Collision Search in a Random Mapping: some asymptotic results

E. Schulte-Geers

BSI

ECC 2000

Outline:

- **(1) Motivation, model, notation**
- **(2) How hard is it to hit the distinguished set?**
- **(3) Double-hits and collision double hits in the distinguished set**
- **(4) Summary**

This talk is about random mappings of a finite set N into itself.

Connections to the discrete log problem (DLP):

"ρ-methods"

Situation:

$G =< g >$ **finite cyclic group of prime order**
$|G| = p$**,**
(group op. written multiplicatively)
$h \in G$
(so, there is a unique $m \in \{0, \ldots, p-1\}$ such that $h = g^m$)

DLP:

given only g, h and G, compute $m = \log_g(h)$ efficiently!

(1) Pollard's ρ-method

(a) Pollard's idea: find a relation

$$g^i h^j = g^k h^l.$$

If $j - l \neq 0 \bmod p$**, then**

$\log_g(h) = \frac{i-k}{l-j} \bmod p$

(b) Find such a relation as follows:

(i) divide G **into sets** S_1, S_2, S_3 **(of roughly equal size).**

(ii) Define an "iteration function" $f : G \longrightarrow G$ **by**

$$x \mapsto \begin{cases} gx & \text{if} \quad x \in S_1 \\ x^2 & \text{if} \quad x \in S_2 \\ hx & \text{if} \quad x \in S_3 \end{cases}$$

(iii) Generate randomly a_0, b_0 **such that** $1 \leq a_0, b_0 \leq p-1$ **and compute** $x_0 := g^{a_0} h^{b_0}$**.**

Iterate f on x_0, i.e. compute $x_i = f(x_{i-1}) = g^{a_i}h^{b_i}$ until we find k, l such that $k \neq l$ and $x_k = g^{a_k}h^{b_k} = x_l = g^{a_l}h^{b_l}$.

Usually assumed:

f *behaves like a random mapping*

(2) Parallel Collision Search

(van Oorschot/Wiener '94,'99)

Collision Search Problem:

given: a mapping $f : N \longrightarrow N$ of a large set N into itself.

Problem: find a collision of f, i.e. find $a, b \in N$ such that $a \neq b$ and $f(a) = f(b)$

Parallelizable algorithm :

select a set $L \subset N$ of distinguished points

(i) (detection of collisions):

iterate f on different starting points x_i until the first time an iterate point $y_i = f^r(x_i)$ is in L (if possible).

Store the triples $(x_i, y_i, r =$ No. of iterations$)$.

Continue until a point y in L is hit twice.

(ii) (localisation of collisions)

if necessary, locate the (alleged) collision

Can be applied to solve the DLP *distributed on many processors* **in the same way as Pollard's ρ-method and gives** *linear* **speed up!**

Remark:

The algorithm finds double-hits in the distinguished set. Such a double hit may be caused by a collision (”collision double-hit”) or by a ”Robin Hood”.

Parallel collision search is currently the best method known to solve the DLP on (general) elliptic curves (e.g. some of the CERTICOM challenges).
Was applied to solve the ECC2K-108 challenge solved by the INRIA group in April 2000.

Questions:

(1) How high is the probability of success of the algorithm?

(2) How many points have to be stored, until the first (collision) double-hit in L occurs?

(3) How many iterations have to be performed until the first (collision) double-hit in L occurs?

(4) What is the effect of ”Robin Hoods”?

(5) How many starting points have missed the distinguished set at the (random) time the first double hit is recorded?

(6) How do the answers depend on the size of L?

Assumption:

*The iteration mapping behaves like a random mapping***.**

Answers are given in the following *model***:**

(two step random experiment)

(i) first step:

the digraph of a random mapping from N to N is generated (in which the points of L are distinguished)

(ii) second step:

non-distinguished **points are sampled** *without replacement* **from the digraph of this mapping. The mapping is iterated on each point until it enters L for the first time. The sampling process is continued, until the goal (first double hit in L or similar) is reached.**

(Sampling of non-distinguished points is assumed to be easy.)

and in the form of

limit theorems **for the distributions of random variables as** $n, l \longrightarrow \infty$**.**

Notation:

$N := \{1, \ldots, n\}$

$L \subset N$ **is the set of distinguished points,**
$card(L) = l$

$\Omega^{(n)}$**:= the set of mappings from** N **to** N**, is the space of elementary events**

"Random" mapping always refers to the uniform distribution on $\Omega^{(n)}$

$T(x) := \sum_{n \geq 1} \frac{n^{n-1}}{n!} x^n$ **is the "tree function"**

$N(0, 1)$ **denotes a standard normal r.v.,** φ **resp.** Φ **denote the density resp. c.d.f. of the standard normal distribution**

Previous and related work:

”Epidemic processes” on a random graph:

Gertsbakh (1977):
introduction, preliminary results

Burtin (1980):
exact and asymptotic distributions for the no. of ancestors of L (inverse epidemic process)

Pittel (1983):
exact and asymptotic distributions for the no. of successors of L and the no. of relatives of L

Jaworski (1990s):
generalization and simplification of *Burtin‘s* and *Pittel‘s* results

”Collision search:”

vanOorschot/Wiener (1994/1999):
answers to the questions above when $\theta = \frac{l}{n}$ is a ”sufficiently small” percentage of all points, some questions left open or dealt with ”ad hoc”.

In this talk:

rigorously proven asymptotic results in the random mapping model, all sizes of L considered, answers to open questions

In the sequel, we will let the size n tend to infinity. The corresponding a distinguished sets L of cardinality l will be called

small	**if** $l = o(\sqrt{n})$
of *critical* **size**	**if** $l = \alpha\sqrt{n}$**,** $\alpha > 0$
large	**if** $\sqrt{n} = o(l)$

Definition:

Let $\omega : N \longrightarrow N$ **be a mapping.**

A *non-distinguished* **point** x **is an ancestor of** L **under** ω**, if some positive power of** ω **maps it into** L**.**

If x **is an ancestor of** L **we also say "**x **hits** L**".**

How hard is it to hit the distinguished set?

Question 1:

(1) How many ancestors does L have?

Answer 1:

Let $A_L^{(n)} : \Omega^{(n)} \longrightarrow \mathbf{N}$ denote the r.v.

$A_L^{(n)}(\omega) :=$

no. of ancestors of L in the digraph of ω

Proposition: (Burtin)

for $j \in \{0, \ldots, n-l\}$

$$P(A_L^{(n)} = j) =$$

$$\binom{n-l}{j} (\frac{j+l}{n})^j (\frac{n-l-j}{n})^{n-l-j} \frac{l}{l+j}$$

Remark:(Jaworski)

$$E(A_L^{(n)}) = l\,\frac{(n-l)!}{n^{n-l}} \sum_{j=0}^{n-l-1} \frac{n^j}{j!},$$

$$E(l + A_l^{(n)})^2 = \frac{l^2\,n}{2} + \frac{1}{2}\,(n - l(l-1))\,E(l + A_l^{(n)})$$

Burtin's Theorem: (no. of ancestors)

(1) let l be fixed, $n \longrightarrow \infty$. Then

$$P(A_L^{(n)} = j) \longrightarrow l\,\frac{(l+j)^{j-1}}{j!}\,e^{-(l+j)}$$

(2) let $l \longrightarrow \infty, n \longrightarrow \infty, l = o(\sqrt{n})$. Then

$$\frac{A_L^{(n)}}{l^2} \longrightarrow \frac{1}{N(0,1)^2}$$

Interpretation of (1) and (2):

”small” distinguished sets L (i.e. $l = o(\sqrt{n})$) will typically have l^2 ancestors

(3) let

$l \longrightarrow \infty, n \longrightarrow \infty, \frac{l}{\sqrt{n}} \longrightarrow \alpha, 0 < \alpha < \infty$**.**

Then

$$\frac{A_L^{(n)}}{n} \longrightarrow \frac{\alpha^2}{\alpha^2 + N(0,1)^2}$$

Interpretation of (3):

"critical" distinguished sets L (i.e. $l = \alpha\sqrt{n}$)

will typically have a representative fraction of all

points as ancestors

(4) let $\frac{l}{\sqrt{n}} \longrightarrow \infty, l = o(n)$. Then

$$\frac{l^2}{n^2}(n - l - A_L^{(n)}) \longrightarrow N(0,1)^2$$

(5) let $\frac{l}{n} \longrightarrow \theta, 0 < \theta < 1$. Then

$$P(n - l - A_L^{(n)} = j) \longrightarrow \theta(1-\theta)^j \, \frac{j^j}{j!} \, e^{-j(1-\theta)}$$

Interpretation of (4) and (5):

”large” distinguished sets L (i.e. $\sqrt{n} = o(l)$)
will typically have only $\frac{n^2}{l^2}$ non-ancestors

Especially (Burtin):

$$\frac{A_L^{(n)}}{n-l} \longrightarrow 0 \quad \textbf{if} \quad \frac{l}{\sqrt{n}} \longrightarrow 0$$

$$\frac{A_L^{(n)}}{n-l} \longrightarrow 1 \quad \textbf{if} \quad \frac{l}{\sqrt{n}} \longrightarrow \infty$$

$\sqrt{n}$ is the threshold value for l:

if $l = o(\sqrt{n})$ only a negligible fraction of the non-distinguished points will hit L,

if $\sqrt{n} = o(l)$ only a negligible fraction of the non-distinguished points will miss L

Question 1':

How probable is it, that a randomly chosen non-distinguished point x hits L?

Answer 1':

$$p_{x,L} = \frac{E(A_L^{(n)})}{n-l} \approx \begin{cases} \frac{l}{\sqrt{n}} \sqrt{\frac{\pi}{2}} & \textbf{if } l = o(\sqrt{n}) \\ \alpha \, \frac{\Phi(\alpha)}{\varphi(\alpha)} & \textbf{if } n = \alpha \sqrt{n} \\ 1 - \frac{n}{l^2} & \textbf{if } \sqrt{n} = o(l) \end{cases}$$

Question 2:

suppose a randomly chosen non-distinguished point x hits L. How many iterations are required?

Answer 2:

Let $T_l^{(n)} := T_{x,L}^{(n)} : \Omega^{(n)} \longrightarrow \mathbf{N} \cup \infty$ **be the r.v.**

$$T_{x,L}^{(n)}(\omega) := \inf\{\, k \in \mathbf{N} \ : \ \omega^k(x) \in L\} \quad (\inf \emptyset := \infty)$$

Proposition: (iteration time)

let $n, l \longrightarrow \infty$

(1) if $l = o(\sqrt{n})$

$$\frac{T_l^{(n)}}{\sqrt{n}} \Big| \{T_l^{(n)} < \infty\} \longrightarrow |N(0,1)|$$

$$E\left(\frac{T_l^{(n)}}{\sqrt{n}} \Big| \{T_l^{(n)} < \infty\}\right) \longrightarrow \sqrt{\frac{2}{\pi}}$$

(2) if $l = \alpha \sqrt{n}$

$$\frac{T_l^{(n)}}{\sqrt{n}} \Big| \{T_l^{(n)} < \infty\} \longrightarrow \frac{\varphi(\alpha + x)}{\Phi(-\alpha)} 1_{]0,\infty[}(x)\, dx$$

$$E\left(\frac{T_l^{(n)}}{\sqrt{n}} \Big| \{T_l^{(n)} < \infty\}\right) \longrightarrow \frac{\varphi(\alpha)}{\Phi(-\alpha)} - \alpha$$

Interpretation of (1) and (2):

if $l = O(\sqrt{n})$**, and a randomly chosen non-distinguished point hits** L**, then typically the number of iterations will be of order** $\sqrt{n}$

(3) if $\sqrt{n} = o(l), l = o(n)$

$$\frac{l}{n} T_l^{(n)} \Big| \{T_l^{(n)} < \infty\} \longrightarrow Exp(1)$$

$$E\left(\frac{l}{n} T_l^{(n)} \Big| \{T_l^{(n)} < \infty\}\right) \longrightarrow 1$$

(4) if $\frac{l}{n} \longrightarrow \theta, 0 < \theta < 1$

$$T_l^{(n)} \Big| \{T_l^{(n)} < \infty\} \longrightarrow 1 + Geom(\theta)$$

$$E\left(T_l^{(n)} \Big| \{T_l^{(n)} < \infty\}\right) \longrightarrow \frac{1}{\theta}$$

Interpretation of (3) and (4):

if l is large (i.e. $\sqrt{n} = o(l)$), then one finds (iterating a random mapping on a random non-distinguished point x), the first point from L after $\frac{n}{l}$ iterations, without a memory for previous iterations.

Question 3:

How many non-distinguished points must be tried, until the first one to hit L is found ?

Answer 3:

Let $W_l^{(n)} := \Omega^{(n)} \longrightarrow \mathbf{N} \cup \infty$ **denote the r.v.**
$W_l^{(n)}(\omega)$ **= "no. of non-distinguished points collected, until the first ancestor of** L **under** ω **is found"**

Theorem (collection time)

(1) let l **be fixed,** $n \longrightarrow \infty$**, then for each** $x, 0 < x < 1$

$$P(W_l^{(n)} > n\,x) \longrightarrow e^{l(T((1-x)e^{-1})-1)}$$

$$E(\frac{W_1^{(n)}}{n} \mid W_1^{(n)} < \infty)) \longrightarrow \frac{1}{2}\,\frac{1-2e^{-1}}{1-e^{-1}}$$

for $l \geq 2$

$$E(\frac{W_l^{(n)}}{n} \mid W_l^{(n)} < \infty))$$

$$\longrightarrow \frac{1 - l\, e^{-l}}{(l-1)^2(1-e^{-l})} - \frac{e^{-l}}{1-e^{-l}}$$

Interpretation:

if l **is fixed : with asymptotic positive probability of** e^{-l}**,** L **cannot be hit at all. If** L **can be hit, an essential fraction of all points must be tried.**

(2) let $l \longrightarrow \infty, n \longrightarrow \infty, l = o(\sqrt{n})$**. Then** $P(W_l^{(n)} < \infty) \longrightarrow 1$ **and for each** $x > 0$

$$P(W_l^{(n)} > \frac{n}{l^2}\, x) \longrightarrow e^{-\sqrt{2x}}$$

further

$$E(\frac{l^2}{n}\, W_l^{(n)} \mid W_l^{(n)} < \infty) \longrightarrow 1$$

Interpretation:

In this case, L **will asymptotically be hit, and a**

number of starting points of magnitude $\frac{n}{l^2}$ must be tried until the first hit is found.

(3) let $\frac{l}{\sqrt{n}} \longrightarrow \alpha$, then

$$P(W_l^{(n)} < \infty) \longrightarrow 1$$

and for each natural number j

$$P(W_l^{(n)} > j) \longrightarrow \frac{1}{2^{j-1}\,(j-1)!} \int_0^\infty z^{2j-1}\, e^{-\alpha z - \frac{1}{2} z^2}\, dz$$

further

$$E(W_l^{(n)} \,|\, W_l^{(n)} < \infty) \longrightarrow 1 + \frac{1}{\alpha^2}$$

Interpretation:

In this case, L will asymptotically be hit after a few tries.

(4) let $\frac{l}{\sqrt{n}} \longrightarrow \infty$ then

$P(W_l^{(n)} = 1) \longrightarrow 1$ **and** $E(W_l^{(n)} \,|\, W_l^{(n)} < \infty) \longrightarrow 1$

Double hits and collision double hits in the distinguished set

Questions 4, 5

How many different starting points must be tried until the first (collision) double hit is recorded?

Answers 4,5

Let $D_l^{(n)} : \Omega^{(n)} \longrightarrow N \cup \{\infty\}$ **be the r.v.**

$D_l^{(n)}(\omega) :=$ **" no. of starting points tried for the first double hit"**

and

let $D_{l,coll}^{(n)} : \Omega^{(n)} \longrightarrow N \cup \{\infty\}$ **be the r.v.**

$D_{l,coll}^{(n)}(\omega) :=$ **" no. of starting points tried for the first collision double hit"**

Theorem (no. of starting points for the first double hit)

(1) (the "critical" case)

let $l \longrightarrow \infty, n \longrightarrow \infty, \frac{l}{\sqrt{n}} \longrightarrow \alpha$**, then**

$$P(D_l^{(n)} < \infty) \longrightarrow 1$$

and for every natural number $k \geq 1$

$$P(D_l^{(n)} > k)$$
$$\longrightarrow \sum_{j=0}^{k} \binom{k}{j} \frac{(2j)!}{j!\,2^j}\,\alpha^{k-j} \int_0^\infty \frac{z^{k+j-1}}{(k+j-1)!}\, e^{-\alpha z - \frac{1}{2}z^2}\, dz$$

further

$$E(D_l^{(n)}) \longrightarrow 2 + \alpha\sqrt{\frac{\pi}{2}} + R(\alpha)$$

where $R(\alpha) \longrightarrow 0$

(monotonely decreasing) as $\alpha \longrightarrow \infty$

Remarks

In the critical case, asymptotically:

(1) the least number of starting points and the least memory are needed.

(2) the starting points miss L with positive probability, and each starting point needs a no. of iterations of order $\sqrt{n}$ to hit L

(3) $P(D_l^{(n)} = 2)$ becomes maximal (among all possible l) for $l = \alpha_2 \sqrt{n}$, $\alpha_2 = 0.777733$, for this α_2

$$P(D_l^{(n)} = 2) \longrightarrow 0.15979$$

(4) the expected no. of needed starting points becomes minimal (among all possible l) for $l = \alpha_{\min} \sqrt{n}$,
(empirically $\alpha_{\min} \approx 2$)

(5) the same asymptotic assertions hold for $D_{l,coll}^{(n)}$. (So, every double hit in L will be a collision double hit).

(2) (double hits and iterations in the ”large” case)

let $l \longrightarrow \infty, n \longrightarrow \infty, \sqrt{n} = o(l), \frac{l}{n} \leq \gamma < 1$**, then**

$$P(D_l^{(n)} < \infty) \longrightarrow 1$$

and for every real $x > 0$

$$P(D_l^{(n)} > x \, \frac{l}{\sqrt{n + 2\,l}}) \longrightarrow e^{-\frac{1}{2}\,x^2}$$

further

$$E(\frac{\sqrt{n + 2\,l}}{l} \, D_l^{(n)}) \longrightarrow \sqrt{\frac{\pi}{2}}$$

and for the number $I_l^{(n)}$ **of iterations needed for the first double hit:**

$$\frac{\sqrt{n + 2\,l}}{n} \, I_l^{(n)} - \frac{\sqrt{n + 2\,l}}{l} \, D_l^{(n)} \longrightarrow 0$$

(3) (collision double hits and iterations in the "large" case)

let $l \longrightarrow \infty, n \longrightarrow \infty, \sqrt{n} = o(l), \frac{l}{n} \leq \gamma < 1$**, then**

$$P(D_{l,coll}^{(n)} < \infty) \longrightarrow 1$$

and for every real $x > 0$

$$P(D_{l,coll}^{(n)} > x \, \frac{l}{\sqrt{n}}) \longrightarrow e^{-\frac{1}{2}\, x^2}$$

further

$$E(\frac{\sqrt{n}}{l} \, D_{l,coll}^{(n)}) \longrightarrow \sqrt{\frac{\pi}{2}}$$

and for the number $I_{l,coll}^{(n)}$ **of iterations needed for the first double hit:**

$$\frac{1}{\sqrt{n}} \, I_{l,coll}^{(n)} - \frac{\sqrt{n}}{l} \, D_{l,coll}^{(n)} \longrightarrow 0$$

Remarks

So, in the large case (asymptotically)

(1) for every run of the parallel detection step:

no. of iterations = $\frac{n}{l} \times$ **no. of stored triples**

(2) effect of ”Robin Hoods”:
the ”average” probability, that a randomly chosen double hit is a ”Robin Hood” is $\approx \frac{2l}{n+2l}$**.**
If $\frac{l}{n} \longrightarrow \theta$**, Robin Hoods are not negligible!**
In this case it is easier to find double hits than collision double hits.

(3) the results of Wiener/vanOorschot are reaffirmed for the domain of l **where**
$\sqrt{n} << l << n$

(4) every starting point hits the distinguished set, therefore the no. of iterations is well defined. In the other cases one must also deal with ”misses” of L**.**

Question 6

How many starting points have missed L at the time the first double hit in L is recorded?

Answer 6

Let $M(D_l^{(n)})$ be the r.v. "no. of starting points that missed L at the (random) time $D_l^{(n)}$"

Theorem (no. of misses at the time of the first double hit)

$l \longrightarrow \infty, n \longrightarrow \infty, \frac{l}{\sqrt{n}} \longrightarrow \alpha, 0 < \alpha < \infty$.

Then $P(M(D_l^{(n)}) = \cdot)$ converges to a discrete probability distribution $q(\alpha, \cdot)$

$$q(\alpha, 0) = 1 - \sqrt{\frac{\pi}{2}} \frac{1}{\alpha} + \frac{1}{\alpha} \frac{\Phi(\alpha)}{\varphi(\alpha)}$$

$$q(\alpha, 0) + q(\alpha, 1) = 1 - \frac{3}{\alpha^2} + O(\alpha^{-3}) \quad (\alpha \longrightarrow \infty)$$

Final remarks

Aim of the algorithm, as applied to elliptic curve logs: to find double hits in L

(1) for success, one must be able to store $\approx \frac{l}{\sqrt{n+2l}}$ tripels, and to perform $\approx \frac{n}{\sqrt{n+2l}}$ iterations

(2) "small" distinguished sets are useless for practical collision search

(3) there are wide possibilities for time-memory trade-off, l may vary from $\alpha\sqrt{n}$ (α large enough) up to $(1-\epsilon)n$

(4) if $\frac{l}{n} \longrightarrow \theta > 0$ double hits are *easier* to find than collision double hits.
Finding double hits needs (on the average) a number of iterations $\frac{1}{\sqrt{1+2\theta}}\sqrt{\frac{\pi}{2}n}$, (and can be made arbitrarily close to $\sqrt{\frac{\pi}{6}\,n}$).
But for this, a memory of size $\sqrt{n}$ is needed.

(5) Limited memory

Situation:

(a) L **is "large"**

(b) only memory of size $p\,\frac{l}{\sqrt{n+2l}}$ **is available**

Strategy:

Phase 1: fill the memory with distinguished points

Phase 2: try to hit a point in the memory

Effects on the performance of the detection step:

(i) the average work to be done increases by a work factor $W(p)$ **compared to the "unlimited memory" situation.**

$$W(p) = 1 - 2\,\Phi(-p) + 2\,\frac{\varphi(p)}{p}$$

(fits experimental data given in Escott et alii.)

(ii) the fluctuations will be much larger

(6) with some more effort, also asymptotic expansions for relevant parameters could be given.