f is not CI, about $n2^m$ trials

Conclude: a "combiner" should be unbiased *and* correlation immune. (Golomb, Siegenthaler.)

How Many Balanced Correlation Immune Functions are There?

Asymptotics:

Let

$$D_n := \frac{1}{2}\left(\frac{8}{\pi}\right)^{(n+1)/2} 2^{-n(n+1)/2} \cdot 2^{2^n}.$$

The number of balanced CI functions of n variables is

$$\sim D_n \left(1 - \frac{(n+1)^2}{4 \cdot 2^n} + O(\frac{n^4}{2^{2n}})\right).$$

Numerics:

n	asympt	exact
1	$1.27e00$	0
2	$1.78e00$	2
3	$6.48e00$	8
4	$2.02e02$	222
5	$7.78e05$	807980
6	$9.37e13$	95259103924394
7	$2.33e31$	23478015754788854439497622689296