A simple method for precisely determining complexity of many Birthday attacks

Ravi Montenegro
University of Massachusetts Lowell
Email: Ravi_Montenegro@uml.edu *

June 2, 2012

Abstract

We show a simple and yet very precise upper bound on self-intersection time of a Markov chain, i.e. the expected number of steps until some state has been visited twice. When applied to certain birthday attacks the bound matches even the lead coefficient in simulation data to over 4 significant digits. This precision makes it possible to explain the differing performance between attacks that are superficially similar, such as Pollard's Rho for discrete logarithm and Teske's additive walks, and can also be used in optimizing birthday attack design.

1 Introduction

Birthday attacks use probabilistic "paradoxes" to solve cryptographic problems. However, the heuristics used to justify these attacks are usually not strong enough to differentiate a more efficient approach from a less efficient one. We remedy this shortcoming with a very precise result that applies to those attacks modeled on a Markov chain.

The Birthday Paradox states that if samples are taken with replacement from a set of N distinct items, then the expected number of samples required until some item is chosen twice is $(1+o(1))\sqrt{\frac{\pi}{2}\,N}$. This can be interpreted as a statement that a Markov chain on the complete graph K_N with transitions $P(i,j) = 1/N$ will have a *self-intersection* or *collision* in expected time $\sqrt{\frac{\pi}{2}\,N} \sim 1.253\sqrt{N}$. For a more general Markov Chain if the Markov chain draws nearly uniform samples after some $\tau \ll N$ steps then a common heuristic is to assume then the standard birthday paradox more-or-less explains what happens subsequently, as each step generates a sample which is nearly uniform and independent of those τ or more steps earlier. However, as noted by Teske [5], simulations show neither this is not quite correct, with some walks colliding nearly as quickly as predicted by the birthday paradox, while other walks are much slower. For instance, Pollard's Rho walk on the cycle $\mathbb{Z}_N$ has transitions $\mathsf{P}(i,i+1) = \mathsf{P}(i,i+k) = \mathsf{P}(i,2i) = 1/3$ for some fixed constant k and reportedly requires around $1.596\sqrt{N}$ steps, while Teske's r-adding walk with $\mathsf{P}(i,i+s_j) = 1/r$ for a random set of elements $\{s_j\}_{j=1}^r$ requires $\omega(\sqrt{N})$ steps in the 3-adding case but a nearly optimal $1.292\sqrt{N}$ in the 20-adding case [5].

Although there are some theoretical results in this area [3, 2, 5] and a sharper heuristic [1], it is not well understood what causes two walks of equal degree to perform very differently, unless one approaches its stationary distribution dramatic slower than the other. Our new result shows,

[*] Much of this work was done while the author was a guest at the Tokyo Institute of Technology.

rather counterintuitively, that if two independent copies of the walk are started from the same state then the faster they are expected to intersect each other the slower the expected collision (self-intersection) time of the Markov chain. Or more simply, self-intersection time is largely determined by how locally tree-like the underlying graph is: the fewer short cycles there are then the faster the collision time.

The paper proceeds as follows. In Section 2 we develop our new heuristic and compare it to previous ideas. The heuristic is then used to study several birthday attacks in Section 3 and found to be extremely precise.

2 Collision time: Old and New Heuristics

In 1978 Pollard introduced his Rho algorithm for finding discrete logarithm [4] over a prime-order cyclic group $G = \langle g \rangle$. Given some $X_0 = g^{a_0} h^{b_0} = g^{a_0 + b_0 k}$ and partition $S_1 \coprod S_2 \coprod S_3 = G$, with S_i all of sizes roughly $|G|/3$, consider the process with

$$X_{i+1} = \begin{cases} X_i g & \text{if } X_i \in S_1 \\ X_i h & \text{if } X_i \in S_2 \\ X_i^2 & \text{if } X_i \in S_3 \end{cases}$$

Floyd's cycle finding method can be used to find when a cycle occurs and the same group element $X_i = X_j$ has been reached twice, i.e. a self-intersection. It is easy to track the exponents (a_i, b_i) with $X_i = g^{a_i} h^{b_i}$, and so once $X_i = X_j$ then $a_i + b_i k = a_j + b_j k \mod |G|$ and so $k = (a_i - a_j)(b_j - b_i)^{-1} \mod |G|$, determining discrete logarithm k except in the degenerate case when $b_i \equiv b_j \mod |G|$.

Typically the partition is fixed by a hash function, so it is not fully random. In this case no method is known for precisely studying the Rho method. However, this is essentially a deterministic implementation of the Markov chain $\mathsf{P}(\mathfrak{g}, \mathfrak{g}g) = \mathsf{P}(\mathfrak{g}, \mathfrak{g}h) = \mathsf{P}(\mathfrak{g}, \mathfrak{g}^2) = 1/3$ on G, and if the partition is chosen by randomly and independently assigning each group element to a partition then this is exactly a Markov chain, at least until it self-intersects by visiting a previously visited state. In this setting the methods of [3, 2, 1] apply, with the analysis simplified by the observation that this is equivalent to a walk on $\mathbb{Z}_N \times \mathbb{Z}_N$ with $\mathsf{P}((a_i, b_i), (a_i + 1, b_i)) = \mathsf{P}((a_i, b_i), (a_i, b_i + 1)) = \mathsf{P}((a_i, b_i), (2a_i, 2b_i)) = 1/3$.

Pollard suggested that the run time of the Rho method should be similar to that of the birthday problem, a heuristic which is correct for order of magnitude but off by about 30% for the lead constant. Much later, Teske proposed speeding this up by considering an alternative "additive" walk, and simulations found this to be only 3% slower than a birthday problem [5].

In recent years several methods have been proposed for studying these walks. Blackburn and Murphy [1] give an improved heuristic suggesting that a walk on a regular digraph of degree r has collision time of $(1 + o(1))\sqrt{\frac{r}{r-1}}\sqrt{\frac{\pi N}{2}}$ steps, a good but not exactly sharp result which cannot differentiate between fast and slow walks of equal degree. Miller and Venkatesan [3] give the first rigorous result of order nearly $O(\sqrt{N})$, that collision time is bounded by $O(\sqrt{N}\,\tau_s(1/2))$ where $\tau_s(\epsilon) = \min\{n : \forall u, v \in V,\, P^n(u, v) \geq (1 - \epsilon)\pi(v)\}$ is mixing time in separation distance. Kim, Montenegro, Peres and Tetali [2] improve this to $O(\sqrt{N}\,\max\{A_\tau, A_\tau^*\})$, where

$$\begin{aligned} \forall x, y \in V \quad &: \quad \frac{1 - \epsilon}{N} \leq \mathsf{P}^\tau(x, y) \leq \frac{1 + \epsilon}{N} \\ A_\tau \quad &= \quad \max_{v \in V} 1 + \sum_{i,j=1}^{\tau} \mathsf{Pr}\left(X_i = Y_j \mid X_0 = Y_0 = v\right) \end{aligned}$$

i.e. $\tau := \tau_\infty(1/2)$ is the L^∞ mixing time, A_τ is the expected number of collisions in the neighborhood of two walks starting at the same state, and A^*_τ is the same for the time-reversal (adjoint) $\mathsf{P}^*(x,y) = \frac{\pi(y)\mathsf{P}(y,x)}{\pi(x)}$. In the special case of Pollard's Rho walk they use this to prove expected collision time of at most $(52.5 + o(1))\sqrt{N}$ steps, the first rigorous proof of $O(\sqrt{N})$.

In spite of these improvements it is still not clear what properties of a random walk govern its collision time. For instance, the heuristic of Blackburn and Murphy cannot explain the differing performance of the Rho and 3-adding walks, whereas the bound of Kim et.al. can explain this but is not sharp enough to differentiate the Rho walk from the 20-adding walk. We remedy this shortcoming with a very precise result that applies to attacks modeled on a Markov chain. In particular, we find the following:

Heuristic 2.1. *The expected collision time of Markov chain* P *on a state space of cardinality* N *is*

$$T + (1 + o(1))\sqrt{\frac{\pi}{2}\, N A_T (1+\epsilon)}$$

The value of A_T can be estimated recursively, since if the walks intersect in some small number of steps t then another $A_{T-t} \approx A_T$ subsequent collisions are to be expected, so that

$$\begin{aligned} & A_T \leq 1 + \mathsf{Pr}\left(\exists i, j \in [1, 2, \ldots, t],\, X_i = Y_j \mid X_0 = Y_0\right) A_T \\ \Rightarrow\quad & A_T \leq \frac{1}{1 - \mathsf{Pr}\left(\exists i, j \in [1, 2, \ldots, t],\, X_i = Y_j \mid X_0 = Y_0\right)} \end{aligned} \tag{1}$$

After a small number of steps the walks are usually sufficiently randomized that if they haven't intersected yet then there is a negligible probability of them colliding, so the early collision probabilities will give a good estimate for A_T.

In Section 3 this recursive estimate is used to study important birthday attacks and it is found that this matches simulation results to at least 4 significant digits.

The accuracy of the heuristic suggests the following counterintuitive principle:

Heuristic 2.2. *If two independent copies of a walk are started from the same state then the faster they are expected to intersect each other the slower the expected collision self-intersection time of a single copy of the Markov chain. Or more simply, self-intersection time is largely determined by how locally tree-like the underlying* undirected *graph is: the fewer short cycles there are then the faster the collision time.*

Remark 2.3. In rare cases the t-step estimate on A_T does not approach a constant as t increases, in which case it is also necessary to determine T. Such a situation arises in Teske's additive walk when there are fewer than 5 generators [5], a case in which we are able to use ideas similar to those here to show $A_T = \omega(1)$ and collision time $\omega(\sqrt{N})$ (unpublished work).

To see why the heuristic is appropriate, note that once a collision occurs then it will be part of a sequence of around A_T collisions on average, so in particular a randomly chosen intersection has about a $1/A_T$ chance of being the first in such a sequence. As a result, if $i \geq j + T$ are far enough apart that X_i and X_j are minimally correlated then while

$$\mathsf{Pr}\left(X_i = X_j \mid i \geq j + T\right) \approx \frac{1}{N}$$

is the probability of a collision,

$$\mathsf{Pr}\left(X_i = X_j \wedge (\text{no intersection in previous } T \text{ steps of } X_i) \mid i \geq j + T\right) \approx \frac{1}{N A_T} \tag{2}$$

is the probability that there have been no recent collisions as well. Then the relation $\mathsf{Pr}\left(X_i = X_j\right) = 1/N$ used in standard proofs for the Birthday Problem can be replaced by (2), leading to a bound on expected time until the first collision which was not recently proceeded by other collisions, i.e. the expected time until first collision. This effectively replaces the N in $\sqrt{\frac{\pi}{2}N}$ with $N\,A_T$ and an initial warm-up of some T steps, leading to Heuristic 2.1 above. Furthermore, after the first T steps the subsequent states will be drawn from a nearly uniform distribution, so if only collisions after the first T steps are considered then once a collision occurs it will be at a fairly typical state, and so A_T need only be calculated for such typical states.

For most walks of interest in number theory there is a canonical notion of direction for the walk. When this is not the case then a collision $X_i = X_j$ may be followed by collisions with $X_{j+\Delta}$ intersection states visited before X_i, i.e. the time-reversed walk P^* may quickly intersect the original non-reversed walk when both start from the same state. When this is not the case, i.e. P^* has a non-trivial chance of intersecting P quickly, then the definition of A_T might be changed to

$$\begin{aligned} A_T = \max_{average\ state\ v} 1+ \quad \sum_{i,j=1}^{T} \quad & \mathsf{Pr}\left(X_i = Y_j \mid X_0 = Y_0 = v\right) \\ & +\mathsf{Pr}\left(X_i = Y_j^* \mid X_0 = Y_0 = v\right) \end{aligned}$$

with (1) redefined to consider when either $X_i = Y_j$ or $X_i = Y_j^*$. However, this is not quite right as the walk referred to as Y_j^* above cannot visit states visited before X_i, as otherwise X_i would not be the time of the first intersection, so this seems unlikely to give the correct result unless the graph has very high degree.

Remark 2.4. Blackburn and Murphy give a heuristic based on an idea of Brent and Pollard [1]. Consider a walk X_i on a directed regular graph of constant in and out-degrees r. If the first collision occurs at time $i+1$ then X_i is an initial vertex on a directed edge into a previously visited state – $\{X_0, X_1, \ldots, X_{i-1}\}$ – and the walk transitions along this edge. There are approximately $(r-1)i$ such directed edges in the graph, and so X_{i+1} is a collision with probability approximately $\frac{(r-1)i}{N}\frac{1}{r}$. In contrast, in the Birthday problem the equivalent probability of X_i colliding with an earlier state is $\frac{i}{N}$, and so the new heuristic effectively replaces N by $\frac{r}{r-1}N$ in the Birthday problem, leading to expected time:

$$(1+o(1))\sqrt{\frac{r}{r-1}}\sqrt{\frac{\pi}{2}N}$$

In our heuristic the simplest non-trivial estimate for A_T is found by using $\mathsf{Pr}\left(X_1 = Y_1\right) \geq \frac{1}{r}$, which implies $A_T \geq \frac{1}{1-1/r} = \frac{r}{r-1}$ and expected time until a collision of at least $(1+o(1))\sqrt{\frac{r}{r-1}}\sqrt{\frac{\pi}{2}N}$, the same as Blackburn and Murphy's heuristic. Adding further terms to the estimate for A_T in (1) improves on the old heuristic.

3 Pollard's Rho and Teske's Walk

In this section we apply our new heuristic to two random walks related to the Discrete Logarithm Problem, Pollard's original Rho walk and Teske's additive version. We finish by using our heuristic to develop a walk for Discrete Logarithm that is (very slightly) faster than either of these, demonstrating how our result may be helpful in speeding up birthday attacks.

Example 3.1 (Pollard's Rho)**.** Recall that Pollard's Rho walk on $\mathbb{Z}_N$ for prime order N has transitions $\mathsf{P}(x, x+1) = \mathsf{P}(x, x+k) = \mathsf{P}(x, 2x) = 1/3$ for some constant k. Simulations by Teske indirectly estimate collision time to be $1.596\sqrt{N}$ steps [5], while a much larger simulation of our own [1] has 95% confidence interval of $1.6252\sqrt{N}$ to $1.6257\sqrt{N}$. These suggest that collision time in the Rho walk is about 30% slower than the $1.2533\sqrt{N}$ steps required in the birthday problem.

To apply our new method suppose that walks X_i and Y_j start at the same state $X_0 = Y_0$ and calculate the probability of intersection within a few steps:

- $\mathsf{Pr}\,(X_1 = Y_1 \mid X_0 = Y_0) = 1/3$
- $\mathsf{Pr}\,(X_2 = X_0 + 1 + k = Y_0 + k + 1 = Y_2,\, X_1 \neq Y_1 \mid X_0 = Y_0) = 2/3^4$
- $\mathsf{Pr}\,(X_3 = X_0 + 1 + k + k = Y_3,$ no prior $X_i = Y_j \mid X_0 = Y_0) = 2/3^6$

  Likewise for $X_3 = X_0 + 1 + 1 + k = Y_3$.
- $\mathsf{Pr}\,(X_3 = 2X_0 + 1 + 1 = 2(Y_0 + 1) = Y_2,$ no prior $X_i = Y_j \mid X_0 = Y_0) = 1/3^5$

  Likewise when the roles of X and Y are switched, and/or when $+k$ is used instead of $+1$, for a total of $\frac{4}{3^5}$.

Then

$$\mathsf{Pr}\,(\exists i, j \leq 3,\, X_i = Y_j \mid X_0 = Y_0) = \frac{1}{3} + \frac{2}{3^4} + \frac{4}{3^6} + \frac{4}{3^5} = 0.37997$$

and so Heuristic 2.1 then suggests collision time of

$$T + \sqrt{\frac{\pi}{2}N} \approx \sqrt{\frac{1.6128\pi}{2}N} = 1.592\sqrt{N}$$

To get a more accurate estimate a computer was used to enumerate all walks of length $t \leq 10$ and from this determine the probability of intersection in $\leq t$ steps. When $t > 10$ then 10^{10} runs of two independent walks were used estimate this probability, for a 95% margin of error of 10^{-5}.

t	$\mathsf{Pr}\,$(**intn in** $\leq t$ **steps**)	**Estimated collision time**
1	0.333333	$1.53499\sqrt{N}$
2	0.358025	$1.56423\sqrt{N}$
3	0.37997	$1.59168\sqrt{N}$
4	0.387289	$1.60115\sqrt{N}$
5	0.394808	$1.61107\sqrt{N}$
10	0.403805	$1.62318\sqrt{N}$
20	0.405482 ± 0.00001	$(1.62546 \pm 0.00001)\sqrt{N}$
40	0.405546 ± 0.00001	$(1.62555 \pm 0.00001)\sqrt{N}$
64	0.405545 ± 0.00001	$(1.62555 \pm 0.00001)\sqrt{N}$

In conclusion, a back of the envelope calculation of the first 3 steps suffices to explain 90% of the deviation from the birthday heuristic, a 10 step estimate accounts for 99.7% of this, and by 20 steps the bound is within our 95% confidence interval for the true collision time.

[1] We did 45 million runs of Pollard's Rho walk on $\mathbb{Z}_N$ with N ranging from 10^8 to 10^{13}, in each case determining the exact number of steps until the first collision. The mean collision time was $1.6254\sqrt{N}$ with sample standard deviation of $0.8495\sqrt{N}$. Simulations were run on an AMD Phenom II desktop CPU with the SIMD oriented Mersenne Prime Twister (SFMT) used to generate pseudorandom numbers.

Remark 3.2 (Optimizing Rho transitions)**.** The previous example can be generalized to find the optimal choice of transition probabilities for the additive and multiplicative steps. If additive steps are taken with probability p each and the multiplicative step with probability $1-2p$ then

$$\begin{aligned} &\mathsf{Pr}\left(\exists i,j\leq 3,\ X_i=Y_j \mid X_0=Y_0\right) \\ = \quad &(1-2p)^2+2p^2+2p^4+\frac{4p^6}{1-p^2}+\frac{4p^3(1-2p)^2}{1-p^2-p^3} \end{aligned}$$

This is minimized when $p=0.3071$, with an expectation of $1.590\sqrt{N}$ steps until collision. This shows that Pollard's suggestion that each transition have type have equal probability is at worst negligibly slower than the optimal choice of transition probabilities.

Example 3.3 (Teske's r-adding walks)**.** Recall the r-adding walk on $\mathbb{Z}_N$ has transitions $\mathsf{P}(x,x+s_k)=1/r$, where $s_k\in\{s_1,\,s_2,\,\ldots,\,s_r\}$ is one of r values fixed (but chosen randomly) from $\{1,2,\ldots,\,N-1\}$. Teske estimates average collision time of around $1.292\sqrt{N}$ steps [5], while our much larger run of simulations [2] have a 95% confidence interval of $1.2877\sqrt{N}$ to $1.2880\sqrt{N}$. These suggest that collision time is about 3% slower than the $1.2533\sqrt{N}$ steps required in the birthday problem.

This time the probability that two independent walks with $X_0=Y_0$ intersect within a few steps is

$$\begin{aligned} &\mathsf{Pr}\left(\exists i,j\leq 3,\ X_i=Y_j \mid X_0=Y_0\right) \\ = \quad &\frac{1}{r}+{}_r\mathrm{P}_2\frac{1}{r^2}\frac{1}{r^2}+\frac{3\,{}_r\mathrm{P}_3+2\,{}_r\mathrm{P}_2}{r^6} \\ = \quad &\frac{1}{r}+\frac{1}{r^2}+\frac{2}{r^3}+O(1/r^4) \end{aligned}$$

When $r=20$ then this leads to an estimate on collision time of

$$T+\sqrt{\frac{\pi A_T}{2}N}\approx\sqrt{\frac{1.0557\pi}{2}N}=1.287709\sqrt{N}$$

which is already within the 95% confidence interval given by simulation data. An exact enumeration of walks of length $t=5$ increases the estimate only negligibly to $1.287765\sqrt{N}$ steps, at $t=10$ to $1.287770\sqrt{N}$ steps, and the sampling based estimate at length $t=100$ gave an estimate of $(1.287769\pm 0.000003)\sqrt{N}$ with 95% confidence.

So in this case a mere 3 steps already explains 99.7% of the 20-additive walk's deviation from the birthday heuristic, and by 5 steps the estimate becomes essentially constant.

Example 3.4 (Mixed walks)**.** Teske also considers mixed walks with r additive and s doubling steps. This walk on $\mathbb{Z}_N$ has probability $\frac{1}{r+s}$ for each additive step, and $\frac{s}{r+s}$ for the doubling step.

The probability of two walks of length $t=3$ intersecting when started from the same state is

$$\begin{aligned} &\mathsf{Pr}\left(\exists i,j\leq 3,\ X_i=Y_j \mid X_0=Y_0\right) \\ = \quad &\frac{r+s^2}{(r+s)^2}+\frac{r(r-1)}{(r+s)^4}+\frac{3\,{}_r\mathrm{P}_3+2\,{}_r\mathrm{P}_2}{(r+s)^6}+\frac{4rs^2}{(r+s)^5} \end{aligned} \tag{3}$$

When $r=16$ and $s=4$ then this is probability 0.081985, and so our result predicts expected collision time of $1.308\sqrt{N}$, quite close to the $1.301\sqrt{N}$ found in Teske's simulations.

[2] We did 75 million runs under the same circumstances as the Rho walk, and found mean collision time of $1.2878\sqrt{N}$ with sample standard deviation of $0.6732\sqrt{N}$.

Remark 3.5 (More efficient attacks)**.** To finish up, observe that in the final example when $r+s = 20$ is fixed then the intersection probability is minimized at $s = 1$, with (3) inducing expected collision time of $1.2875\sqrt{N}$ versus $1.2877\sqrt{N}$ for the 20-additive walk ($r = 20$ and $s = 0$). Sampling 10^{10} random pairs of length 20 paths improves this to $(1.287571 \pm 0.000003)\sqrt{N}$, again better than the 20 step estimate of $(1.287771 \pm 0.000003)\sqrt{N}$ for the 20-adding walk.

While this is a negligible improvement, and within the margin of error of sampling based estimates for the true complexity of both walks, it demonstrates how our theoretical result may be useful for fine tuning proposed birthday attacks.

References

[1] S. Blackburn and S. Murphy, "The number of partitions in Pollard Rho," *Technical report RHUL-MA-2011-11 (Department of Mathematics, Royal Holloway, University of London, 2011)*, `http://www.ma.rhul.ac.uk/tech`

[2] J-H. Kim, R. Montenegro, Y. Peres and P. Tetali, "A Birthday Paradox for Markov chains, with an optimal bound for collision in the Pollard Rho Algorithm for Discrete Logarithm," *The Annals of Applied Probability*, vol. 20(2), pp. 495–521 (2010).

[3] S. D. Miller and R. Venkatesan, "Spectral analysis of Pollard rho collisions," In *Algorithmic number theory.* Lecture Notes in Comput. Sci., vol. 4076. Springer, Berlin, 573–581. (2006)

[4] J. Pollard, "Monte Carlo methods for index computation mod p," *Mathematics of Computation*, vol. 32(143), pp. 918–924 (1978).

[5] E. Teske, "Speeding Up Pollard's Rho Method for Computing Discrete Logarithms," *Proceedings of the 3rd International Symposium on Algorithmic Number Theory (ANTS-III)*, LNCS 1423, Springer, pp. 541–554 (1998).