Algorithms of Information Security: Key generation algorithms

Faculty of Information Technology
Czech Technical University in Prague

September 22, 2020

Pseudorandom bit generator (PRBG)

Definition

A *random bit generator* is a device or algorithm which outputs a sequence of statistical independent and unbiased binary digits.

Definition

A *pseudorandom bit generator (PRBG)* is a deterministic algorithm which,given a truly random sequence of length k, outputs a binary sequence of length $l \geq k$ which "appears" to be a random. The input to the PRBG is called the *seed*, while the output of the PRBG is called a *pseudorandom bit sequence*.

Deterministic here means that for the given the same initial seed, the generator will always produce the same output sequence.

Polynomial-time statistical tests

Definition

A pseudorandom bit generator is said to pass all *polynomial*-time statistical tests* if no polynomial-time algorithm can correctly distinguish between an output sequence of the generator and a truly random sequence of the same length with probability significantly greater that $\frac{1}{2}$.

* The running time of the test is bounded by a polynomial in the length l of the output sequence.

CSPRBG

Definition

A pseudorandom bit generator is said to pass the *next bit test* if there is no polynomial-time algorithm which, on input of the first l bits of an output sequence s, can predict the $(l+1)^{st}$ bit of s with probability significantly greater than $\frac{1}{2}$.

Theorem

(Universality of the next-bit tests). A pseudorandom bit generator passes the next-bit test if and only if it passes all polynomial-time statistical tests.

Definition

A PRBG that passes the next-bit test (possibly under some plausible but unproved mathematical assumption such as the intractability of factoring integers) is called a *cryptographically secure pseudorandom bit generator (CSPRBG)*.

Sources of random bits

1. **Hardware-based generators.** Hardware-based random bit generators exploit the randomness which occurs in some physical phenomena. Examples of such physical phenomena include:
   - elapsed time between emission of particles during radioactive decay;
   - thermal noise from a semiconductor diode or resistor;
   - the frequency instability of a free running oscillator.
2. **Software-based generators.** Designing a random bit generator in software is even more difficult than doing so in hardware. Processes upon which software random bit generators may be based include include:
   - the system clock;
   - elapsed time between keystrokes or mouse movement;
   - content of input/output buffers.

   The behavior of such processes can vary considerably depending on various factors, such as the computer platform.

RSA pseudorandom bit generator

Algorithm 1 Algorithm RSA PRBG

```
Output: $z_1, z_2, \dots, z_l \in \mathbb{Z}_2$ a pseudorandom bit sequence
Setup. Generate two secret RSA-like primes $p$ and $q$, and compute $n = pq$ and
  $\phi = (p-1)(q-1)$. Select a random integer $e, 1 < e < \phi$, such that GCD$(e, \phi) = 1$.
Select a random integer $x_0$ (the seed) in the interval $[1, n-1]$.
for $i = 1$ to $l$ do
    $x_i = x_{i-1}^e$ mod $n$
    $z_i =$ the least significiant bit of $x_i$
end for
return the output sequence $z_1, z_2, \dots, z_l$
```

RSA Generator is provably secure

"It is difficult to predict the next number in the sequence given the previous numbers, assuming that it is difficult to invert the RSA function" (Shamir)

Prime numbers generator

- Goal is to generate n-bit prime number
- Assumption: existence of an efficient primality test
- Idea of the algorithm is to repeat:
  1. pick a random number $x \in [2^{n-1}, 2^n - 1]$
  2. apply the primality test on x

  until x is prime

Theorem

(Prime Number Theorem). The total number of primes $< N$ *is (roughly)* $P_N = N/\log N$.

- So, the number of primes in the range $[2^{n-1}, 2^n - 1]$ is $P_{2^n} - P_{2^{n-1}} \approx 2^{n-1} \cdot \frac{1}{n}$.
- If we pick a number at random from the range $[2^{n-1}, 2^n - 1]$, then the probability that the number is a prime is $\frac{1}{n}$.

Statistical tests

The normal and χ^2 distributions are widely used in statistical applications.

Definition

If the result X of an experiment can be any real number, then X is said to be a *continuous* random variable.

Definition

A *probability density function* of a continuous random variable X is a function $f(x)$ which can be integrated and satisfies:

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$;
2. $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$; and
3. for all $a, b \in \mathbb{R}, P(a < X \leq b) = \int_a^b f(x)\,\mathrm{d}x$.

The normal distribution

The normal distribution arises in practise when a large number of independent random variables having the same mean and variance are summed.

Definition

A (continuous) random variable X has a *normal distribution* with *mean* μ and *variance* σ^2 if its probability density function is defined by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \text{exp}\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}, -\infty < x < \infty.$$

Notation. X is said to be $N(\mu, \sigma^2)$. If X is $N(0,1)$ then X is said to have a *standard normal distribution.*

Fact. If the random variable X is $N(\mu, \sigma^2)$, then the random variable $Z = \frac{(X-\mu)}{\sigma}$ is $N(0,1)$.

χ^2 distribution

The χ^2 distribution can be used to compare the *goodness-of-fit* of the observed frequencies of events to their expected frequencies under a hypothesized distribution. The χ^2 distribution with v degrees of freedom arises in practice when the squares of v independent random variables having standard normal distributions are summed.

Definition

Let $v \geq 1$ be an integer. A (continuous) random variable X has a χ^2 *(chi-square) distribution* with v degrees of freedom if its probability density function is defined by

$$f(x) = \begin{cases} \frac{1}{\Gamma(v/2)2^{v/2}} x^{(v/2)-1} e^{-x/2}, & 0 \leq x < \infty, \\ 0, & x < 0, \end{cases}$$

where Γ is the gamma function. The *mean* and *variance* of this distribution are $\mu = v$, and $\sigma^2 = 2v$

χ^2 distribution

Definition

The gamma function is defined by $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \mathrm{d}x$, for $t > 0$.

Fact. If the random variable X is $N(\mu, \sigma^2), \sigma^2 > 0$ then the random variable $Z = \frac{(X-\mu)^2}{\sigma^2}$ has a χ^2 distribution with 1 degree of freedom. In particular, if X is $N(0,1)$, then $Z = X^2$ has a χ^2 distribution with 1 degree of freedom.

Hypothesis testing

A *statistical hypothesis,* denoted H_0, is an assertion about a distribution of one or more random variables. A *test* of a statistical hypothesis is a procedure, based upon observed values of the random variables, that leads to the acceptance or rejection of the hypothesis H_0. When we testing hypotheses, we always compare two hypotheses. One of them we're testing, we called the *null hypothesis* H_0 and in contrast to it we build the so-called *alternative hypothesis* H_A.The test only provides a measure of the strength of the evidence provided by the data against the hypothesis; hence, the conclusion of the test is not define, but rather probabilistic.

Hypothesis testing

Let $X = (X_1, \ldots, X_n)$ be a random sequence from some distribution $R(\theta)$, where θ is a parameter that can be multidimensional. Let $h(\theta)$ be a parametric function and k is a real constant. Then the null hypothesis be in the following form:

$$H_0 : h(\theta) = k.$$

The alternative hypothesis can be defined in the following three ways:

- Right-handed alternative hypothesis: $H_A : h(\theta) > k$
- Left alternative hypothesis: $H_A : h(\theta) < k$
- Two-sided alternative hypothesis: $H_A : h(\theta) \neq k$.

Hypothesis testing

If the result of the test corresponds with reality, then a correct decision has been made. However, if the result of the test does not correspond with reality, then an error has occurred. There are two situations in which the decision is wrong. The null hypothesis may be true, whereas we reject H_0. On the other hand, the alternative hypothesis H_A may be true, whereas we do not reject H_0. Two types of error are distinguished: type I error and type II error.

Table of error types

Reality	**Decision about null hypothesis (H_0)**	
	Do not reject H_0	Reject H_0
H_0 is true	true positive	type I error
H_0 is not true	type II error	true negative

Hypothesis testing

Definition

The *significance level* α of the test of a statistical hypothesis H_0 is the probability of rejecting H_0 when it is true.

A statistical test is implemented by specifying a *statistic* on the random sample. A statistic is a function of the elements of a random sample. For example, the number of 0's in binary sequence is statistic. Statistics are generally chosen so that they can be efficiently computed, and so that they (approximately) follow an $N(0,1)$ or a χ^2 distribution. The value of the stistic for the sample output sequence is computed and compared with the value expected for a random sequence.

Hypothesis testing

1. Suppose that a statistic X for a random sequence follows a χ^2 distribution with v degrees of freedom, and suppose that the statistic can be expected to take on larger values for nonrandom sequences. To achieve a significance level for α, a threshold value x_α is chosen so that $P(X > x_\alpha) = \alpha$. If the value of $X_S > x_\alpha$, then the sequence *fails* the test; otherwise, it *passes* the test. Such a test is called a *one-sided test.*
2. Suppose that a statistic X for a random sequence follows a $N(0, 1)$ distribution, and suppose that the statistic can be expected to take on both larger and smaller values for nonrandom sequences. To achieve a significance level for α, a threshold value x_α is chosen so that $P(X > x_\alpha) = P(X < -x_\alpha) = \frac{\alpha}{2}$. If the value of X_S of the statistic for the sample output sequence satisfies $X_S > x_\alpha$ or $X_S < -x_\alpha$, then the sequence *fails* the test; otherwise, it *passes* the test. Such a test is called a *two-sided test.*

Chi-Square Goodness-of-Fit Test

The χ^2 (chi-square) goodness of fit test is used to compare the observed distribution to an expected distribution, in a situation where we have two or more categories in a discrete data. In other words, it compares multiple observed proportions to expected probabilities. The chi-square test is defined for the hypothesis:

- H_0 : There is no significant difference between the observed and the expected value.
- H_A : There is a significant difference between the observed and the expected value.

Chi-Square Goodness-of-Fit Test

For the chi-square goodness-of-fit computation, the value of the test-statistic is

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} = N \sum_{i=1}^{n} \frac{(\frac{O_i}{N} - p_i)^2}{p_i}$$

where

- O_i the number of observations of type i,
- N is the total number of observations,
- $E_i = N.p_i$ the expected (theoretical) count of type i, asserted by the null hypothesis that the fraction of type i in the population is p_i and n is the number of cells in the table.

Chi-Square Goodness-of-Fit Test

The test can be used provided that all values of $N \cdot p_i$ are at least 5. The random variable X^2 has an approximate χ^2 distribution of $n-1$ degrees of freedom. We reject the null hypothesis at the level of significance α, if $X^2 > {\chi^2}_{1-\alpha;n-1}$, where the value of ${\chi^2}_{1-\alpha;n-1}$ is a quantile χ^2 distribution of $n-1$ degrees of freedom.

The chi-squared statistic can then be used to calculate a p-value by comparing the value of the statistic to a chi-squared distribution.

The number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

Chi-Square Goodness-of-Fit Test

Example. Suppose you flip two coins 100 times. The results are 20 HH, 27 HT, 30 TH, and 23 TT. Are the coins fair? Test at a 5% significance level.

Solution. This problem can be set up as a goodness-of-fit problem. The sample space for flipping two fair coins is HH, HT, TH,TT. Out of 100 flips, you would expect 25 HH, 25 HT, 25 TH, and 25 TT. This is the expected distribution. The question, "Are the coins fair?" is the same as saying, "Does the distribution of the coins (20 HH, 27 HT, 30 TH, 23 TT) fit the expected distribution?"

Random Variable: Let X is the number of heads in one flip of the two coins. X takes on the values 0, 1, 2. (There are 0, 1, or 2 heads in the flip of two coins.) Therefore, the number of cells is three. Since X is the number of heads, the observed frequencies are 20 (for two heads), 57 (for one head), and 23 (for zero heads or both tails). The expected frequencies are 25 (for two heads), 50 (for one head), and 25 (for zero heads or both tails). This test is right-tailed.

- H_0 : The coins are fair.
- H_A : The coins are not fair.

Distribution for the test: χ^2, where df = 3 − 1 = 2.
Calculate the test statistic: $X = 2.14$.
Probability statement: p-value = $P(X > 2.14) = 0.3430$.
Compare α *and the* p*-value:*

- $\alpha = 0.05$
- p-value = 0.3430.
- $\alpha <$ p-value.

Make a decision: Since $\alpha <$ p-value, do not reject H_0.
Conclusion: There is insufficient evidence to conclude that the coins are not fair.

Golomb's randomness postulates

Deciding the pseudorandomness of a sequence is a difficult task. Golomb's randomness postulates were one of the first attempts to establish some necessary conditions for a periodic pseudorandom sequence to look random.

Definition

Let $s = s_0, s_1, s_2, \ldots$ be an infinite sequence. The subsequence consisting the first n terms of s is denoted by $s^n = s_0, s_1, s_2, \ldots, s_{n-1}$.

Definition

The sequence $s = s_0, s_1, s_2, \ldots$ is said to be N-*periodic* if $s_i = s_{i+N}$ for all $i \geq 0$. The sequence s is *periodic* if it is N-periodic for same positive integer N. The *period* of a periodic sequence s is the smallest positive integer N for which s is N-periodic. If s is a periodic sequence of period N, then the *cycle* of s is the subsequence s^N.

Definition

Let s be a sequence. A *run* of s is a subsequence of s consisting of consecutive 0's or consecutive 1's which is neither preceded nor succeeded by the same symbol. A run of 0's is called a *gap*, while a run of 1's is called a *block*.

Definition

Let $s = s_0, s_1, s_2, \ldots$ be a periodic sequence of period N. The *autocorrelation function* of s is the integer-valued function $C(t)$ defined as

$$C(t) = \frac{1}{N} \sum_{i=0}^{N-1} (2s_i - 1)(2s_{i+t} - 1), \text{for } 0 \leq t \leq N - 1.$$

The autocorrelation function $C(t)$ measure the amount of similarity between the sequence s and a shift of s by t positions.

Definition

Let s be a periodic sequence of period N. *Golomb's randomness postulates* are the following.

- R1: In the cycle s^N of s, the number of 1's differs from the number of 0's by at most 1.
- R2: In the cycle s^N, at least $\frac{1}{2}$ the runs have length 1, at least $\frac{1}{4}$ have length 2,at least $\frac{1}{8}$ have length 3, etc., as long as the number of runs so indicated exceeds 1. More over for each of these lengths, there are (almost) equally many gaps and blocks.
- R3: The autocorrelation function $C(t)$ is two-valued. That is for some integer K,

$$N.C(t) = \sum_{i=0}^{N-1}(2s_i - 1)(2s_{i+t} - 1) = \begin{cases} N, & \text{if } t = 0, \\ K, & \text{if } 1 \leq t \leq N-1 \end{cases}$$

Definition

A binary sequence which satisfies Golomb's randomness postulates is called a *pseudo-noise sequence* or a *pn-sequence*

Basic tests

This subsection presents statistical tests that are commonly used for determining whether the binary sequence s possesses some specific characteristics that a truly random sequence would be likely exhibit. It is emphasized again that the outcome of each test is not definite, but rather probabilistic.

Frequency test(monobit test)

Let $s = s_0, s_1, s_2, \ldots, s_{n-1}$ be a binary sequence of length n. The purpose of this test is to determine whether the number of 0's and 1's in s are approximately the same, as would be expected for a random sequence. First, we count the number of 1's in the sequence s and their number denote as $n_1 = \sum_{i=0}^{n-1} s_i$. The number of 0's will be $n_0 = n - n_1$, where n is the length of the sequence s. In a random sequence of bits of length n, the number of 1's is expected to be the same as the number of 0's and will be equal to $n_0 = n_1 = \frac{n}{2}$. In other words, at zero hypothesis, we assume that the number of units of the sequence s is described by a random quantity with binomial distribution with parameters n and $p = \frac{1}{2}$. Next we apply Chi-square goodness of fit test.

Frequency test. Testing hypothesis

Let's define the null hypothesis H_0 and the alternative hypothesis H_A:

- $H_0 : n_1 = \frac{n}{2}$
- $H_A : n_1 \neq \frac{n}{2}$.

The statistic used is

$$X^2 = \frac{(n_0 - \frac{n}{2})^2}{\frac{n}{2}} + \frac{(n_1 - \frac{n}{2})^2}{\frac{n}{2}} = \frac{(n_0 - n_1)^2}{n}.$$

We decide to reject or not reject the null hypothesis on the basis of the value of the quantile $\chi^2{}_{1-\alpha}; f$. We could choose the level of significance $\alpha = 0,05$ and we consider $f = 1$ degree of freedom if $n \geq 10$.

Serial test(two-bit test)

Let $s = s_0, s_1, s_2, \ldots, s_{n-1}$ be a binary sequence of length n. The purpose of this test is to determine whether the number of occurrences of 00, 01, 10 and 11 as subsequences of s are approximately the same, as would be expected for a random sequence. Let n_0, n_1 denote the number of 0's and 1's in s, respectively, and let $n_{00}, n_{01}, n_{10}, n_{11}$ denote the number of ccurrences of 00, 01, 10, 11 in s, respectively. Note that $n_{00} + n_{01} + n_{10} + n_{11} = (n-1)$ since the subsequences are allowed to overlap. The statistic used is

$$X^2 = \frac{4}{(n-1)}(n_{00}^2 + n_{01}^2 + n_{10}^2 + n_{11}^2) - \frac{2}{n}(n_{00}^2 + n_{11}^2) + 1$$

which approximately follows a χ^2 distribution.

Poker test

Let m be a positive ineteger such that $\lfloor \frac{n}{m} \rfloor \geq 5\dot{(}2^m)$, and let $k = \lfloor \frac{n}{m} \rfloor$. Divide the sequence s into k non-overlapping parts each of length m, and let n_i be the number of occurences of the i^{th} type of sequence of length $m, 1 \leq i \leq 2^m$. The poker test determines whether the sequences of length m each appear approximately the same number of times in s, as would be expected for a random sequence. The statistic used is

$$X^2 = \frac{2^m}{k}(\sum_{i=1}^{2^m} {n_i}^2) - k$$

which approximately follows a χ^2 distribution with $2^m - 1$ degrees of freedom. Note that the poker test is a generalization of the frequency test; setting $m = 1$ in the poker test yields the frequency test.

Runs test

The purpose of the runs test is to determine whether the number of runs (of either zeros or ones) of various lengths in the sequence s is as expected for a random sequence. The expected number of gaps (or blocks) of length i in a random sequence of length n is $e_i = \frac{(n-i+3)}{2^{i+2}}$. Let k be equal to the largest integer i for which $e_i \geq 5$. Let B_i, G_i be the number of blocks and gaps, respectively, of length i in s for each $i, 1 \leq i \leq k$. The statistic used is

$$X^2 = \sum_{i=1}^{k} \frac{(B_i - e_i)^2}{e_i} + \sum_{i=1}^{k} \frac{(G_i - e_i)^2}{e_i}$$

which approximately follows a χ^2 distribution with $2^k - 2$ degrees of freedom.

Autocorrelation test

The purpose of this test is to check for correlations between the sequence s and (non-cyclic) shifted versions of it. Let d be a fixed integer, $1 \leq d \leq \lfloor \frac{n}{2} \rfloor$. The number of bits in s not equal to their $d-$shifts is $A(d) = \sum_{i=0}^{n-d-1} s_i \oplus s_{i+d}$, where $\oplus$ denotes the XOR operator. The statistic used is

$$X^2 = 2\frac{(A(d) - \frac{n-d}{2})}{\sqrt{n-d}}$$

which approximately follows a $N(0,1)$ distribution if $n - d \geq 10$. Since small values of $A(d)$ are as unexpected as large values of $A(d)$, a two-sided test should be used.

Source

[1] A. Menezes, P. vanOorschot, and S. Vanstone, *Handbook of Applied Cryptography, CRC Press, 1996.*

Linear congruential generator - duplicity (see MI-KRY)

1. The linear congruential generator is defined by recurrence relation: $x_i = ax_{i-1} + c \bmod m$.
2. If $c = 0$, the generator is often called a multiplicative congruential generator (MCG), or Lehmer RNG. If $c \neq 0$, the method is called a mixed congruential generator.
3. Generates numbers in the interval $[1, m - 1]$
4. While LCGs are capable of producing pseudorandom numbers which can pass formal tests for randomness, the quality of the output is extremely sensitive to the choice of the parameters m and a.

Linear congruential(LCG) PRBG - duplicity (see MI-KRY)

Algorithm 2 Algorithm LCG PRBG

```
Input: $l-$length generated sequence
Output: $x_1, x_2, \ldots, x_l \in \mathbb{Z}_2$ a pseudorandom bit sequence
select a random integer $x_0$ (the seed) and select parameters $a, c, m$, where $a, c$ are
    in the interval $[1, m]$
for $i = 1$ to $l$ do
    $x_i = ax_{i-1} + c$ mod $m$
end for
return the output sequence $x_1, x_2, \ldots, x_l$
```

Multiplicative congruential generator (MCG) choosing parameters - duplicity (see MI-KRY)

- Definition: $x_i = ax_{i-1}$ mod m.
- When m is a power of 2 (most often $m = 2^{32}$ or $m = 2^{64}$) and $c = 0$ and $a = 3$ mod 8 or $a = 5$ mod 8 then the maximal period is $m/4$. The seed x_0 must be odd.
- When m is a prime and $c = 0$ then the period is $m - 1$ if the multiplier a is chosen to be a primitive element of the integers modulo m. The seed x_0 must be chosen between 1 and $m - 1$.

LCG choosing parameters - duplicity (see MI-KRY)

- Definition: $x_i = ax_{i-1} + c$ mod m.
- For $c = 0$ and $m = 2^b$, correctly chosen parameters allow a period to be equal to $m = 2^b$, for all seed values. This will occur if and only if:

  1. GCD$(c, m) = 1$,
  2. $a - 1$ is divisible by all prime factors of m
  3. $a - 1$ is divisible by 4 if m is divisible by 4

  These three requirements are referred to as the Hull–Dobell Theorem.

Blum-Blum-Shub pseudorandom bit generator - duplicity (see MI-KRY)

Blum Blum Shub (B.B.S.) is a pseudorandom number generator proposed in 1986 by Lenore Blum, Manuel Blum and Michael Shub that is derived from Michael O. Rabin's one-way function.
Blum Blum Shub takes the form $x_{i+1} = {x_i}^2$ mod n where $n = pq$ is the product of two large primes p and q. At each step of the algorithm, some output is derived from x_{i+1}; the output is commonly either the bit parity of x_{i+1} or one or more of the least significant bits of x_{i+1}.
The seed s should be an integer that is co-prime to n(i.e. p and q are not factors of x_0) and not 1 or 0.

Blum-Blum-Shub pseudorandom bit generator - duplicity (see MI-KRY)

Algorithm 3 Algorithm Blum-Blum-Shub PRBG

```
Input: $l-$length generated sequence
Output: $z_1, z_2, \ldots, z_l \in \mathbb{Z}_2$ a pseudorandom bit sequence
generate large secret random (and distinct) primes $p$ and $q$, each congruent to $3$
    modulo $4$ and compute $n = pq$
select a random integer $s$ (the seed) in the interval $[1, n-1]$ such that GCD$(s, n) =$
    $1$, and compute $x_0 = s^2$ mod $n$.
for $i = 1$ to $l$ do
    $x_i = x_{i-1}^2$ mod $n$
    $z_i =$ the least significiant bit of $x_i$
end for
return the output sequence $z_1, z_2, \ldots, z_l$
```