Reproducing And Improving MD5 Fast Collision Attack Algorithm

Meng Zheng
Electrical Engineering
Delft University of Technology
Delft, Netherlands
M.Zheng-3@student.tudelft.nl

Yuhang Tian
Embedded Systems
Delft University of Technology
Delft, Netherlands
Y.Tian-13@student.tudelft.nl

***Abstract*—This paper is the survey for exploring the MD5 fast collision based on M. Stevens' paper published in 2012 where he extends the research carried out by Wang et al. and shows that the original structure of differential path cannot be proved as satisfying the "sufficient conditions". Wang's collision method has the limitation of the identical prefix, whereas Steven breaks this limitation to make the method become chosen-prefix collision attacks, which is the main contribution for him compared to primitive achievements. Our work, in this survey, is mainly 1) to explain the significance of MD5 collision, 2) to reproduce the fast collision attack algorithm for MD5, 3) to improve the current collision method or 3) to apply the methods in a real attack.**

***Index Terms*—collision attack, MD5, cryptography**

I. Introduction

For a one-way hash function, the collision attack is to find two different messages M and M' that have the identical hash values $H(M) = H(M')$. In terms of the Merkle-Damgard 4 (MD4) family, the most efficient technology to find the collisions for them is *differential cryptanalysis*. The core idea is to investigate and control the differences between two preimages when they propagate through the hash blocks and to generate two identical digests at the end. Following this idea, Xiaoyun Wang et al., published a paper in 2004 and proposed an effective attack for MD5 and achieved the collision attack. The paper points out that using two consecutive blocks with some constraints can generate two 1024-bit messages with identical hash digests - M_1 varies from M_2 ($M_1 \neq M_2$) but $MD5(M_1, IHV_i) = MD5(M_2, IHV_i)$.

MD5 structure has the property that if $MD5(x) = MD5(y)$ then $MD5(x||z) = MD5(y||z)$ where "||" denotes concatenation. According to this structure-property and combining the result provided by Wang, Dan Kaminsky in 2005 created *Stripwire* to mislead integrity checking, and so did Ondrej Mikle. This kind of attacks based on Wang's knowledge is *identical-prefix collision attack*. However, this type of attacks is quite limited as it requires that the initial IHV for those two consecutive blocks ought to be the same.

Later on, in 2007, Marc Stevens made great progress who removed that constraint - no requirement for identical prefix anymore - and he generated *chosen-prefix collision attacks* for MD5 which was also his main contribution. Without the restriction of the identical prefix, he applied this attack to forgery certificates such as rogue X.509 CA. X.509 was widely adopted to guarantee the security of the HTTPS website. The result reassured that MD5 could not be a secure certificating scheme anymore.

The following report will be generally divided into 4 sections. Firstly, it will detail the structure of the MD5 algorithm and give demos with both python and java versions. Following that, it will demonstrate the identical-prefix collision attack created by Wang et al. which is a brilliant attacking method different from the original methods based on brute-forcing or birthday-paradox and will show how this affects the secrecy of MD5. After that, it will present another elegant and advanced method - chosen-prefix collision attack with its application. Finally, it will give a summary that consists of the reflection for the project and suggestions for future work.

II. MD5 Message-Digest Algorithm

The MD5 message-digest algorithm, MD5 algorithm for short, takes the number of N 512 bits messages including padding and an initial 128 bits IHV as inputs to generate a 128 bits fingerprint. The whole algorithm process can be separated into three parts.

Part I: Initially, it uses method 1 padding - append a bit '1' to the message and then append many bits '0' - to make the length of the padded message become 448 (mod 512). After that, for the last remaining 64 bits, it appends the length of the original unpadded message. As a consequence, the message length is now $512N$ where N is an integer.

Part II: It will divide the padded message into N 512-bit segments, for each of which it will assign a block that takes a segment and a 128-bit IHV_{in} as inputs and output a 128-bit IHV_{out}. Except for the first block which IHV is set manually, the input IHV_{in} of a block is given by the output IHV_{out} from the former block. All the blocks are concatenated in serial like Fig. 1.

Part III: For each block, it will continue to divide the input 512-bit segment into 16 words (32 bits). It will run 4 rounds and each round consists of 16 steps; as a result, there are 64 steps in total. Each step includes some bit-wise operations and their values may vary among steps. One-step structure has been shown in Fig. 2. More details can be found in the appendix. In the final step of a block, it will add the current A', B', C', D' to the very beginning $IHV_{in} = A, B, C, D$ as

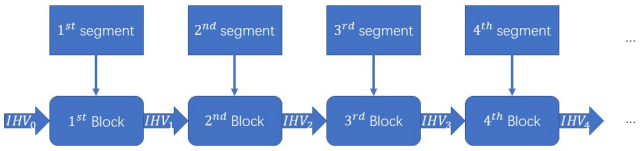

Fig. 1. Blocks Ordering

the output IHV_{out}. If the current block is not the last, it will propagate the output to the next block as the input for the next one; otherwise, it will output it as the final result.

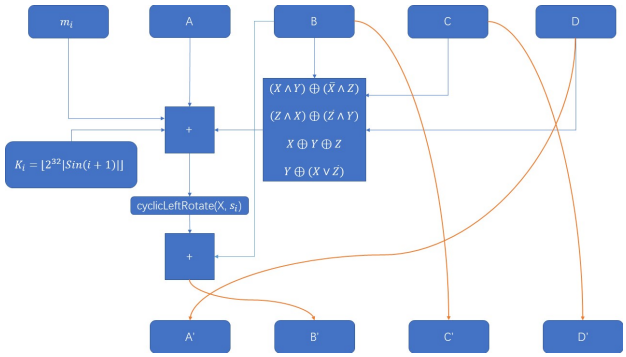

Fig. 2. MD5 Main Structure

For example,

III. Introduction

This document is a model and instructions for LaTeX. Please observe the conference page limits.

IV. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

V. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections V-A–V-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: “Wb/m^2” or “webers per square meter”, not “webers/m^2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
- Use a zero before decimal points: “0.25”, not “.25”. Use “cm^3”, not “cc”.)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

D. LaTeX-Specific Advice

Please use “soft” (e.g., `\eqref{Eq}`) cross references instead of “hard” references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don’t use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you’ve discovered a new method of counting.

BIBTEX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BIBTEX to produce a bibliography you must send the .bib files.

LATEX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LATEX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 3", even at the beginning of a sentence.

TABLE I
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 3. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an

example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks . . .”. Instead, try “R. B. G. thanks. . .”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.