パラメータ固定ハッシュ関数の理論的安全性評価に関する一考察 — 関数の近似可能性の観点から —

縫田 光司 † 阿部 拓郎 ‡ 鍛冶 静雄 ◇ 沼田 泰英 ⋆≀ 前野 俊昭 △

† 産業技術総合研究所情報セキュリティ研究センター
305-8568 茨城県つくば市梅園 1-1-1 中央第 2
k.nuida@aist.go.jp

‡ 京都大学大学院工学研究科機械理工学専攻
606-8501 京都市左京区吉田本町
abe.takuro.4c@kyoto-u.ac.jp

◇ 山口大学理学部数理科学科
753-8511 山口市吉田 1677-1
skaji@yamaguchi-u.ac.jp

⋆ 東京大学大学院情報理工学系研究科
113-8656 東京都文京区本郷 7-3-1
numata@stat.t.u-tokyo.ac.jp

≀ 独立行政法人科学技術振興機構（JST） CREST

△ 京都大学大学院工学研究科電気工学専攻
606-8501 京都市左京区吉田本町
maeno@kuee.kyoto-u.ac.jp

あらまし ＳＨＡ－１やＭＤ５に代表される、パラメータを具体的に固定したハッシュ関数の衝突耐性の解析は、実社会における情報セキュリティ技術の安全性保証のために重要度の高い問題であるが、その理論的な取り扱いは極めて困難であることが以前から知られていた。本研究では、パラメータ固定型ハッシュ関数の衝突耐性に関する理論的解析手法の確立に向けた端緒とすべく、ハッシュ関数の衝突耐性解析に関する従来手法のエッセンスを抽出する形で新たな数学的問題の定式化を行い、その問題について考察を行った。

On Security Evaluation of Hash Functions from a Viewpoint of Function Approximation

Koji Nuida† Takuro Abe‡ Shizuo Kaji◇ Yasuhide Numata⋆≀ Toshiaki Maeno△

† Research Center for Information Security (RCIS), National Institute of Advanced Industrial Science and Technology (AIST) k.nuida@aist.go.jp
‡ Department of Mechanical Engineering and Science, Kyoto University abe.takuro.4c@kyoto-u.ac.jp
◇ Department of Mathematical Sciences, Faculty of Science, Yamaguchi University skaji@yamaguchi-u.ac.jp
⋆ Department of Mathematical Informatics, The University of Tokyo numata@stat.t.u-tokyo.ac.jp
≀ Japan Science and Technology Agency (JST), CREST
△ Department of Electrical Engineering, Kyoto University maeno@kuee.kyoto-u.ac.jp

Abstract Security evaluation of hash functions with fixed parameters, such as SHA-1 and MD5, is a significant problem in real-life use of information security technology, but its theoretical treatment has been regarded as highly difficult. In this talk, we propose a new mathematical problem that is motivated from and relevant to abstraction of existing cryptanalytic techniques for hash functions, and show some preliminary observation of this problem.

1 はじめに

ハッシュ関数は情報セキュリティ分野における代表的な要素技術の一つであり、情報通信において受信したデータの破損や改竄の有無を確認する用途などに広く利用されている。通常、複雑な暗号プロトコルの一部分としてハッシュ関数を用いる際には、ハッシュ関数族の要素を指定するパラメータをランダムに選び、そのパラメータを攻撃者に対して秘匿することで安全性の理論的保証を可能とすることが多い。一方、より直接的なハッシュ関数の利用法としては、例えばインターネット経由で電子ファイルを配布したい場合に、ウェブサイト上にそのファイルとファイルのハッシュ値を一緒に載せておき、ユーザがファイルをダウンロードした後にハッシュ値をウェブサイトに記載された値と照合することで、ファイルの破損の有無を確認するといった利用法が広く行われている。この利用法では、ファイルの配布元がハッシュ値の計算に用いたハッシュ関数を公開する必要があるため、前述のようなパラメータのランダムな選択と秘匿による安全性保証ができないという問題がある。このようなパラメータを含めた内部アルゴリズムが固定され公になっている状況におけるハッシュ関数の安全性評価は、暗号・情報セキュリティ分野の主要な研究テーマの一つであり、多くの研究が進められている。

パラメータ固定型ハッシュ関数の理論的安全性評価に関しては、以下のような根源的な困難性が知られている。ハッシュ関数の主要な安全性要件の一つである衝突耐性（collision resistance）を例に挙げると、ハッシュ関数 H の定義域が値域よりも大きくないような自明（かつ実用的に無意味）な場合を除いて、鳩の巣原理により $H(x) = H(y)$ を満たす定義域の異なる元 x, y（衝突対、collision pair）は常に**存在する**。そのため、この x と y を予め内部に保持した攻撃者（このような攻撃者は — それを現実に構成する方法はさておき — 理論的には常に**存在する**）は H の衝突耐性を短時間で破ることが可能である。従って、通常の暗号学的議論で用いられるような「（○○の仮定の下で）このプロトコルの安全性を破る攻撃者は理論的に**存在しない**」という具合に定式化された安全性概念は、パラメータ固定型ハッシュ関数には適さないことになる。

このように、パラメータ固定型ハッシュ関数については、その安全性の理論的評価を行うことはおろか、厳密かつ意味のある形で安全性概念を定式化すること自体が困難である。そのため現状では、理論的な安全性評価の代わりに、与えられたハッシュ関数 H（MD5、SHA-1、SHA-3 の候補など）に対して実際に衝突対の構成を試みて、もし衝突対が構成できたら H は安全でない、構成できなければ暫定的に H を安全と看做す、という運用が行われている。換言すると、「H に関する衝突対の構成法を**我々が知らなければ**、H は安全である」と看做していることになる。しかし、この「我々が知らない」という安全性の根拠づけは、理論的な安全性評価に比べると極めて不安定なものである。安全な情報通信におけるパラメータ固定型ハッシュ関数の重要性を鑑みると、より理論的にしっかりした安全性の根拠を与えることは、理論的にも実用的にも重要な課題であると考えられる。なお、理論的な非存在証明ではなく「我々が知らない」ことを根拠とするという点は、現在の計算量的安全性理論における困難性仮定の取り扱いと共通するように感じられるかもしれない。しかし、素因数分解や離散対数問題については「効率的な解法が本当に存在しない」可能性が（古典計算機の範囲では）存在する一方、パラメータ固定型ハッシュ関数については「効率的な解法が本当に存在しない」可能性は上記の通り否定されるため、「我々が知らない」という状況の意味するところが異なると考えられる。もっとも、もしあるハッシュ関数に対する衝突対の構成法が（現時点での楕円曲線暗号の歴史と同じく）30 年以上の間発見されなければ、そのハッシュ関数の安全性は現時点での楕円曲線暗号と同程度には信頼されるのかもしれないが、今のところそこまで安全性が長持ちするパラメータ固定型ハッシュ関数が得られる見通しは立っていないものと思われる。

本研究では、このように非常に困難であるパラメータ固定型ハッシュ関数の安全性の定式化

と理論的評価について、どうにかして有効な手掛かりが得られないか検討を行った。上述の通り、存在するかしないかの二分法で考える限り、どんなパラメータ固定型ハッシュ関数も、潜在的な攻撃法が確実に存在するという意味で違いは見出せない。それでもなお、各々のハッシュ関数に対する潜在的な脆弱性の度合いを相対的に評価することにより、より「ましな」ハッシュ関数を選び出す一つの目安を提示できるのではないか、というのがその基本的なアイデアである。より詳しくは、「**既知の攻撃法＋ほんのちょっとの工夫**」（この工夫は、理論的な考察によるものかもしれないし、あるいは偶然や勘によるものかもしれない）で破れてしまうハッシュ関数とそうでないハッシュ関数では、後者の方がましな安全性を持つと期待できるであろうという考えである。本稿では、この「ほんのちょっとの工夫」の必要量を、ハッシュ関数の近似可能性という観点から定量的に評価するための枠組みの定式化を試みる。具体的には、与えられたハッシュ関数 H が、既知の攻撃法のみで破れるハッシュ関数の集合 $\mathcal{C}'$ の要素とどの程度似ているかによって、H の潜在的な脆弱性を評価しようという試みである。このハッシュ関数の類似度を評価する問題については、独立した数学的問題（「**関数密度問題**」、“**Function Density Problem**”）として新たに抽象化・定式化を行う。このように新たな数学的問題を提示することで、この課題に関する数学分野との連携研究の呼び水となることを期待している。

なお、関連研究として、Rogaway [3] もハッシュ関数の安全性に関して上述のような「存在しない」と「我々が攻撃法を知らない」の違いに着目し、ハッシュ関数を部品として含む暗号プロトコルの安全性を「我々が部品となるハッシュ関数の攻撃法を知らない」事実に帰着させる枠組みを提案している。本稿で提案する安全性評価の枠組みを Rogaway の手法と組み合わせることで、より広範な応用が得られるものと考えられる。

本稿の構成は以下の通りである。2.1 節では、パラメータ固定型ハッシュ関数の近似可能性の評価を抽象化する形で、「関数密度問題」という数学的問題を定式化し、2.2 節ではその問題に関する具体例を示す。3 節では、パラメータ固定型ハッシュ関数の安全性評価への関数密度問題の応用について論じる。4 節では、その他の題材に対する関数密度問題の応用の可能性を論じる。なお、本稿では頁数の都合上割愛した内容（命題等の証明を含む）も多いため、詳細については本年 11 月に行われる IWSEC 2011（The 6th International Workshop on Security）における本稿著者の発表 [2] を参照されたい。

2 関数密度問題

本節では、パラメータ固定型ハッシュ関数の安全性評価に応用する数学的問題である関数密度問題を定義し（2.1 節）、またその問題に慣れるために簡単な具体例を紹介する（2.2 節）。関数密度問題をハッシュ関数の安全性評価にどう応用するかについては次節で議論する。

2.1 定義

$\mathcal{C}$ をある種の関数からなる（有限）集合、$\mathcal{C}'$ をその部分集合とする。また、$\mathcal{C}$ に属する二つの関数 f, g について、それらの「距離」$d(f, g)$ が定められているものとする（大半の状況では d は実際に距離の公理を満たすであろう）。以上の状況において、次の数学的問題を定義する：

定義 1 (関数密度問題、Function Density Problem)**.** 以上の状況において、以下の量

$$r(\mathcal{C}; \mathcal{C}') = \max\{d(f; \mathcal{C}') \mid f \in \mathcal{C}\} \qquad (1)$$

（もしくはその上界や下界）を求める問題を**関数密度問題**（***Function Density Problem***）と呼ぶ。ここで $d(f; \mathcal{C}')$ は

$$d(f; \mathcal{C}') = \min\{d(f, g) \mid g \in \mathcal{C}'\} \qquad (2)$$

で定義される量（f と $\mathcal{C}'$ の距離）である。

$\mathcal{C}' \subset \mathcal{C}$ という関係から、d が通常の意味での距離関数である場合には $r(\mathcal{C}; \mathcal{C}')$ は $\mathcal{C}$ と $\mathcal{C}'$ の Hausdorff 距離に一致する。直感的な説明とし

ては、$r(\mathcal{C};\mathcal{C}')$ は部分集合 $\mathcal{C}'$ を中心としたときの全体集合 $\mathcal{C}$ の半径に相当する（'r' は "radius" の頭文字）。図 1 を参照されたい。勿論、これだけの問題設定ではあまりに抽象的過ぎて、なんら意味のある結果を導くものではないと容易に想像されるため、実際には $\mathcal{C}$、$\mathcal{C}'$ や「距離」d の性質を（応用したい具体的問題に合わせて）適度に規定した上で問題を考察することになる。

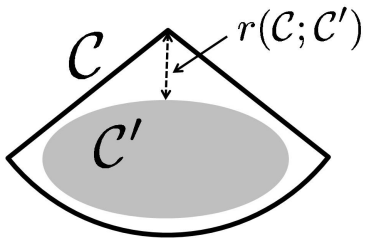

図 1: 量 $r(\mathcal{C};\mathcal{C}')$ の模式図

一つの設定としては、$\mathcal{C}$ をある有限集合 X から有限集合 Y への写像 $f\colon X \to Y$ からなる集合とし、距離 $d(f,g)$ を

$$d(f,g) = |\{x \in X \mid f(x) \neq g(x)\}| \qquad (3)$$

で定義することが考えられる（各 $f \in \mathcal{C}$ を自然な形で長さ $|X|$ の Y の元の有限列と同一視すると、この定義は一般化した Hamming 距離と一致する）。以下、本稿ではこの設定を採用する。

更に、例えば部分集合 $\mathcal{C}'$ として（ある意味で）「単純な」写像 $X \to Y$ の集合を取ると、$d(f;\mathcal{C}')$ が小さい $f \in \mathcal{C}$ は「単純な写像に良く似た」写像ということになり、$d(f;\mathcal{C}')$ が大きい場合に比べると f 自身が比較的「単純な」構造をしていると考えることができる。第 3 節ではこの考えをパラメータ固定型ハッシュ関数の安全性評価に応用することを試みる。

2.2 具体例

上述の関数密度問題に馴染むため、ここでは簡単な具体例を紹介する。$\mathcal{C}$ を n ビット入力 1 ビット出力の関数全体の集合とする（つまり $X = \{0,1\}^n$、$Y = \{0,1\}$ である）。$\mathcal{C}$ の部分集合 $\mathcal{C}'$ を定義する前に、Y を 2 元からなる有限体 $\mathbb{F}_2 = GF(2)$ と同一視することで、関数 $f\colon X \to Y$ は n 変数の square-free な多項式として以下のように表されることを注意しておく：

$$f(x_1,\ldots,x_n) = \sum_{a\in\{0,1\}^n} f(a_1,\ldots,a_n) \prod_{i;a_i=0} (1-x_i) \prod_{i;a_i=1} x_i \qquad (4)$$

（右辺に現れる積 $\prod_{i;a_i=0}(1-x_i)\prod_{i;a_i=1} x_i$ は、全ての i について $a_i = x_i$ のときにのみ 1 となりそれ以外では 0 となることに注意されたい）。以上を踏まえて、$\mathcal{C}' = \mathcal{C}'_k$ を上の多項式表示の次数が k 以下となる関数 $f \in \mathcal{C}$ 全体の集合と定義する。例えば $\mathcal{C}'_0$ は定数関数の集合、$\mathcal{C}'_1$ はアフィン関数の集合となる。この状況で、関数密度問題の対象である量 $r(\mathcal{C};\mathcal{C}')$ について以下が成り立つ（証明については [2] を参照されたい）：

命題 1. $u_{n,k} = \sum_{i=k+1}^{n} \binom{n}{i}$ とおき、また $2^{u_{n,k}} \leq \sum_{i=0}^{\ell} \binom{2^n}{i}$ を満たす最小の整数 ℓ を $\ell_{n,k}$ とおく。このとき

$$\ell_{n,k} \leq r(\mathcal{C};\mathcal{C}'_k) \leq \min\{u_{n,k}, 2^{n-1}\} \qquad (5)$$

が成り立つ。

なお、命題 1 の上界と下界については $\ell_{n,k} \leq u_{n,k} \leq n\ell_{n,k}$ という関係が成り立ち、概ね n 倍程度しか違わないことになる。この関係の証明についても [2] を参照されたい。参考までに、命題 1 の下界 $\ell_{n,k}$ の数値例を表 1 に記しておく。

表 1: 小さなパラメータに関する $\ell_{n,k}$ の数値例

		$n-k$							
		1	2	3	4	5	6	7	8
	2	1	2						
	3	1	2	4					
	4	1	2	4	8				
n	5	1	2	5	10	16			
	6	1	2	5	13	22	32		
	7	1	2	6	16	31	49	64	
	8	1	2	6	19	43	75	105	128

注意 1. 上の例で扱った 1 ビット出力の関数自体はハッシュ関数としては無意味なものであるが、多ビット出力のハッシュ関数を 1 ビット出力

関数の列として捉えることも可能であろうし、また [2] で提示する擬似乱数生成器の安全性評価への応用においては 1 ビット出力関数についての関数密度問題が充分な意味を持つことを注意しておく。

3 パラメータ固定型ハッシュ関数の安全性

本節では、2.1 節で導入した関数密度問題を応用して、パラメータ固定型ハッシュ関数の理論的な安全性評価について議論する。

2.1 節の後半と同様に、$\mathcal{C}$ として有限集合 X から有限集合 Y への写像からなる集合を考える。今の文脈では、X がハッシュ値を計算したいメッセージの集合、Y がハッシュ値の集合と解釈される。そして、ハッシュ関数 $H \in \mathcal{C}$ の衝突耐性、即ち $H(x_1) = H(x_2)$ を満たす異なる $x_1, x_2 \in X$ を構成することの困難性について考察する。

まず、H の衝突対を構成する戦略として以下のようなものを考える：

1. ハッシュ関数 H を、現実的な時間で衝突対が得られるように単純化された関数 $H' \in \mathcal{C}$ で近似する。
2. H' の衝突対 (x'_1, x'_2) をランダムに構成する。
3. (x'_1, x'_2) を基に、H の衝突対の候補 (x_1, x_2) を計算する（最も素朴な方針としては、単に $(x_1, x_2) = (x'_1, x'_2)$ とおく）。
4. この (x_1, x_2) が実際に $H(x_1) = H(x_2)$ を満たすかどうか確認する。
5. $H(x_1) = H(x_2)$ ならばそれが求める H の衝突対である。もし $H(x_1) \neq H(x_2)$ ならば手順 2 に戻って繰り返す。

この戦略はだいぶ抽象化されてはいるものの、ハッシュ関数に対する既存の攻撃法の多くの共通要素を含んでいると考える。以下、最も単純な $(x'_1, x'_2) = (x_1, x_2)$ の場合のみを取り扱う。

上記の戦略については、もし手順 1 で与えられた関数 H' が元のハッシュ関数 H を良く近似しているならば、手順 2 以降を繰り返す回数の期待値が小さくなることが直感的に予想されるであろう。この予想される傾向を定量的に評価したのが次の結果である：

補題 1. H と H' を X から Y への写像とし、$|Y| = n \geq 2$ と仮定する。もし $d(H, H') = d$ $(0 < d < |X|)$ ならば、H' の衝突対を一様ランダムに選んだときにそれが H の衝突対でもある確率は、少なくとも

$$\frac{2\alpha_0|X| - n(\alpha_0+1)\alpha_0 - 2d\alpha_0}{2\alpha_0|X| + 2d|X| - n(\alpha_0+1)\alpha_0 - 2d\alpha_0 - d^2 - d} \tag{6}$$

（ただし $\alpha_0 = \lfloor (|X| - d - 1)/n \rfloor$）である。そして $|X| \geq d + (n-1)^2$ であれば、式 *(6)* の値は d に関して単調減少する。

X が d や n よりも充分大きい場合、式 (6) は

$$1 - \frac{4dn}{2|X| + 4dn - 3d - 2n + 1} \tag{7}$$

（もしくは、より簡単に $1 - 2dn|X|^{-1}$）と近似できる。補題 1 はややこしいながらも初等的な計算によって証明できるが、頁数の関係で証明は省略する。詳細は [2] を参照されたい。

以上を踏まえた上で、次のような状況を想像してみよう。ハッシュ関数の新たな候補として、二つのパラメータ固定型ハッシュ関数 H_1 と H_2 が与えられている。H_1 はメッセージ長、ハッシュ値長やその他のパラメータ等に関して特徴 P_1 を、H_2 は特徴 P_2 を有している。今、特徴 P_i を有する全てのハッシュ関数の集合を $\mathcal{C}_i$ $(i = 1, 2)$ とおく（つまり $H_i \in \mathcal{C}_i$）。また、$\mathcal{C}_i$ の元で、ある既知の攻撃法によって衝突対が現実的な計算時間で得られるようなものの集合を $\mathcal{C}'_i$ とおく。更に、以下の性質が成り立っているものとする：

- $r(\mathcal{C}_1; \mathcal{C}'_1)$ は充分小さい。従って、$\mathcal{C}_1$ に属するどんなハッシュ関数 H も、良い近似 $H' \in \mathcal{C}'_1$ を見つけさえすれば、H' の衝突対を構成する既知の攻撃法と上述の戦略との組合せによって H の衝突耐性が高い確率で破られる、という潜在的な脆弱性が存在する。
- $r(\mathcal{C}_2; \mathcal{C}'_2)$ は充分大きい。従って、$\mathcal{C}_2$ に属する少なくとも一つのハッシュ関数 H につい

ては、良い近似 $H' \in \mathcal{C}'_2$ を見つけて H' の既知の攻撃法を適用するという戦略は確率的には上手くいかないということになる。

このとき、件の既知の攻撃法とハッシュ関数の近似の組合せ、という戦略に対しては、$\mathcal{C}_1$ に属している H_1 については前者の通り脆弱性を秘めていることが確実であるのに対し、$\mathcal{C}_2$ に属している H_2 については後者の通り安全である可能性が残されている。従って、もし H_1 と H_2 のどちらを採用するかについて他の判断材料から考えて五分五分の状況なのであれば、上記の理由から H_2 の方を選んでおいた方が相対的に安全である可能性が高いものと考えられる。このように、ハッシュ関数に対する近似可能性という観点から、潜在的な脆弱性の度合いを相対的に評価するというのが本稿の提案内容である。

勿論、現実の具体的な状況においていかにしてハッシュ関数の集合 $\mathcal{C}$ と部分集合 $\mathcal{C}'$ を適切に定め、量 $r(\mathcal{C};\mathcal{C}')$ の有効な評価を与えるかという点は依然として大きな課題として残されている。それでもなお、パラメータ固定型ハッシュ関数の安全性の理論的取り扱いという困難な課題に対して、本稿の提案手法が一つの足掛かりとなることを期待するものである。

4 関数密度問題のその他の応用

関数密度問題の情報セキュリティ分野への応用としては、本稿で議論したパラメータ固定型ハッシュ関数の理論的安全性評価の他に、多値出力を持つ識別者に対して安全な擬似乱数生成器（pseudorandom generators that fool non-boolean distinguishers, “nb-PRGs”）[1] の理論へ応用できることが著者らの研究により明らかとなっている。より具体的には、「ある安全性パラメータ（の組）P を持つ擬似乱数生成器（PRG）は、安全性パラメータ（の組）P' を持つ nb-PRG でもある」という形の関係について、関数密度問題を応用して P と P' の関係式を導出する手法を見出した。この成果については、本年 11 月に行われる IWSEC 2011（The 6th International Workshop on Security）における発表 [2] を参照されたい。

またそれ以外にも、例えば McEliece 暗号などの符号ベース暗号、松本–今井暗号などの多次多変数暗号、NTRU 暗号などの格子ベース暗号のように、一様ランダムな入力に対して NP-困難な計算問題を背景としているものの実際の暗号系に付随する問題の分布は一様ランダムではない類の暗号系について、関数密度問題と同様の手法により、実際の問題の分布と一様ランダムな分布との「距離」を評価できないだろうかと考えている。この問いについては今後の研究課題としたい。

5 まとめ

本稿では、パラメータ固定型ハッシュ関数の近似可能性の評価という問題を抽象化する形で、関数密度問題という数学的問題を新たに定式化・提案した。またその問題を応用する形で、パラメータ固定型ハッシュ関数の理論的安全性評価に関する一つの手法を提案した。更に、情報セキュリティ分野の別の課題に対する関数密度問題の応用の可能性について論じた。

謝辞 4 節で論じた関数密度問題の別の応用については、花岡悟一郎氏より有益なコメントを頂いたのでこの場を借りて感謝の意を表する。

参考文献

[1] B. Dubrov, Y. Ishai, “On the randomness complexity of efficient sampling,” in: Proc. STOC 2006, pp. 711–720 (2006)

[2] K. Nuida, T. Abe, S. Kaji, T. Maeno, Y. Numata, “A mathematical problem for security analysis of hash functions and pseudorandom generators,” to appear in IWSEC 2011, Tokyo, Japan, Nov. 8–10, 2011.

[3] P. Rogaway, “Formalizing human ignorance - collision-resistant hashing without the keys,” in: Proc. VIETCRYPT 2006, pp. 211–228 (2006)