Precise Probabilities for Hash Collision Paths

Max Gebhardt, Georg Illies and Werner Schindler

Bundesamt für Sicherheit in der Informationstechnik (BSI)
Godesberger Allee 185–189
53175 Bonn, Germany
{Maximilian.Gebhardt,Georg.Illies,Werner.Schindler}@bsi.bund.de

Abstract. We describe a generally applicable methodology to determine the probability of given differential (near-)collision paths in MD-type hash collision attacks (cf. [WY,WYiY,Kli2]). For MD5 this program is worked out explicitly. The probabilities of three (near-)collision paths are computed. Experiments confirm these results.

Keywords: Hash function, collision path, post addition, probability.

1 Introduction

In [WLFCY], [WY] and [WYuY] efficient collision search methods are described for the hash functions HAVAL, RIPEMD, MD4 and MD5 and SHA-0; improvements can be found in [St], [SNKO], [LiLa], [Kli1], [Kli2], [BCH]. In [WYiY] a collision attack on SHA-1 is sketched with a predicted workload of 2^{69} hash calculations and [WYaYa] announces an improvement with a workload of only 2^{63}.

At first we sketch a typical 2-block attack as described in the mentioned papers. Roughly speaking, the definition of a dedicated hash function H (e.g. MD5 or SHA-1) consists of the $IV \in \{0,1\}^t$, the padding rule (which is irrelevant for our considerations) and the compression function

$$h: \ \{0,1\}^t \times \{0,1\}^s \longrightarrow \{0,1\}^t.$$

The letters t and s denote the length of the hash value and the block length in bits, resp. Usually, t and s are multiples of 32 (e.g. $(t,s) = (128,512)$ for MD5 and $(t,s) = (160,512)$ for SHA-1). The compression function itself is a composition of simple step functions $h_i : \{0,1\}^t \times \{0,1\}^s \to \{0,1\}^t$ and a post addition given by a modular addition $+_w$ on (32-bit) words: If we define

$$h^{(N)}(a,b) = h_N(h_{N-1}(\ldots h_2(h_1(a,b),b),\ldots,b),b)$$

with N the number of steps ($N = 64$ for MD5, $N = 80$ for SHA-1) then

$$h(a,b) = h^{(N)}(a,b) +_w \widetilde{a} \tag{1}$$

where $\widetilde{a}$ denotes a word-wise permutation of a. A 2-block collision for H is a pair $(M_1||M_2, M_1'||M_2')$ with $M_i, M_i' \in \{0,1\}^s$ and $(M_1||M_2) \neq (M_1'||M_2')$ such that

$$h(h(IV,M_1),M_2) = h(h(IV,M_1'),M_2') \tag{2}$$

which implies $H(M_1||M_2) = H(M_1'||M_2')$ after padding.

The algorithms in [WY] and [WYiY] work as follows (we neglect details at this point):

1. (first block) Pairs of blocks (M_1, M_1') are produced in a specific manner (a part of this procedure is referred to as "message modification") until a pair $(h(IV,M_1), h(IV,M_1'))$ is found that satisfies specific bit conditions.

2. (second block) Depending on the pair (M_1, M_1') found in the first block message blocks (M_2, M_2') are generated in a specific manner (in particular, by another message modification) until one pair is found that satisfies (2).

[WY] and [WYiY] list a set of conditions on intermediate results during the step-by-step calculation of the value $h(h(IV, M_1), M_2)$ (called the "sufficient conditions" in [WY] and [WYiY]) such that the fulfillment of all these bit conditions shall not only imply the conditions after the post additions and (2) (so that a collision is obtained). Moreover, these conditions (shall) additionally ensure that intermediate values follow a specified differential path (a so-called collision path, resp. near-collision path). A more precise definition of the term "differential path" will be given later. Basically it is a combination of a 32-bit word modular differential path and a kind of signed XOR differential path.

To estimate the actual workload for finding a collision that follows a given differential path the authors of the papers mentioned above simply count the number of those bit conditions per block that are not automatically satisfied due to the generation process of the block pairs ("message modification" etc.). If, for instance, r_1 bit conditions remain in the first block then they argue that 2^{r_1} first-block message pairs have to be generated on average until one of them satisfies all conditions and thus the differential path. Accordingly, if r_2 conditions remain in the second block one has to produce about 2^{r_2} second-block pairs on average.

Experimental results in the MD5 case showed that these probability estimates are not exact. Hence we developed a more precise stochastic model and a generally applicable method which allows to determine (at least almost) exact probabilities of given differential paths, resp. that a given set of "sufficient conditions" is satisfied. For concrete computations we restricted ourselves to MD5.

There are some minor errors in the list of "sufficient" bit conditions contained in [WY] as was already observed e.g. in [YaSh], [SNKO],[LiLa]. Apart from that two other effects are neglected in [WY], which significantly affect the probability. The first one is that several additional conditions are not mentioned in [WY] (but also described in [HPR] and [Th]), which have to be satisfied to "stay on the differential path": The modular difference of some cyclically shifted 32-bit intermediate values must be equal to the modular difference of two register values. The second effect is the influence of the postadditions. We point out that the IV influences the success probability of concrete (near-)collision paths as was quantified in [GIS2] and qualitatively also mentioned in [St]. To our best knowledge, in particular the second effect has not been quantified elsewhere although it is non-negligible.

For the near-collision paths 1, 2 and 3 specified in the appendix the computed path probabilities were confirmed by experiments, underlining that the applied stochastic model may be regarded as appropriate. We point out that the probability of a particular (near-) collision path gives an upper bound for the workload of a collision attack since different (near-)collision paths may result in equal bit conditions after the postadditions. When the details of the announced attack on SHA-1 [WYaYa] become available it should also be possible to apply similar methods to get an upper bound for the actual workload. It should also be pointed out that the impact of the IV may be interesting for "prefix" attacks as described in [DL] and [GIS1].

The rest of this paper is organized as follows. After introducing some notation for registers, step functions etc. and sketching the general goal (Section 2) in Section 3 we prove three theorems that will turn out to be useful for the envisaged probability calculations. In Section 4 we introduce the stochastic model for MD5, i.e. we formulate a stochastic assumption. We quantify the impact of the post additions and calculate the overall probabilities of three MD5-near-collision paths.

Acknowledgement: We would like to thank Søren Thomsen for making his paper [Th] available to us.

2 The Goal

Generically, the compression function $h\colon \{0,1\}^t \times \{0,1\}^s \to \{0,1\}^t$ of a dedicated hash function H consists of the following steps:

1. (Input) chaining value $r_{(0)}$ (IV for the first block) and message block m
2. (Message Expansion) $m = (m_1, \ldots, m_{s/32}) \mapsto \widetilde{m} = (\widetilde{m}_1, \ldots, \widetilde{m}_N)$
3. (Initialization of the registers) for $i = 1$ to k do
   $r_{-k+i} := r_{(0),i}$
   where $r_{(0),i}$ denotes a particular word of the IV, resp. the chaining value.
4. (Step functions) for $i = 1$ to N do
   $r_i := F_i(r_{i-1}, \ldots, r_{i-k}, \widetilde{m}_i)$.
5. (Postadditions) for $i = N - k + 1$ to N do
   $r_i^p := r_i + r_{i-N} (\mathrm{mod} 2^{32})$
6. (Output) $(r^p_{N-k+1}, \ldots, r^p_N)$ (new chaining value).

Remark 1. (i) (Example) MD5: $(s,t,N,k) = (512,128,64,4)$, SHA-1: $(s,t,N,k) = (512,160,80,5)$, SHA-256: $(s,t,N,k) = (512,256,64,8)$.
(ii) The step function F_i usually depends on the Step number i.

For any hash function H a (one-block) collision can be found with complexity $O(2^{-t/2})$ ("birthday paradox"). Roughly speaking, the goal of a collision attack is to determine sufficient conditions on related message blocks (m, m') and on the intermediate register values $(r_1, r'_1), \ldots, (r_N, r'_N)$, $(r^p_{N-k+1}, r^{p\prime}_{N-k+1}), \ldots, (r^p_N, r^{p\prime}_N)$ such that $h(c,m) = h(c,m')$ (collision) or at least that $h(c,m)$ and $h(c,m')$ assume a determined difference (near-collision) 'preparing' a collision in one of the next blocks. Usually, there exists a number $N_1 < N$ such that a suitable (random) choice of (m, m') guarantees the conditions on the register values (r_j, r'_j) and the expanded message blocks $(\widetilde{m}_j, \widetilde{m}'_j)$ for all $j \leq N_1$ (message modification). The conditions after step N_1 shall be satisfied with a considerably larger probability than $2^{-t/2}$.

From Step $N_1 + 1$ to N (including the postadditions) the attacker just checks whether the intermediate register values (and maybe the message blocks) fulfil the given sufficient conditions (with the option of stopping the calculation of $(h(c,m), h(c,m'))$ early), or at least whether $h(c,m) = h(c,m')$, resp. (in a multiblock-collision) whether the register values after postaddition step meet specific requirements. In this paper we are interested in the probabilities of (near-)collision paths. That is, we are interested in the probability that the sufficient conditions after Step N_1 (end of the message modification) are fulfilled. Therefore, we interpret the register values and the extended message blocks as values that are assumed by random variables, which we denote with the respective capital letters. For the example of MD5 we work out a stochastic model which allows to determine (almost) exact probabilities. In particular, we show how to calculate the transition probabilities between the particular steps and quantify the impact of the postadditions, i.e. of the IV, on the probability of near-collision paths. In order to compute the exact probabilities of collision paths one has to determine the conditional probabilities (= transition probabilities)

$$\begin{aligned} &\mathrm{Prob}((R_i, R'_i) \mid R_{i-1}, R'_{i-1}, \ldots, \widetilde{M}_i, \widetilde{M}'_i, \ldots) \quad \text{and} \\ &\mathrm{Prob}((R^p_i, R'^p_i) \mid R_i, R'_i, R_{i-N}, R'_{i-N}, \ldots) \text{ (post addition).} \end{aligned} \tag{3}$$

The conditional parts comprise the whole prehistory up to Step i where the random variables $R_i, R'_i, \ldots$ meet specific path-dependent requirements. The following section provides three theorems that support this goal.

3 Some Useful Observations

In this section we derive three stochastic theorems that will turn out to be useful later. We begin with some definitions.

Notation. In the following $w[j]$ stands for the j^{th} bit of a 32-bit word w. The numbering starts at the least significant bit with 1.

For $M \in \mathbb{N}$ we define $Z_M := \{0, 1, \ldots, M-1\}$. For $a, b \in \mathrm{Z}_{2^{32}}$ the term $\Delta(a, b)$ denotes the modulo 2^{32}-difference of a and b, i.e. $\Delta(a, b) := (b - a) (\mathrm{mod}\ 2^{32})$.

Similarly as in [WY] we define $\Delta_B(a, b) := [\pm j_1, \ldots, \pm j_k]$ where $j_1, \ldots, j_k$ denote those bit positions where a and b are different. Here '$+j$', resp. simply 'j', means that $(a[j], b[j]) = (0, 1)$ while '$-j$' means that $(a[j], b[j]) = (1, 0)$.

The letter X denotes a random variable that assumes values on $\mathrm{Z}_{2^{32}}$. The random variable X is said to be uniformly distributed on $A \subseteq \mathrm{Z}_{2^{32}}$ iff $\mathrm{Prob}(X = a) = \mathrm{Prob}(X \in A)/|A|$. We write $X \sim \nu$ if X is uniformly distributed on $\mathrm{Z}_{2^{32}}$.

Convention. In the following $F_+, F_-, F_0, F_1 \subseteq \{1, \ldots, 32\}$ and $F_{32,\neq} \subseteq \{32\}$ denote disjoint subsets. Further, $F_= := \{1, \ldots, 32\} \setminus (F_+ \cup F_- \cup F_0 \cup F_1 \cup F_{32,\neq})$.

Apparently,

$$S_+ := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid (m[j], m'[j]) = (0, 1) \text{ for all } j \in F_+\} \quad (4)$$

$$S_- := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid (m[j], m'[j]) = (1, 0) \text{ for all } j \in F_-\} \quad (5)$$

$$S_0 := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid (m[j], m'[j]) = (0, 0) \text{ for all } j \in F_0\} \quad (6)$$

$$S_1 := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid (m[j], m'[j]) = (1, 1) \text{ for all } j \in F_1\} \quad (7)$$

$$S_{32,\neq} := \begin{cases} \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid m[32] = m'[32]\} & \text{if } F_{32,\neq} = \{32\} \\ \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} & \text{if } F_{32,\neq} = \{\} \end{cases} \quad (8)$$

$$S_= := \{(m, m') \in \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \mid m[j] = m'[j] \text{ for all } j \in F_=\} \quad (9)$$

define 1-1-correspondences between the index sets $F_+, \ldots, F_=$ and the subsets $S_+, \ldots, S_= \subseteq \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}}$. In the notation of [WY] the index sets $F_+, F_-, F_0, F_1, F_{32,\neq}, F_=$ express bit conditions. Note that $(a, b) \in S(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=)$ $:= S_+ \cap S_- \cap S_0 \cap S_1 \cap S_{32,\neq} \cap S_=$ iff (a, b) meets the bit conditions implied by $F_+, F_-, F_0, F_1, F_{32,\neq}, F_=$.

Example 1. The bit conditions $\Delta_B(a, b) = [30, -26]$ and a[4]=b[4]=1 correspond to $F_+ = \{30\}, F_- = \{26\}, F_0 = \{\}, F_1 = \{4\}, F_{32,\neq} = \{\}, F_= = \{1, \ldots, 32\} \setminus \{4, 26, 30\}$.

Convention. In the following $G_0, G_1 \subseteq \{1, \ldots, 32\}$ denote disjoint subsets.

Similarly as above, for $q \in \{0, 1\}$

$$T_q := \{m \in \mathrm{Z}_{2^{32}} \mid m[j] = q \text{ for all } j \in G_q\} \quad (10)$$

implies a 1-1-correspondence between the index set G_q and $T_q \subseteq \mathrm{Z}_{2^{32}}$. Further, $T(G_0, G_1) := T_0 \cap T_1$.

Lemma 1. *(i) The mappings*

$$(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=) \mapsto \quad (11)$$
$$S(F_+, F_-, F_0, F_1, F_{32,\neq}, F_=) = S_+ \cap S_- \cap S_0 \cap S_1 \cap S_{32,\neq} \cap S_= \text{ and}$$
$$(G_0, G_1) \mapsto T(G_0, G_1) = T_0 \cap T_1. \quad (12)$$

are injective.
(ii) For any $(a,b) \in S(F_+, F_-, F_0, F_1, F_{32,=}, F_=)$

$$\Delta(a,b) \equiv b - a \equiv \sum_{j \in F_{32,\neq}} 2^{31} + \sum_{j \in F_+} 2^{j-1} - \sum_{j \in F_-} 2^{j-1} (\mathrm{mod}\, 2^{32}) \tag{13}$$

In particular, the function $\Delta(\cdot,\cdot)$ *is constant on the set* $S_+ \cap S_- \cap S_{32,=} \supseteq S(F_+, F_-, F_0, F_1, F_{32,=}, F_=)$*, assuming a value* $\Delta(F_+, F_-, F_{32,=})$*.*
(iii) Let $\mathrm{pr}_1 \colon \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \to \mathrm{Z}_{2^{32}}$ *denote the projection onto the first component. Then*

$$\mathrm{pr}_1(S(F_+, F_-, F_0, F_1, F_{32,=}, F_=)) = T(F_+ \cup F_0, F_- \cup F_1). \tag{14}$$

(iv) The mapping

$$(F_+, F_-, F_0, F_1, F_{32,=}, F_=) \mapsto (\Delta(F_+, F_-, F_{32,=}), (F_+ \cup F_0, F_- \cup F_1)) \tag{15}$$

is injective.
(v)

$$\begin{aligned} &(a,b) \in S(F_+, F_-, F_0, F_1, F_{32,=}, F_=) \iff \\ &\left((b - a \equiv \Delta(F_+, F_-, F_{32,=}) (\mathrm{mod}\, 2^{32})), \, (a \in T(F_+ \cup F_0, F_- \cup F_1)) \right). \end{aligned} \tag{16}$$

(vi) Let X, X' *denote random variables that assume values on* $\mathrm{Z}_{2^{32}}$*, and let* $S := S(F_+, F_-, F_0, F_1, F_{32,=}, F_=)$*,* $\Delta := \Delta(F_+, F_-, F_{32,=})$ *and* $T := T(F_+ \cup F_0, F_- \cup F_1)$ *for the moment. Then*

$$\mathrm{Prob}((X, X') \in S) = \mathrm{Prob}(X' - X \equiv \Delta (\mathrm{mod}\ 2^{32}) \mid X \in T) \cdot \mathrm{Prob}(X \in T). \tag{17}$$

Proof. Assertions (i) and (ii) are obvious since $+2^{31} \equiv -2^{31} (\mathrm{mod}\, 2^{32})$. Assertion (iii) is true since $j \in F_+ \cup F_0$ implies $m[j] = 0$ for all $(m, m') \in S_+ \cap S_0 \supseteq S(F_+, F_-, F_0, F_1, F_{32,=}, F_=)$ etc. The "⇒" direction of (v) is obvious from (ii) and (iii). For the other direction observe that because of (iii) for every $a \in T(F_+ \cup F_0, F_- \cup F_1)$ there is at least one b with $(a,b) \in S(F_+, F_-, F_0, F_1, F_{32,=}, F_=)$, but (ii) implies that there is only one, namely, with $b - a \equiv \Delta(F_+, F_-, F_{32,=}) (\mathrm{mod}\, 2^{32})$. (iv) follows from (v) and also (vi) is an immediate consequence of (v) and the definition of conditional probabilities.

The following theorems will turn out to be very useful since under assumptions that can be viewed as fulfilled in the intended applications (cf. Chapter 4) they allow to move the calculation of transition probabilities (cf. (3) and Sec. 4) from the product space $\mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}}$ to $\mathrm{Z}_{2^{32}}$. Note that for the special case addressed in Theorem 2(ii) the exact probability equals $2^{-\#\,\text{bit conditions}}$.

Definition 1. *We introduce the following abbreviations which will be used in the remainder of this section:* $S_{(i)} := S(F_{(i)+}, F_{(i)-}, F_{(i)0}, F_{(i)1}, F_{(i)32,=}, F_{(i)=})$*,* $\Delta_{(i)} := \Delta(F_{(i)+}, F_{(i)-}, F_{(i)32,=})$ *and* $T_{(i)} := T(F_{(i)+} \cup F_{(i)0}, F_{(i)-} \cup F_{(i)1})$*. The index* i *ranges from* 1 *to* 3*.*

Theorem 1. *Let* X, X', Y, Y' *denote random variables that assume values in* $\mathrm{Z}_{2^{32}}$*, where* (X, X') *and* (Y, Y') *are independent. Assume further that the random vectors* (X, X') *and* (Y, Y') *are uniformly distributed on the sets* $S_{(1)}$ *and* $S_{(2)}$*, resp. and that* X *and* Y *are uniformly distributed on* $T_{(1)}$ *and* $T_{(2)}$*, respectively.*
(i) The conditional probability

$$\mathrm{Prob}\left([(X, X') + (Y, Y')] (\mathrm{mod}\, 2^{32}) \in S_{(3)} \mid (X, X') \in S_{(1)}, (Y, Y') \in S_{(2)} \right) \tag{18}$$

simplifies to

$$\begin{cases} \mathrm{Prob}\left([X + Y] (\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X \in T_{(1)}, Y \in T_{(2)} \right) & \text{if } \Delta_{(3)} \equiv \Delta_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32}) \\ 0 & \text{else} \end{cases} \tag{19}$$

(ii) If $T_{(1)} = \mathrm{Z}_{2^{32}}$ *or* $T_{(2)} = \mathrm{Z}_{2^{32}}$ *then*

$$\text{Prob } [X+Y] \,(\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X \in T_{(1)}, Y \in T_{(2)} \;\; = 2^{-|F_{(3)+} \cup F_{(3)0} \cup F_{(3)-} \cup F_{(3)1}|} \tag{20}$$

Proof. We assume $\mathrm{Prob}((X,X') \in S_{(1)}, (Y,Y') \in S_{(2)}) > 0$ since otherwise the conditional probability (18) may be defined arbitrarily. Obviously, this conditional probability is zero if $\Delta_{(3)} \not\equiv \Delta_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32})$. We assume $\Delta_{(3)} \equiv \Delta_{(1)} + \Delta_{(2)} (\mathrm{mod}\, 2^{32})$ in the remainder of this proof. Due to Lemma 1(v) the term (18) equals

$$\sum_{(x,x') \in S_{(1)}, (y,y') \in S_{(2)}} \text{Prob } [X+Y] \,(\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid (X,X') = (x,x'), (Y,Y') = (y,y') \;\; \times$$
$$\times \frac{\mathrm{Prob}((X,X') = (x,x'), (Y,Y') = (y,y'))}{\mathrm{Prob}((X,X') \in S_{(1)}, (Y,Y') \in S_{(2)})}.$$

To any $(x,y) \in T_{(1)} \times T_{(2)}$ there exists a unique quadruple $(x,x',y,y') \in S_{(1)} \times S_{(2)}$. Since (X,X') and (Y,Y') are independent the above term simplifies to

$$\begin{aligned}
&\sum_{x \in T_{(1)}, y \in T_{(2)}} \text{Prob } [X+Y] \,(\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X = x, Y = y \;\; \times \\
&\qquad \times \frac{\mathrm{Prob}((X,X') = (x,x'))}{\mathrm{Prob}((X,X') \in S_{(1)})} \cdot \frac{\mathrm{Prob}((Y,Y') = (y,y'))}{\mathrm{Prob}((Y,Y') \in S_{(2)})} \\
= &\sum_{x \in T_{(1)}, y \in T_{(2)}} \text{Prob } [X+Y] \,(\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X = x, Y = y \;\; \times \\
&\qquad \times \frac{\mathrm{Prob}(X = x)}{\mathrm{Prob}(X \in T_{(1)})} \cdot \frac{\mathrm{Prob}(Y = y)}{\mathrm{Prob}(Y \in T_{(2)})} \\
= &\,\text{Prob } [X+Y] \,(\mathrm{mod}\, 2^{32}) \in T_{(3)} \mid X \in T_{(1)}, Y \in T_{(2)}
\end{aligned}$$

The first equality follows from the condition that (X,X') and X are uniformly distributed on $S_{(1)}$ and $T_{(1)}$, resp. that (Y,Y') and Y are uniformly distributed on $S_{(2)}$ and $T_{(2)}$, repectively. In (ii) the random variable $[X+Y] (\mathrm{mod}\, 2^{32})$ is uniformly distributed on $\mathrm{Z}_{2^{32}}$ regardless of $T_{(2)}$ (first assumption), resp., regardless of $T_{(1)}$ (second assumption). Consequently, in both cases the conditional probability equals $|T_{(3)}|/|\mathrm{Z}_{2^{32}}|$.

Definition 2. *For* $1 \le sh < 32$ *the term* $w^{<<<sh}$ *denotes the cyclic shift of the word* w *by* sh *positions to the left. Similarly* $(w,w')^{<<<sh}$ *stands for* $(w^{<<<sh}, w'^{<<<sh})$.

Remark 2. (i) Clearly, if $F_{32,=} = \{\}$ the image $S(F_+, F_-, F_0, F_1, F_{32,=}, F_=)^{<<<sh}$ equals $S(F_+^{<<<sh}, F_-^{<<<sh}, F_0^{<<<sh}, F_1^{<<<sh}, \{\}, F_=^{<<<sh})$. We have $j \in F_*$ iff $j+sh$, resp. $j+sh-32 \in F_*^{<<<sh}$. This condition is not too restrictive since the set $S(F_+, F_-, F_0, F_1, \{32\}, F_=)$ equals the disjoint union $S(F_+ \cup \{32\}, F_-, F_0, F_1, \{\}, F_=) \cup S(F_+, F_- \cup \{32\}, F_0, F_1, \{\}, F_=)$.
(ii) Note that $\Delta(F_+, F_-, \{\}) = \Delta(F'_+, F'_-, \{\})$ does not necessarily imply $\Delta(F_+^{<<<sh}, F_-^{<<<sh}, \{\}) = \Delta(F'^{<<<sh}_+, F'^{<<<sh}_-, \{\})$.
Counterexample: $F_+ = \{20\}, F'_+ = \{21\}, F'_- = \{20\}, sh = 12$. Then $\Delta(\{20\}, \{\}, \{\})$ $= 2^{19} = \Delta(\{21\}, \{20\}, \{\})$ but $\Delta(\{20\}^{<<<12}, \{\}, \{\}) = \Delta(\{32\}, \{\}, \{\}) = 2^{31}$ whereas $\Delta(\{21\}^{<<<12}, \{20\}^{<<<12}, \{\}) = \Delta(\{1\}, \{32\}, \{\}) \equiv -2^{31} + 1 \equiv 2^{31} + 1 (\mathrm{mod}\, 2^{32})$.

Theorem 2. *Let* X, X', Y, Y' *denote random variables that assume values in* $\mathrm{Z}_{2^{32}}$*, where* (X,X') *and* (Y,Y') *are independent, and let* $F_{(1)32,=} = \{\}$*. Assume*

further that (Y, Y') and Y are uniformly distributed on $S_{(2)}$ and $T_{(2)}$, respectively.
(i) Let (X, X') and X be uniformly distributed on $S_{(1)}$ and $T_{(1)}$, respectively. Then the conditional probability

$$\text{Prob}\left[(X,X')^{<<<sh} + (Y,Y')\right] (\text{mod } 2^{32}) \in S_{(3)} \mid (X,X') \in S_{(1)}, (Y,Y') \in S_{(2)} \tag{21}$$

simplifies to

$$\begin{array}{ll} \text{Prob}\left[X^{<<<sh} + Y\right] (\text{mod } 2^{32}) \in T_{(3)} \mid X \in T_{(1)}, Y \in T_{(2)} & \text{if } \Delta_{(3)} \equiv \widetilde{\Delta}_{(1)} + \Delta_{(2)} (\text{mod } 2^{32}) \\ 0 & \text{else} \end{array} \tag{22}$$

Here $\widetilde{\Delta}_{(1)} := \Delta(F_{(1)+}^{<<<sh}, F_{(1)-}^{<<<sh}, \{\})$.
(ii) Additionally to (i) we assume $T_{(1)} = \mathrm{Z}_{2^{32}}$ or $T_{(2)} = \mathrm{Z}_{2^{32}}$. Then

$$\text{Prob}\left[X^{<<<sh} + Y\right] (\text{mod } 2^{32}) \in T_{(3)} \mid X \in T_{(1)}, Y \in T_{(2)} \;= 2^{-|F_{(3)+} \cup F_{(3)0} \cup F_{(3)-} \cup F_{(3)1}|} \tag{23}$$

(iii) Assume that (X, X') and X are uniformly distributed on $\{(x, x + \Delta_{[1]} (\text{mod } 2^{32})) \mid x \in \mathrm{Z}_{2^{32}}\}$ and on $\mathrm{Z}_{2^{32}}$, respectively. Setting $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh) := \{u \in \mathrm{Z}_{2^{32}} \mid \Delta((u, u + \Delta_{[1]} (\text{mod } 2^{32}))^{<<<sh}) + \Delta_{(2)} \equiv \Delta_{(3)} (\text{mod } 2^{32})\}$ we obtain

$$\begin{aligned} &\text{Prob}\left[(X,X')^{<<<sh} + (Y,Y')\right] (\text{mod } 2^{32}) \in S_{(3)} \mid \Delta(X,X') = \Delta_{[1]}, (Y,Y') \in S_{(2)} \\ &= \text{Prob}\left[X^{<<<sh} + Y\right] (\text{mod } 2^{32}) \in T_{(3)} \mid X \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh), Y \in T_{(2)} \;\times \\ &\quad \times \text{Prob}(X \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)). \end{aligned} \tag{24}$$

(iv) If $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh) = \mathrm{Z}_{2^{32}}$ then (24) equals $2^{-|F_{(3)+} \cup F_{(3)0} \cup F_{(3)-} \cup F_{(3)1}|}$.
(v) If $T_{(2)} = \mathrm{Z}_{2^{32}}$ then (24) equals
$2^{-|F_{(3)+} \cup F_{(3)0} \cup F_{(3)-} \cup F_{(3)1}|} \cdot \text{Prob}(X \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh))$.
(vi) Assertion (iv) does also hold if we drop the assumption that (Y,Y') and Y are uniformly distributed on $S_{(2)}$ and $T_{(2)}$, resp. Moreover, $(Z := X^{<<<sh} + Y (\text{mod} 2^{32}), Z := X'^{<<<sh} + Y' (\text{mod} 2^{32}))$ and Z are uniformly distributed on $S_{(3)}$ and $T_{(3)}$, respectively.

Proof. We point out that $(X, X') \in S_{(1)}$ iff $(X, X')^{<<<sh} \in {S_{(1)}}^{<<<sh}$ since $F_{32,=} = \{\}$. Since (X, X') and X are uniformly distributed on $S_{(1)}$ and $T_{(1)}$, resp., $(\widetilde{X}, \widetilde{X}') := (X, X')^{<<<sh}$ and $\widetilde{X}$ are uniformly distributed on $\widetilde{S}_{(1)} := {S_{(1)}}^{<<<sh}$ and $\widetilde{T}_{(1)} := {T_{(1)}}^{<<<sh}$, respectively. Hence (i) follows immediately from Theorem 1(i). To prove (iii) we first note that the set $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$ is well-defined. As in the proof of Theorem 1(i) we may assume that $\text{Prob}((Y, Y') \in S_{(2)}) > 0$ in the remainder. Due to Lemma 1(v) the left-hand side in (24) equals

$$\begin{aligned} &\text{Prob}\left[X^{<<<sh} + Y\right] (\text{mod } 2^{32}) \in T_{(3)} \mid \Delta(X,X') = \Delta_{[1]}, X \in M(\cdots), (Y,Y') \in S_{(2)} \;\times \\ &\qquad \times \frac{\text{Prob}(\Delta(X,X') = \Delta_{[1]}, X \in M(\cdots))}{\text{Prob}(\Delta(X,X') = \Delta_{[1]})} \end{aligned}$$

Due to the uniformity assumptions in (iii) the second factor equals $\text{Prob}(X \in M(\cdots))$. Consequently, the above probability equals

$$\begin{aligned} &\sum_{x \in M(\cdots), y \in T_{(2)}} \text{Prob}\left[X^{<<<sh} + Y\right] (\text{mod } 2^{32}) \in T_{(3)} \mid X = x, Y = y \;\times \\ &\quad \times \frac{\text{Prob}((X,X') = (x, x + \Delta_{[1]}))}{\text{Prob}((X,X') : X \in M(\cdots), \Delta(X,X') = \Delta_{[1]})} \cdot \frac{\text{Prob}((Y,Y') = (y, y + \Delta_{(2)}))}{\text{Prob}((Y,Y') \in S_{(2)})} \times \\ &\quad \times \text{Prob}(X \in M(\cdots)) \\ = &\sum_{x \in M(\cdots), y \in T_{(2)}} \text{Prob}\left[X^{<<<sh} + Y\right] (\text{mod } 2^{32}) \in T_{(3)} \mid X = x, Y = y \;\times \\ &\qquad \times \frac{\text{Prob}(X = x)}{\text{Prob}(X \in M(\cdots))} \cdot \frac{\text{Prob}(Y = y)}{\text{Prob}(Y \in T_{(2)})} \cdot \text{Prob}(X \in M(\cdots)) \end{aligned}$$

which finishes the proof of (iii). The assertions (ii), (iv) and (v) are obvious (cf. the proof of Theorem 1). To prove (iv) for arbitrary distribution of (Y, Y') in the first line of (22) $Y \in T_{(2)}$ has to be replaced by $(Y \mid (Y, Y') \in S_{(2)}) \in T_{(2)}$. Since X is uniformly distributed on $M(...) := Z_{2^{32}}$ the random variable Z is uniformly distributed on $T_{(3)}$. Moreover, due to the construction of $S_{(3)}$ hence also (Z, Z') is uniformly distributed on $S_{(3)}$.

In the remainder of this section we derive a characterization of the set $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$ which is more suitable for concrete computations.

Definition 3. *For $a \in \mathrm{Z}$ and $n \in N$ we set a div n to be $\lfloor a/n \rfloor$ where $\lfloor r \rfloor$ denotes the largest integer that is $\leq r$.*
For $a \in \mathrm{Z}$ the term $a(\mathrm{mod}\, M)$ stands for the representative of $a + \mathrm{Z/MZ}$ in $\mathrm{Z_M}$, i.e. for that element in $\mathrm{Z_M}$ that is congruent to the integer a modulo M.

Lemma 2. *Within this lemma let $x' \in \mathrm{Z}_{2^{32}}$ with $x' \equiv x + \Delta (\mathrm{mod}\, 2^{32})$ for fixed $\Delta \in \mathrm{Z}$. Let further $x = x_1 \cdot 2^{32-sh} + x_0$ and $x' = x'_1 \cdot 2^{32-sh} + x'_0$ with $0 \leq x_0, x'_0 < 2^{32-sh}$ and $0 \leq x_1, x'_1 < 2^{sh}$. Further, we decompose $\Delta = \Delta_1 \cdot 2^{32-sh} + \Delta_0$ where Δ_0 and Δ_1 may assume arbitrary integer values. This implies:*
(i) For $a \in \mathrm{Z}$ and $n \in N$ we have $a(\mathrm{mod} n) = a - (a \text{ div } n)n$.
(ii) For $a \in \mathrm{Z}$ and $n, m \in \mathbb{N}$ we have $(a(\mathrm{mod} n))m = am(\mathrm{mod}\, nm)$.
(iii) $x^{<<<sh} = x_0 \cdot 2^{sh} + x_1$.
(iv) Let $\mathrm{ca}(a_1, \ldots, a_k) := (a_1 + \cdots + a_k)$ div 2^{32-sh} ('carry'). Then $x' = ((x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0))(\mathrm{mod}\ 2^{sh})) \cdot 2^{32-sh} + (x_0 + \Delta_0)(\mathrm{mod}\ 2^{32-sh})$ (integer equation!)
(v) $x'^{<<<sh} \equiv\ x_0 + \Delta_0 -\ (x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0))\ \text{div}\ 2^{sh}\ \ \cdot 2^{sh} +$ $x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0)\ (\mathrm{mod}\, 2^{32})$.
(vi) Let $k \cdot 2^{32-sh} \leq \Delta_0 < (k+1)2^{32-sh}$ for a particular $k \in Z$. Then $\mathrm{ca}(x_0, \Delta_0) \in \{k, k+1\}$.

Proof. Assertion (i) follows from its definition, and $(a(\ \mathrm{mod}\ n))m = (a - [a \text{ div } n]n)m$ $= am - [a \text{ div } n]nm = am(\mathrm{mod}\, nm)$. Assertions (iii), (iv) and (vi) are obvious. From (iv) we immediately obtain

$$x'^{<<<sh} \equiv\ (x_0 + \Delta_0)\,(\mathrm{mod}\, 2^{32-sh})\ \cdot 2^{sh} + (x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0))\,(\mathrm{mod}\, 2^{sh}).$$

Applying (i) to the right-hand summand and (ii) to the left-hand summand proves (v).

Theorem 3. *(Continuation of Theorem 2) Assume that $\Delta_{(3)} - \Delta_{(2)} \equiv \left(\widetilde{\Delta}_0 * 2^{sh} + \widetilde{\Delta}_1\right)(\ \mathrm{mod}\ 2^{32})$ and $\Delta_{[1]} \equiv\ \Delta_{1[1]} * 2^{32-sh} + \Delta_{0[1]}\ \ (\ \mathrm{mod}\ 2^{32})$ for suitable (but not necessarily nonnegative) numbers $\widetilde{\Delta}_0, \widetilde{\Delta}_1, \Delta_{0[1]}, \Delta_{1[1]}$. Then*

$$\begin{aligned} x = x_1 \cdot 2^{32-sh} + x_0 \in M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh) \quad \textit{iff} \quad (25) \\ \widetilde{\Delta}_0 * 2^{sh} + \widetilde{\Delta}_1 \equiv\ \Delta_0 - [x_1 + \Delta_1 + \mathrm{ca}(x_0, \Delta_0)]\ \text{div}\ 2^{sh}\ \ \cdot 2^{sh} + \\ [\Delta_1 + \mathrm{ca}(x_0, \Delta_0)]\ (\mathrm{mod}\, 2^{32}) \end{aligned}$$

Proof. Theorem 3 follows immediately from the definition of $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$ and Lemma 2(iii), (v).

Theorem 3 provides the announced alternative characterization of the set $M(\Delta_{[1]}, \Delta_{(2)}, \Delta_{(3)}, sh)$. In fact, it provides sufficient and necessary conditions on x_0 and x_1. By (25) and Lemma 2(vi) we have

$$\mathrm{ca}(x_0, \Delta_0) \equiv \widetilde{\Delta}_1 - \Delta_1 (\ \mathrm{mod}\ 2^{sh}), \quad \mathrm{ca}(x_0, \Delta_0) \in \{\Delta_0 \text{ div } 2^{32-sh}, \Delta_0 \text{ div } 2^{32-sh} + 1\} \tag{26}$$

which determines $\mathrm{ca}(x_0, \Delta_0)$. This gives an inequality for x_0. In a second step one derives a characterization for the 'upper' part x_1. We point out that the term $\mathrm{ca}(x_0, \Delta)$ compensates the 'non-uniqueness' of Δ_0, Δ_1.

Remark 3. Theorem 2 and Theorem 3 will be very useful in the next section. We point out that both theorems can be extended to handle bit conditions that affect (Y, Y') and $(Z := X^{<<<sh} + Y (\mathrm{mod} 2^{32}), Z' := X^{<<<sh} + Y (\mathrm{mod}\ 2^{32}))$ simultaneously. For instance, the (additional) condition $Y[3] = Y'[3] = Z[3] = Z'[3]$ can be decomposed into two disjoint cases, namely into $Y[3] = Y'[3] = 0 = Z[3] = Z'[3]$, and $Y[3] = Y'[3] = 1 = Z[3] = Z'[3]$, respectively. Both cases can be expressed in the form $S(F_+, F_-, F_0 \cup \{3\}, F_1, F_{32,=}, F_=)$ and $S(F_+, F_-, F_0, F_1 \cup \{3\}, F_{32,=}, F_=)$ with suitable subsets $F_+, F_-, F_0, F_1, F_{32,=}, F_=$.

4 Example: Concrete Collision Paths in MD5

In this section we illustrate the use and the usefulness of the observations from Section 3 by three MD5 near-collision paths. We do not claim that the most probable of these paths is optimal.

After the initialization of four registers by the $IV = (IV_0, IV_1, IV_2, IV_3)$

$$r_{-3} := IV_0, \quad r_{-2} := IV_3, \quad r_{-1} := IV_2, \quad r_0 := IV_1 \tag{27}$$

the MD5 algorithm processes 64 steps. In Step i the MD5 step function has the form

$$\text{(Step i)} \quad r_i \equiv r_{i-1} + (\Phi_i(r_{i-1}, r_{i-2}, r_{i-3}) + r_{i-4} + \widetilde{m}_i + const_i)^{<<<sh(i)} (\mathrm{mod}\ 2^{32}) \tag{28}$$

where $\Phi_i \colon \mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}} \to \mathrm{Z}_{2^{32}}$ is a bit-oriented, step-dependent function. Also the constant $const_i$ and the number of shift positions $sh(i)$ depend on the particular step. Finally, the four registers are updated by

$$\text{(post addition)} \quad r_i^p \equiv r_i + r_{i-64} (\mathrm{mod}\, 2^{32}) \quad i \in \{61, 62, 63, 64\} \tag{29}$$

The known MD5 attacks are two-block attacks (see, e.g. [WY,Kli2,YaSh]), i.e. after block 1 the pairs $(r_{61}^p, r'^p_{61}), (r_{62}^p, r'^p_{62}), (r_{63}^p, r'^p_{63}), (r_{64}^p, r'^p_{64})$ shall meet specified bit conditions that shall 'prepare' a collision after the compression of the second block. E.g. in [WY,Kli2,YaSh,Th] conditions on the message blocks, the register bits and intermediate values are formulated that shall ensure this goal. The conditions for the first 20 steps can be guaranteed by a random (but sophisticated) choice of the message blocks, the so-called message modification ([WY,Kli2] etc.). Our goal is to compute the probability for concrete (near-)collision paths from Step 21 to Step 64 (including the post additions).

4.1 Step Transition Probabilities

Since at least large parts of the message blocks $m_1, \ldots, m_{16}$ are chosen randomly we interpret the register values $r_1, \ldots, r_{64}^p$ and the extended message blocks $\widetilde{m}_1, \ldots, \widetilde{m}_{64}$ as realizations of random variables $R_1, \ldots, R_{64}^p$ and $\widetilde{M}_1, \ldots, \widetilde{M}_{64}$ with specific distributions. In the notion of random variables (28) reads

$$\text{(Step i)} \quad R_i \equiv R_{i-1} + (\Phi_i(R_{i-1}, R_{i-2}, R_{i-3}) + R_{i-4} + \widetilde{M}_i + const_i)^{<<<sh(i)} (\mathrm{mod}\ 2^{32}) \tag{30}$$

If we replace $const_i$ by an independent random variable that is uniformly distributed on $\mathrm{Z}_{2^{32}}$ the terms R_{i-1} and $(\cdots)^{<<<sh(i)}$ were independent and the latter uniformly distributed on $\mathrm{Z}_{2^{32}}$. Although $const_i$ assumes a constant value the following stochastic model is yet reasonable.

Stochastic Model. We assume that the random variables $(R_{i-1}, R'_{i-1}), (R_{i-2}, R'_{i-2}),$ ... lie on a particular near-collision path, i.e. that they meet specific suifficient conditions. Let $X_i := \Phi_i(R_{i-1}, R_{i-2}, R_{i-3}) + R_{i-4} + \widetilde{M}_i + const_i \pmod{2^{32}}$

and $X'_i := \ \Phi_i(R'_{i-1}, R'_{i-2}, R'_{i-3}) + R'_{i-4} + \widetilde{M}'_i + const_i \ \pmod{2^{32}}$. We assume (a) that the pairs (X_i, X'_i) and (R_{i-1}, R'_{i-1}) are independent, (b) that X_i is uniformly distributed, and (c) that $(X_i, X'_i) \mid \{(x, x + \Delta_i (\mathrm{mod}\, 2^{32})) \mid x \in \mathrm{Z}_{2^{32}}\}$ is uniformly distributed.

Remark: If $M_{(i-1)} = Z_{2^{32}}$ then Theorem 2(vi) implies that (R_{i-1}, R'_{i-1}) and R_{i-1} are uniformly distributed on $S_{(i-1)}$ and $T_{(i-1)}$, respectively, with $M_{(i)} := M(\Delta_{[i]}, \Delta_{(i-1)}, \Delta_{(i)}, sh(i))$. We note that for the paths specified in Table 2 we have $M_{(i)} = Z_{2^{32}}$ only for $i = 23, 35, 62$.

Justification of the Stochastic Model. (i) We add R_{i-4} and $\widetilde{M}_i$ which have no 'obvious' (at least no linear) dependencies with R_{i-1} and $\Phi(R_{i-1}, R_{i-2}, R_{i-3})$ (merging the last three register values in a non-linear manner), while the modular addition is a $\mathrm{Z}_{2^{32}}$-linear operation on $\mathrm{Z}_{2^{32}}$. As the same argumentation holds for the related message M' instead of M this justifies Condition (a).
(ii) Even under weak heuristic assumptions the modular sum of three random variables is very close to the uniform distribution. distributed, justifying (b).
(iii) Condition (c) is induced by the fact that the particular register values 'spread' rapidly for different messages. For 'purely' random input M and M' (without message modification) and neglecting any bit condition up to step $i-1$ we would assume that (X, X') is uniformly distributed on $\mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}}$. In our scenario, i.e. where we focus on a small subset of near-collision paths that fulfil a sequence of bit conditions we assume a weaker assumption, namely that the conditional random vectors $(X, X') \mid \{(x, x + \Delta_i (\mathrm{mod}\, 2^{32})) \mid x \in \mathrm{Z}_{2^{32}}\}$ is uniformly distributed.

In this subsection we consider about the step transition probabilities, i.e. the first type of conditional probabilities in (3). Therefore, we apply Theorem 2 with $\Delta_{[1]} := \Delta_i$, $S_{(2)} = S_{(i-1)}$, $S_{(3)} = S_{(i)}$, $(X, X') := (X_i, X'_i)$ and $(Y, Y') := (R_{i-1}, R'_{i-1})$ where we assume (cf. the stochastic model) that also the predecessors $(R_{i-2}, R'_{i-2}), \ldots$ meet the path requirements quantified by sets $S_{(i-2)}, \ldots$.

The bit conditions from Step 21 to Step 64 are listed in Table 2. For $i > 21$, resp. $i \geq 61$, the sets $S_{(i)}$, resp. $S_{(i)p}$, can be expressed in the form $S_{(i)} := S(F_{(i)+}, F_{(i)-}, F_{(i)0}, F_{(i)1}, F_{(i)32,=}, F_{(i)=})$, resp. $S_{(i),p} := S(F_{(i)+,p}, F_{(i)-,p}, F_{(i)0,p}, F_{(i)1,p}, F_{(i)32,=,p}, F_{(i)=,p})$ (see Sect. 3). In Step 21 we have a specific equality condition (namely, $r_{21}[18] = r'_{21}[18] = r_{20}[18] = r'_{20}[18]$) which yet can also be handled by applying Theorem 3 (see Remark 3).

Example 2. (Step 48)
As in Theorem 3 we decompose $x_{(48)} = x_1 \cdot 2^{32-sh(48)} + x_0$ with $0 \leq x_0 < 2^{32-sh(48)}$ and $0 \leq x_1 < 2^{sh(48)}$. Elementary calculations give $X'_{(48)} - X_{(48)} \equiv 0 (\mathrm{mod}\, 2^{32})$ and $\Delta_{(48)} - \Delta_{(47)} \equiv 2^{31} + 2^{31} \equiv 0 (\mathrm{mod}\, 2^{32})$. Using the notation from Theorem 3 (with $(X_{(i)}, X'_{(i)}), (R_{(i-1)}, R'_{(i-1)}), (R_{(i)}, R'_{(i)})$ corresponding to $(X, X'), (Y, Y'), (Z, Z')$) we conclude $\Delta_0 = \Delta_1 = \widetilde{\Delta}_0 = \widetilde{\Delta}_1 = 0$. In particular, $\mathrm{ca}(x_0, \Delta_0) = \mathrm{ca}(x_0, 0) = 0$ for all x_0, and the second condition of (25) simplifies to $0 \equiv -(x_1 \ \mathrm{div}\ 2^{sh(48)}) \cdot 2^{sh(48)} + 0 (\mathrm{mod}\, 2^{32})$ which obviously is fulfilled for all $0 \leq x_1 < 2^{sh(48)}$. In other words, $M(\Delta_{[1]}, \Delta_{(47)}, \Delta_{(48)}, sh(48)) = M(0, 2^{31}, 2^{31}, 23)) = \mathrm{Z}_{2^{32}}$. For path 1 we have $S_{(48)} = (\{32\}, \{\}, \{\}, \{\}, \{\}, \{1, \ldots, 31\})$ and $S_{(47)} = (\{\}, \{32\}, \{\}, \{\}, \{\}, \{1, \ldots, 31\})$ and by Theorem 2 (vi) we finally obtain the conditional probability (transition probability) $\mathrm{Prob}((R_{48}, R'_{48}) \in S_{(48)} \mid (R_{47}, R'_{47}) \in S_{(47)}, \Delta(X_{48}, X'_{48}) = \Delta_{48}) = 2^{-|F_{(48)+}|} = 2^{-1}$. The analogous calculation for path 2 and 3 gives the same transition probablity 2^{-1}.

In Step 48 the exact conditional probability equals the value that follows from 'condition counting'. This is also true for each Step $i \in \{21, \ldots, 63\} \setminus \{23, 35, 62\}$.

We point out that the conditions in Steps 36 to 45 are fulfilled with probability 1 (no 'real' bit conditions), which is obvious, resp. can be verified with formula (24). (Note that $\Delta(X_i, X'_{i-1}) = 0$ and $\Delta_{(i-1)} = \Delta_{(i)} = 2^{31}$ in these steps.) The next example treats the exceptional set $\{23, 35, 62\}$. In these steps the sums within the bracket (cf. (28)) must satisfy additional conditions. In other words: The list of conditions in [WY] is not sufficient. We point out that [HPR] and [Th] mentioned these additional bit conditions.

The situation in Step 64 is different from that in the other steps since r_{64} has no impact on any other register. Hence only the modulo 2^{32}-difference $(r'_{64} - r_{64} (\mathrm{mod}\, 2^{32})$ is relevant (cf. Example 4(iv)).

For our paths $M_{(i)} = \mathrm{Z}_{2^{32}}$ for $i \notin \{23, 35, 62\}$, and since $\Delta(X_i, X'_i) = 0$ the modulo 2^{32} condition in Step 64 is fulfilled with probability 1. By Theorem 2 the path transition probability from Step 21 to 64 (before post addition) reads

$$\prod_{i\in\{21,\ldots,63\}\setminus\{23,35,62\}} 2^{-|F_{(i)+}\cup F_{(i)0}\cup F_{(i)-}\cup F_{(i)1}|} \times \tag{31}$$
$$\prod_{i\in\{23,35,62\}} \mathrm{Prob}(X_i^{<<<sh(i)} + U_{i-1} (\mathrm{mod} 2^{32}) \in T_{(i)} \mid X_i \in M_{(i)}, U_{i-1} \in T_{(i-1)}) \mathrm{Prob}(X_i \in M_{(i)})$$

(to be modified in Step 21, see above). For $i = 23, 35, 62$ the random variables U_{i-1} are assumed to be independent and uniformly distributed on $T_{(i-1)}$ (cf. Theorem 2(vi), applied in Step $i-1$). The following example treats the exceptional steps 23, 35 and 62.

Example 3. (i) (Step 23): Here $sh(23) = 14$ and hence $x_{(23)} = x_1 \cdot 2^{18} + x_0$ with $0 \le x_0 < 2^{18}$ and $0 \le x_1 < 2^{14}$. Elementary calculations give $X'_{(23)} - X_{(23)} \equiv 2^{31} + 2^{31} + 2^{17} \equiv 2^{17} (\mathrm{mod}\, 2^{32})$ and $\Delta_{(23)} - \Delta_{(22)} \equiv 0 - 2^{31} \equiv 2^{31} (\mathrm{mod}\, 2^{32})$. We conclude $\Delta_0 = 2^{17}, \Delta_1 = 0, \widetilde{\Delta}_0 = 2^{17}, \widetilde{\Delta}_1 = 0$. From (26) we obtain the condition $\mathrm{ca}(x_0, \Delta_0) = \mathrm{ca}(x_0, 2^{17}) = 0$, or equivalently, $0 \le x_0 < 2^{17}$. Substituting into the second condition of (25) we obtain $2^{31} \equiv (2^{17} - (x_1 + 0 + 0) \mathrm{\ div\ } 2^{14})) \cdot 2^{14} + 0 \equiv 2^{31} + 0 (\mathrm{mod}\, 2^{32})$ for all x_1. In other words, $M_{(23)} := M(\Delta_{[1]}, \Delta_{(22)}, \Delta_{(23)}, sh(23)) = M(0, 2^{31}, 2^{31}, 14) = \{x \in \mathrm{Z}_{2^{32}} \mid x[18] = 0\}$. Hence $\mathrm{Prob}(X_{(23)} \in M_{(23)}) = 0.5$. Next we point out that $F_{(22)+} = \{32\}$ and $F_{(23)0} = \{32\}$. To finally apply (24) it remains to determine the conditional probability $\mathrm{Prob}((X^{<<<14} + R_{22}) (\mathrm{mod}\, 2^{32}) < 2^{31} \mid R_{22}[32] = 0, X \in M_{(23)}\}$. This probability is yet ≈ 0.5 since bit 18 of X has only marginal influence on bit 32 of the sum $X^{<<<14} + R_{22}$. Hence we use the approximation $\mathrm{Prob}((R_{23}, R'_{23}) \in S_{(23)} \mid (R_{22}, R'_{22}) \in S_{(22)}, \Delta(X_{23}, X'_{23}) = 2^{17}) \approx 2^{-1} \cdot 2^{-1} = 2^{-2}$ in the following.
(ii) (Step 35): In Step 35 we have $sh(35) = 16$ and $M_{(35)} := \{x \in \mathrm{Z}_{2^{32}} \mid x[16] = 0\}$. As in (i) we obtain $\mathrm{Prob}(X_{(35)} \in M_{(35)}) = 0.5$. Since (R_{35}, R'_{35}) and hence R_{35} need not satisfy any condition the transition probability from Step 34 to Step 35 equals 2^{-1}.
(iii) (Step 62): In Step 62 we have $sh = 10$. Similarly as for Step 23 and Step 35 we conclude that $x = x_1 \cdot 2^{22} + x_0 \in M_{(62)}$ iff $0 \le x_0 < 2^{22} - 2^{15}$. Consequently, $\mathrm{Prob}(X_{62} \in M_{(62)}) = 1 - 2^{-7}$. Since $F_{(62)+} = \{26, 32\}$ and since the influence of the least 22 significant bits of X on the sum $X + R_{61}$ is negligible $\approx (1 - 2^{-7})2^{-2}$ is a very good approximation for the transition probability from Step 61 to Step 62.

4.2 The Impact of the Postadditions

In this subsection we quantify the impact of bit conditions on the chaining values on the probabilities of hash collision paths. For Step 61 to 63 we apply Theorem 1 with $(X, X') = (R_i, R'_i), (Y, Y') = (r_{i-64}, r'_{i-64})$ and $(Z, Z') = (R_i^p, R'^p_i)$ and

for Step 64 Theorem 2(vi) with $sh = 0$. With regard to Theorem 2(vi) we may assume that (R_i, R'_i) and R_i are (only 'almost' for $i = 62$) uniformly distributed on $S_{(i)}$ and $T_{(i)}$, respectively. Note that $(Y, Y') = (r_{i-64}, r'_{i-64})$ means that the random variables (Y, Y') assume constant values, i.e. that they are contained in a singleton subset of $\mathrm{Z}_{2^{32}} \times \mathrm{Z}_{2^{32}}$. In the first message block, $r_{i-64} = r'_{i-64}$ are determined by the IV. As already pointed out in Section 3 singleton subsets can be expressed in the form $S(\cdots)$.

For the last block of a multiblock collision (= the first block in a one-block collision) we have $S_{(i),p} = S(\cdots)$ with $F_{(i)=,p} = \{1, \ldots, 32\}$ and hence $T_{(i),p} = \mathrm{Z}_{2^{32}}$. Consequently, the transition probability

$$\mathrm{Prob}((R_i^p, R'^p_i) \in S_{(i),p} \mid (R_i, R'_i) \in S_{(i)}, (R_{i-64}, R'_{i-64}) = (r_{i-64}, r'_{i-64})) = 1, \tag{32}$$

provided, of course, that $\Delta_{(i)} + \Delta_{[i-64]} \equiv \Delta_{(i),p} (\mathrm{mod}\ 2^{32})$ holds. In contrast, for near-collisions $R^p_{(i)} = R'^p_{(i)}$ for at least not all $i = N - k + 1, \ldots, N$, but the $(R^p_{(i)}, R'^p_{(i)})$ shall satisfy specific bit conditions. In that case the combination of the IV (or, the previous chaining value) and the bit conditions $\Delta_B(R_i, R'_i)$ have relevant impact on the transition probabilities. The probabilities in Example 4(i) to (iii) refer to the standard $IV = (0x\,67452301,\ 0x\,efcdab89,\ 0x\,98badcfe,\ 0x\,10325476)$, i.e. $r_{-3} = 0x\,67452301$, $r_{-2} = 0x\,10325476$, $r_{-1} = 0x\,98badcfe$, and $r_0 = 0x\,efcdab89$.

Example 4. (i) (Postaddition in Step 61): In collision path 1 (see Table 2) we have $F_{(61)+} = \{32\}, F_{(61)0} = \{27\}, F_{(61)1} = \{26\}$ and $F_{(61)+,p} = \{32\}$. Since the modulo 2^{32}-conditions are obviously fulfilled, by Theorem 1(i) (and as a consequence of our Stochastic Model) it remains to determine the probability $\mathrm{Prob}([X + r_{-3}](\mathrm{mod}\, 2^{32}) \in [0, 2^{31} - 1) \mid X[26] = 1, X[27] = X[32] = 0\}$ for uniformly distributed X. Let X_1 and X_3 denote independent random variables that are uniformly distributed on $Z_{2^{25}}$, resp. on Z_{2^4}. The last probability equals $\mathrm{Prob}([X_1 + 2^{25} + 2^{27}X_3 + r_{-3}](\mathrm{mod} 2^{32}) \in [0, 2^{31}))$. Similarly, $r_{-3} = c_1 + c_2 2^{25} + c_3 2^{27} + c_4 2^{31}$ with $c_1 \in [0, 2^{25})$, $c_2 \in [0, 4)$, $c_3 \in [0, 16)$, and $c_4 \in [0, 2)$. For the standard IV we have $c_2 = 3$, $c_3 = 12$, and $c_4 = 0$. Since $r_{-3} < 2^{31}$ the above probability simplifies to $\mathrm{Prob}((X_1 + c_1) + (X_3 + 12 + 1)2^{27} \in [0, 2^{31}))$. As $0 < c_1 + X_1 < 2^{26}$ this expression equals $\mathrm{Prob}((X_3 + 12 + 1)2^{27} \in [0, 2^{31}) = \mathrm{Prob}(X_3 + 13 < 16) = 3/16 = 0.1875$. For collision path 2 and collision path 3 from Table 2 we have $F_{(61)-} = \{32\}$ instead of $F_{(61)+} = \{32\}$. The same argumentaion as above then yields $\mathrm{Prob}((X_1 + c_1) + (X_3 + 12 + 1)2^{27} + 2^{31} \in [2^{32}, 2^{32} + 2^{31}))$ which can be reduced to $\mathrm{Prob}(X_3 + 13 \geq 16) = 13/16 = 0.81250$.
(ii) (Postadditions in Step 62): With the same techniques as in (i) we obtain the transition probability 0.789 for all three paths.
(iii) (Postadditions in Step 63): For Path 1, Path 2 and Path 3 we obtain the transition probabilities 0.034, 0.148, and 0.516, respectively.
(iv) (Postadditions in Step 64): Before postaddition only $\Delta(R_{64}, R'_{64})$ is relevant. Hence the postaddition transition probability equals 2^{-4} for all paths ('condition counting').
(v) The probabilities for the postadditions change when IVs are used that are not standard-conformant. For collision path 2, for example, for IV=$(0x\,80000000,\ 0x\,efcdab89,\ 0x\,82000000,\ 0x\,00000000)$ the joint transition probability for the postadditions in Step 61 - 63 equals 0.5. In contrast, IV=$(0x\,00000000,\ 0x\,efcdab89,\ 0x\,80000000,\ 0x\,82000000)$ gives the joint transition probability 0 (impossible transition).

4.3 Overall Collision Path Probabilities

The results from Subsects. 4.1 and 4.2 provide the overall probabilities for the near-collision paths 1, 2, and 3 in Table 2 after message modification. Table 1

below contains all the values calculated above (for the standard IV), the resulting collision path probabilities $2^{-41.64}$, $2^{-37.41}$ and $2^{-36.61}$, resp., and the relative frequencies obtained by practical experiments. The number of samples was $2^{41.866}$. Interestingly, although near-collision path 3 even demands one bit condition more than the near-collision paths 1 and 2 (39 instead of 38) it is the most probable one (cf. Example 4).

steps	23	35	62	61p	62p	63p	64p	rest	theor. prob.	emp. value
Path 1	2^{-2}	2^{-1}	$(1-2^{-7})2^{-2}$	0.1875	0.789	0.034	2^{-4}	2^{-25}	$2^{-41.64}$	$2^{-40.86}$
Path 2	2^{-2}	2^{-1}	$(1-2^{-7})2^{-2}$	0.8125	0.789	0.148	2^{-4}	2^{-25}	$2^{-37.41}$	$2^{-37.11}$
Path 3	2^{-2}	2^{-1}	$(1-2^{-7})2^{-2}$	0.8125	0.789	0.516	2^{-4}	2^{-26}	$2^{-36.61}$	$2^{-36.25}$

Table 1 Transition probabilities for the three paths of Table 2

Our experiments revealed that also other (slightly different) near-collision paths than listed in Table 1 may lead to the near-collisions that satisfy identical conditions after the postaddition. This means that the path probabilties of concrete near-collision paths only give upper bounds for the workload of collision attacks. It appears that this effect also diminishes the impact of the IV.

Due to (32) the probabilities for the collision paths in the second block are significantly larger than those for the near-collision paths in the first block. We note that a specific sample path after message modification in Steps 1 to 20 occurs with probability $2^{-30.01}$ as it saves bit conditions on the postadditions.

5 Conclusion

We presented a stochastic model and a general method for an explicit computation of the probability of concrete (near-)collision paths after message modification. The computed probabilities for the MD5-near-collision paths 2 and 3 were found to be in good conformance with experimental results. It may thus be expected that similar calculations for SHA-1 (once the details of the attack anounced in [WYaYa] are published) should deliver reliable estimates for the probability of concrete collision paths. An interesting observation in the MD5 case was the significant impact of the post additions and the IV on these probabilities. If several near-collision-paths result in the same near-collision (i.e. in equal bit conditions after the post additions) this effect yet may diminish.

References

[BCH] J. Black, M. Cochran, T. Highland *A Study of the MD5 Attacks: Insights and Improvements* , FSE 2006, to appear in Springer LNCS

[Daum] M. Daum, *Cryptanalysis of Hash Functions of the MD4-Family*, PhD thesis, Ruhr-Universität Bochum, June 2005

[DL] M. Daum, S. Lucks, *The Story of Alice and Bob*, Presented at the rump session of Eurocrypt '05, May 2005, online at http://www.cits.rub.de/imperia/md/content/magnus/rump_ec05.pdf

[GIS1] M. Gebhardt, G. Illies, W. Schindler, *A Note on the Practical Value of Single Hash Collisions for Special File Formats*, Sicherheit 2006 — 'Sicherheit — Schutz und Zuverlässigkeit', Köllen, LNI P-77 (2006), 333-344.
Extended version: NIST Cryptographic Hash Workshop 2005, online at http://www.csrc.nist.gov/pki/Hashworkshop/2005/Oct31_Presentations/..
..Illies_NIST_05.pdf

[GIS2] M. Gebhardt, G. Illies, W. Schindler: The Impact of the IV on Multiblock Hash Collision Paths, FSE 2006, rump session, 16 Mar 2006.
http://fse2006.iaik.tugraz.at/rumpsession.html

[HPR] P. Hawkes, M. Paddon, G. D. Rose *Musing on the Wang et. al. MD5 Collision*, Cryptology ePrint Archive, Report 2004/264, http://eprint.iacr.org/2004/264.

[Kli1] V. Klima, *Finding MD5 Collisions on a Notebook PC Using Multi-messagenModifications*, Cryptology ePrint Archive, Report 2005/102, http://eprint.iacr.org/2005/102.

[Kli2] V. Klima, *Tunnels in Hash-Functions: MD5 Collisions Within a Minute*, Cryptology ePrint Archive, Report 2006/105, http://eprint.iacr.org/2006/105.

[LiLa] J. Liang, X. Lai, *Improved Collision Attack on Hash Function MD5*, Cryptology ePrint Archive, 23 Nov 2005 , Report 2005/425, http://eprint.iacr.org/2005/425.

[MOV] A. Menezes, P. C. van Oorschot, S. A. Vanstone, *Handbook of Applied Cryptography*, CRC Press, Boca Raton, 1997.

[MPRR] F. Mendel, N. Pramstaller, C. Rechberger, V. Rijmen *The Impact of Carries on the Complexity of Collision Attacks*, FSE 2006,to appear in Springer LNCS

[SNKO] Y. Sasaki, Y. Naito, N. Kunihiro, K. Ohta *Improved Collision Attack on MD5*, Cryptology ePrint Archive, 07 Nov 2005 , Report 2005/400, http://eprint.iacr.org/2005/400.

[SO] M. Schläffer, E. Oswald *Searching for Differential Paths in MD4*, FSE 2006,to appear in Springer LNCS

[St] M. Stevens *Fast Collision Attack on MD5*, Cryptology ePrint Archive, 17 Mar 2006 , Report 2006/104, http://eprint.iacr.org/2006/104.

[Th] S. Thomsen *Cryptographic Hash Functions*, Master thesis, Technical University of Denmark, November 2005

[WLFCY] X. Wang, X. Lai, D. Feng, H. Chen and X. Yu, *Cryptanalysis of the Hash Functions MD4 and RIPEMD*, EuroCrypt 2005, Springer LNCS 3494 (2005), 118.

[WY] X. Wang and H. Yu , *How to Break MD5 and Other Hash Functions*, EuroCrypt 2005, Springer LNCS 3494 (2005), 1935.

[WYaYa] X.Wang, A. Yao, F. Yao *New Collision Search for SHA-1*, Presented by Adi Shamir at the rump session of Crypto '05, Aug 2005, online at http://www.iacr.org/conferences/crypto2005/rumpSchedule.html

[WYiY] X. Wang, Y. L. Yin, H. Yu, *Collision Search Attacks on SHA-1*, Crypto 2005, Springer LNCS 3621 (2005), 17-36.

[WYuY] X. Wang, H. Yu, Y. L. Yin *Efficient Collision Search Attacks on SHA0*, Crypto 2005, Springer LNCS 3621 (2005), 1-16.

[YaSh] J. Yajima, T. Shimoyama *Wang' s sufficient conditions on MD5 are not sufficient*, Cryptology ePrint Archive, 10 Aug 2005 , Report 2005/236, http://eprint.iacr.org/2005/236.

Appendix

In Table 1 below bit conditions for three MD5-near-collision paths for block 1 are given. If the conditions for Step i are the same for each path the three columns are merged to a single column. The terms [j] and [-j] were already defined in Sect. 3. Further, $r_{i,j}$ denotes the j^{th} bit of r_i, and [*32] stands for $r_{i,32} = r'_{i,32}$. The additional conditions in Step 21, Step 35, and Step 62 (cf. Example 3) are not listed in Table 1. The conditions for Step 1 to Step 20 are as in [WY] (apart from additional conditions as in Steps 21, 35 and 62). Due to the lack of space these conditions are omitted.

Path 1 corresponds to the published bit conditions in [WY] while their published collision satisfies the bit conditions of path 2.

Step	Path 1	Path 2	Path 3
21	$[32]$, $r_{21,18} = r_{20,18}$		
22	$[32]$		
23	$r_{23,32} = 0 = r'_{23,32}$		
24	$r_{24,32} = 1 = r'_{24,32}$		
25... 34			
35	$[*32]$		
36	$[*32]$		
37	$[*32]$		
38	$[*32]$		
39	$[*32]$		
40	$[*32]$		
41	$[*32]$		
42	$[*32]$		
43	$[*32]$		
44	$[*32]$		
45	$[*32]$		
46	$[32]$		
47	$[32]$	$[-32]$	$[-32]$
48	$[32]$		
49	$[32]$	$[-32]$	$[-32]$
50	$[-32]$		
51	$[32]$	$[-32]$	$[-32]$
52	$[-32]$		
53	$[32]$	$[-32]$	$[-32]$
54	$[-32]$		
55	$[32]$	$[-32]$	$[-32]$
56	$[-32]$		
57	$[32]$	$[-32]$	$[-32]$
58	$[-32]$		
59	$[32]$	$[-32]$	$[-32]$
60	$[32]$, $r_{60,26} = 0 = r'_{60,26}$		
61	$[32]$	$[-32]$	$[-32]$
61	$r_{61,27} = 0 = r'_{61,27}$, $r_{61,26} = 1 = r'_{61,26}$		
62	$[32, 26]$		
63	$[32, 26]$	$[-32, 26]$	$[-32, 27, -26]$
64	$r'_{64} - r_{64} = 2^{31} + 2^{25} \pmod{2^{32}}$		
61,p	$[32]$		
62,p	$[32, 26]$		
63,p	$[32, 27, -26]$		
64,p	$[32, 26]$, $r^p_{64,27} = 0 = r'^p_{64,27}$, $r^p_{64,6} = 0 = r'^p_{64,6}$		

Table 2 Three different MD5 near-collision paths in the 1st block after message modification