Limit theorem for the image size of a subset under compositions of random mappings

A. M. Zubkov, A. A. Serov

Steklov Mathematical Institute of Russian Academy of Sciences

St. Petersburg 2017

Exhaustive search

- Cryptanalytic attacks based on exhaustive search need a lot of computing power or a lot of time to complete.
- When the same attack has to be carried out multiple times, it may be possible to execute the exhaustive search in advance and store all results in memory.
- Once this precomputation is done, the attack may be carried out almost instantly. Unfortunately, this method cannot be realizable practically because of the large amount of memory needed.

Time-memory trade-off

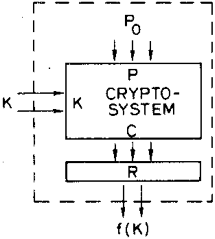
- In 1980 Martin Hellman described a cryptanalytic time-memory trade-off which reduces the time of cryptanalysis by using precalculated data stored in memory.
- For a cryptosystem having N keys this method recovers a key in $N^{2/3}$ operations using $N^{2/3}$ words of memory.
- The typical application of this method is the recovery of a key in cases when the password hash or the plaintext-ciphertext pair are known.
- The time-memory trade-off is a probabilistic method. Success is not guaranteed and the success probability depends on the amount of time and memory available.

Hellman method

Construction of the function f

P_0 — given fixed plaintext,
C_0 — corresponding ciphertext,
$\mathcal{X}$ — key space, $|\mathcal{X}| = N$, $K \in \mathcal{X}$,
R — reduction function which creates a new key from a cipher text.

$$f(K) = R(S_K(P_0))$$



The method tries to find the key K which was used to encipher the plaintext using the cipher S:

$$C_0 = S_K(P_0),$$

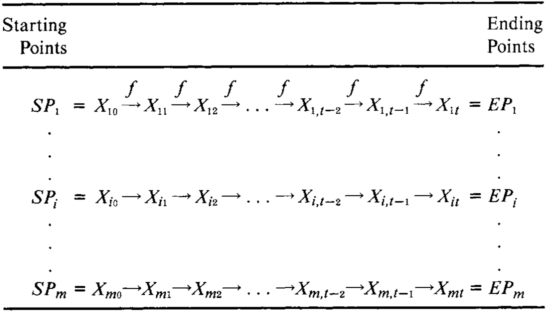
and tries to generate all possible ciphertexts in advance by enciphering the plaintext with all N possible keys.

Time-memory tradeoff table

The ciphertexts are organised in chains:

$$\mathrm{SP}_i = X_{i,0} \xrightarrow{f} X_{i,1} \xrightarrow{f} \ldots \xrightarrow{f} X_{i,t} = \mathrm{EP}_i, \quad 1 \le i \le m,$$

$$X_{i,j+1} = f(X_{i,j}) = R(S_{X_{i,j}}(P_0)).$$



Starting Points		Ending Points
SP_1	$= X_{10} \xrightarrow{f} X_{11} \xrightarrow{f} X_{12} \xrightarrow{f} \ldots \xrightarrow{f} X_{1,t-2} \xrightarrow{f} X_{1,t-1} \xrightarrow{f} X_{1t} =$	EP_1
$\vdots$		$\vdots$
SP_i	$= X_{i0} \to X_{i1} \to X_{i2} \to \ldots \to X_{i,t-2} \to X_{i,t-1} \to X_{it} =$	EP_i
$\vdots$		$\vdots$
SP_m	$= X_{m0} \to X_{m1} \to X_{m2} \to \ldots \to X_{m,t-2} \to X_{m,t-1} \to X_{mt} =$	EP_m

m chains of length t are created and their first and last elements $(\mathrm{SP}_i, \mathrm{EP}_i)$, $1 \le i \le m$, are stored in the table.

The starting points $\mathrm{SP}_1, \ldots, \mathrm{SP}_m$ are randomly chosen from $\mathcal{X}$.

The probability of success

M. E. Hellman had shown that the chance of finding a key by using a table of m rows and t keys in the row is the following:

$$\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{t}\left(1-\frac{it}{N}\right)^{j} < \mathbf{P}_{table} \leqslant \frac{mt}{N}\,.$$

The efficiency of a single table rapidly decreases with its size. The probability of success for at least one of l tables is estimated as

$$\mathbf{P}_{success} > 1-\left(1-\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{t}\left(1-\frac{it}{N}\right)^{j}\right)^{l}\,.$$

For $mt^2 \approx N$ and $m, t \gg 1$

$$\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{t}\left(1-\frac{it}{N}\right)^{j} \geqslant \frac{mt}{N}\int_{0}^{1}\frac{1-e^{-x}}{x}dx \approx 0.796599\,\frac{mt}{N}.$$

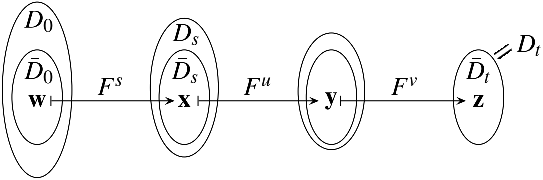
Coverage rate

Let $F : \mathcal{X}_N \to \mathcal{X}_N$ be a random function

$F^u = F(\dots F(F(\cdot)))$ be the u-times iteration of F

$$D_0, D_s, D_t \subset \mathcal{X}_N,\ D_k = F^k(D_0),\ k = 1, 2, \dots,$$

$$\bar{D}_0 \subset D_0,\ \bar{D}_s \subset D_s$$

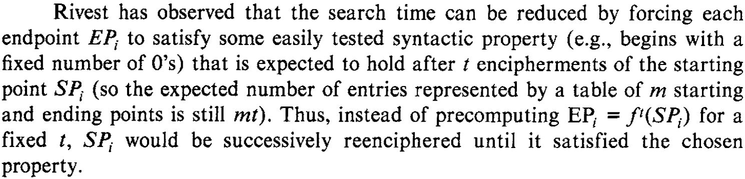


$$\mathbf{x} = F^s(\mathbf{w}),\ \mathbf{y} = F^u(\mathbf{x}),\ \mathbf{z} = F^\nu(\mathbf{y})$$

Distinguished points

The Martin Hellman technique was improved by Rivest before 1982 with the introduction of distinguished points which drastically reduces the number of memory lookups during cryptanalysis.

Denning D. E., *Cryptography and Data Security*, Addison-Wesley, 1982, 419 pp.:



> Rivest has observed that the search time can be reduced by forcing each endpoint EP_i to satisfy some easily tested syntactic property (e.g., begins with a fixed number of 0's) that is expected to hold after t encipherments of the starting point SP_i (so the expected number of entries represented by a table of m starting and ending points is still mt). Thus, instead of precomputing $\mathrm{EP}_i = f^t(SP_i)$ for a fixed t, SP_i would be successively reenciphered until it satisfied the chosen property.

This suggestion greatly reduces the number of memory lookups!

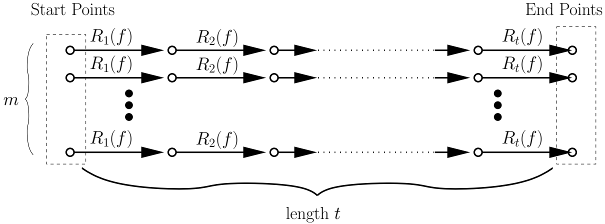
This improved technique has been studied extensively but no new optimisations have been published until 2003.

Rainbow tables

The main limitation of the original Hellman scheme is the fact that when two chains collide in a single table they merge.

In 2003 Philippe Oechslin has proposed a new way of precalculating the data which reduces by two the number of calculations needed during cryptanalysis.

$$f(K) = R(S_K(P_0)),$$
$$R_1, R_2, \ldots, R_t \colon \mathcal{X}_N \to \mathcal{X}_N \text{ — reduction functions}$$



Moreover, since the method does not make use of distinguished points, it reduces the overhead due to the variable chain length, which again significantly reduces the number of calculations.

Mathematical model and heuristic estimate

$F_1, F_2, \ldots$ be a sequence of random independent equiprobable mappings $\mathcal{X}_N \to \mathcal{X}_N$, $\mathcal{X}_N = \{X_1, \ldots, X_N\}$, $|\mathcal{X}_N| = N$.

$$S_0 \subset \mathcal{X}_N, \quad |S_0| = m,$$
$$S_1 = F_1(S_0),\ S_2 = F_2(F_1(S_0)), \ldots, S_t = F_t(F_{t-1}(\ldots(F_1(S_0))\ldots)),$$
$$\Psi_t = S_1 \cup S_2 \cup \ldots \cup S_t\,.$$

Heuristic estimate for the success probability of the Rainbow table method: for any $x \in \mathcal{X}_N$, $S_0 \subset \mathcal{X}_N$

$$\mathbf{P}\left\{x \in S_0 \cup \Psi_t\right\} \approx 1 - \prod_{i=1}^{t+1}\left(1 - \frac{m_i}{N}\right),$$

where $m_1 = |S_0| = m$, $m_{i+1} = N\left(1 - e^{-\frac{m_i}{N}}\right)$ for $i \geqslant 1$.

Markov chains $\{\varphi_k\}_{k\geqslant 0}$ and $\{\zeta_k\}_{k\geqslant 0}$

$$S_0 \subset \mathcal{X}_N,\ |S_0| = m, \quad S_k = F_k(S_{k-1}), \quad \Psi_k = \cup_{j=1}^{k} S_j,\ k \geqslant 1. \qquad (1)$$
$$\varphi_0 = |S_0|,\ \zeta_0 = 0,\ \varphi_k = |S_k|,\ \zeta_k = |\Psi_k|,\ k \geqslant 1.$$

The transition probability matrix of the Markov chain $\{\varphi_k\}_{k\geqslant 0}$ has the form

$$P = \|p_{i,j}\|_{i,j=1}^{N},$$

$$p_{i,j} = \begin{cases} \binom{N}{j}\left(\frac{j}{N}\right)^i \sum\limits_{u=0}^{j} (-1)^u \binom{j}{u} \left(1 - \frac{u}{j}\right)^i, & 1 \leqslant j \leqslant i \leqslant N, \\ 0, & j > i. \end{cases}$$

The transition probability matrix of the Markov chain $\{(\varphi_k, \zeta_k)\}_{k\geqslant 0}$ has the form

$$Q = \|q_{(i,r),(j,s)}\|_{i,j,r,s=1}^{N},$$

$$q_{(i,r),(j,s)} = \begin{cases} p_{i,j} \frac{\binom{N-r}{s-r}\binom{r}{j-s+r}}{\binom{N}{j}}, & \substack{1\leqslant j \leqslant i \leqslant N, \\ 1 \leqslant r \leqslant s \leqslant \min\{N, r+j\}}, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem 1 (A.M. Zubkov, A.A. Serov)

Let $F_1, F_2, \ldots$ be the random independent equiprobable mappings of the set X_N to itself, $S_0 \subseteq X_N$, $|S_0| = m$, $S_k = F_k(\ldots(F_1(S_0))\ldots)$, $k \geqslant 1$. For any element $x \in X_N$, which does not depend on F_1, $F_2, \ldots$, for all $1 \leqslant k$, $m \leqslant N$ we have

$$\tfrac{m}{N} - C_m^2 \tfrac{k}{N^2} \leqslant \mathbf{P}\{x \in S_k \,|\, \varphi_0 = m\} < \tfrac{m}{N} - C_m^2 \tfrac{k}{N^2} + \tfrac{m^3 k^2}{4N^3}\,,$$

$$\tfrac{mt}{N} - C_{t+1}^2 \tfrac{3\,m^2}{2N^2} < \mathbf{P}\left\{x \in \Psi_t \,\middle|\, \varphi_0 = m\right\} < \tfrac{mt}{N} - C_m^2 C_{t+1}^2 \tfrac{1}{N^2} + \tfrac{m^3 (t+1)^3}{12N^3}\,.$$

The following inequalities are valid also:

$$m - C_m^2 \tfrac{k}{N} \leqslant \mathbf{M}\{\varphi_k \,|\, \varphi_0 = m\} < m - C_m^2 \tfrac{k}{N} + \tfrac{m^3 k^2}{4N^2}\,,$$

$$mt - C_{t+1}^2 \tfrac{3\,m^2}{2N} < \mathbf{M}\left\{\zeta_t \,\middle|\, \varphi_0 = m\right\} < mt - C_m^2 C_{t+1}^2 \tfrac{1}{N} + \tfrac{m^3 (t+1)^3}{12N^2}\,.$$

Assertions

Assertion 1

If the images of the initial subset $S_0 \subset \mathcal{X}_N$, $|S_0| = m$, are calculated according to the formulas (1)*, then the following identities are true:*

$$\mathbf{P}\left\{|S_t| = m \,\middle|\, |S_0| = m\right\} = \left(\prod_{q=1}^{m-1}\left(1 - \frac{q}{N}\right)\right)^t,$$

$$\mathbf{P}\left\{|S_t| = m-1 \,\middle|\, |S_0| = m\right\} = \frac{m}{2}\left(1 - \left(1 - \frac{m-1}{N}\right)^t\right)\left(\prod_{q=1}^{m-2}\left(1 - \frac{q}{N}\right)\right)^t.$$

Let $p_0(m) = \mathbf{P}\left\{|S_1| = m \,\middle|\, |S_0| = m\right\}$, $p_1(m) = \mathbf{P}\left\{|S_1| = m-1 \,\middle|\, |S_0| = m\right\}$.

From the Assertion 1 and $\left(1 - \sum_{i=1}^k x_i\right) \leqslant \prod_{i=1}^k (1 - x_i)$, $x_i \in [0,1)$, it follows, that

$$p_0(m) \geq 1 - \frac{m(m-1)}{2N}, \quad p_1(m) \geq \frac{m(m-1)}{2N} - \frac{m(m-1)^2(m-2)}{4N^2}. \qquad (2)$$

Consider the event

$$A_{m,t} = \left\{ |S_0| = m, \bigcap_{k=0}^{t-1} \{|S_{k+1}| \geq |S_k| - 1\} \right\}.$$

From (2) for $t\,m^4 \leq 4N^2$ it follows that

$$\mathbf{P}\{A_{m,t}\} > (p_0(m) + p_1(m))^t \geq \left(1 - \frac{m(m-1)^2(m-2)}{4N^2}\right)^t > 1 - \frac{tm^4}{4N^2}\,.$$

Thus, if $m, t, N \to \infty$, $m < CN^{1/4}$, $t = o(N)$, then $\mathbf{P}\{A_{m,t}\} \to 1$.

Consider the auxiliary Markov chain $\{S_k^*\}_{k=0}^{\infty}$ with $S_0^* = |S_0| = m$ and transition probabilities

$$\mathbf{P}\{S_{k+1}^* = j \,|\, S_k^* = j\} = p_0(j),$$

$$\mathbf{P}\{S_{k+1}^* = j-1 \,|\, S_k^* = j\} = 1 - p_0(j) = \mathbf{P}\{|S_{k+1}| \leq j-1 \,|\, |S_k| = j\} \geq$$

$$\geq \mathbf{P}\{|S_{k+1}| = j-1 \,|\, |S_k| = j\},\ j = 2, \ldots, m.$$

So, for any nonincreasing sequence $m_0 = m \geq m_1 \geq \ldots \geq m_t \geq 1$ such that $\max_{0\leq k<t}(m_k - m_{k+1}) \leq 1$ we have

$$\mathbf{P}\{S_k^* = m_k\,(1 \leq k \leq t)\,|\,S_0^* = m\} \geq \mathbf{P}\{|S_k| = m_k\,(1 \leq k \leq t)\,|\,|S_0| = m\},$$

and $\mathbf{P}\{S_k^* = m_k\,(1 \leq k \leq t)\,|\,S_0^* = m\} = 0$ otherwise. Thus

$$\sum_{m_0=m\geq m_1\geq\ldots\geq m_t\geq 1} |\mathbf{P}\{S_k^* = m_k\,(1 \leq k \leq t)\,|\,S_0^* = m\} \\ -\mathbf{P}\{|S_k| = m_k\,(1 \leq k \leq t)\,|\,|S_0| = m\}| \\ = 2\mathbf{P}\left\{\max_{0\leq k<t}(|S_k| - |S_{k+1}|) > 1\right\} = 2(1 - \mathbf{P}\{A_{m,t}\}),$$

that is if m, t and N tend to ∞ in such a way that $\mathbf{P}\{A_{m,t}\} \to 1$, then the total variation distance between the distributions of trajectories of Markov chains $\{|S_k|\}_{k=0}^t$ and $\{S_k^*\}_{k=0}^t$ tends to 0. Consequently, the total variation distance between the distributions of any functions of these trajectories tends to 0.

Consider the random variables

$$T_n = \min\{k \geqslant 1\colon S_k^* = n\},\ n = 1, \ldots, m.$$

If $m, n, N \to \infty$ in such a way that m is of the order $N^{1/4}$ and $n = o(m)$, then

$$\mathbf{E}T_n = \sum_{j=n+1}^{m} \frac{1}{1-\lambda_j} = 2N\left(\frac{1}{n} - \frac{1}{m}\right)\left(1 + O\left(\frac{n^2}{N}\right)\right),$$

$$\mathbf{D}T_n = \sum_{j=n+1}^{m} \frac{\lambda_j}{(1-\lambda_j)^2} = \frac{4N^2}{3}\left(\frac{1}{n^3} - \frac{1}{m^3}\right)(1 + o(1))\ ,$$

$$C_3(n,m) = \sum_{j=n+1}^{m} \mathbf{E}|\delta_j - \mathbf{E}\delta_j|^3 < 10N^3\left(\frac{1}{n^5} - \frac{1}{m^5}\right).$$

If $0 < \varepsilon < \frac{n}{m} < 1 - \varepsilon$, then the Lyapunov ratio

$$\frac{C_3(n,m)}{(\mathbf{D}T_n)^{3/2}} = O(n^{3\cdot 3/2-5}) = O(n^{-1/2})$$

tends to 0 as $N, m, n \to \infty$, $n = o(m)$.

According to the Lyapunov theorem the distribution of T_n is asymptotically normal with parameters $(\mathbf{E}T_n, \mathbf{D}T_n)$.

The equalities $\{S_t^* \le n\} = \{T_n \le t\}$ allow to find the asymptotic behavior of distribution of S_t^* for $N, m \to \infty, n = o(m)$, since

$$\mathbf{P}\left\{\frac{T_n - \mathbf{E}T_n}{\sqrt{\mathbf{D}T_n}} \le x\right\} = \mathbf{P}\left\{T_n \le \mathbf{E}T_n + x\sqrt{\mathbf{D}T_n}\right\}$$
$$= \mathbf{P}\{S^*_{\mathbf{E}T_n + x\sqrt{\mathbf{D}T_n}} \le n\} \to \Phi(x),$$

where $\Phi(x)$ is the standard normal distribution function.

Theorem 2

If m, n, t, $N \to \infty$ in such a way that m has the order $N^{1/4}$ and $n = o(m)$, then for any fixed $x \in \mathbb{R}$ and

$$t = 2N\left(\frac{1}{n} - \frac{1}{m}\right) + (1 + o(1))x\,\frac{2N}{\sqrt{3}}\sqrt{\left(\frac{1}{n^3} - \frac{1}{m^3}\right)},$$

the following relation is true:

$$\mathbf{P}\left\{|S_t| \le \frac{N}{N/m + t/2}\right\} \to \Phi(x)\,.$$

Avoine G., Junod P., Oechslin P. result

Hong J., “The cost of false alarms in Hellman and rainbow tradeoffs”, *Designs, Codes and Cryptography*, **57:3** (2010), 293–327:

«We will use the closed form approximation

$$\frac{m_k}{N} \approx \frac{1}{N/m_0 + k/2}$$

which can be found[1] in»

Avoine G., Junod P., Oechslin P., “Time-memory trade-offs: false alarm detection using checkpoints”, *Lect. Notes Comput. Sci.*, **3797** (2005), 183-196.

[1] The statement in the referenced paper is somewhat different, but this it due to multiple typographic errors. The version presented here can easily be obtained by following their proofs.

Thank you for attention!