New combinatorial bounds for universal hash functions

Abstract

Using combinatorial analysis, we introduce a new lower bound for the key length in an *almost* universal hash function, which is tighter than another similar bound derived from an well-studied equivalence between almost universal hashes and error-correcting codes. To the best of our knowledge, this is the first time when combinatorial analysis yields a better universal hash bound than the use of the relation, and we will explain why there is a mismatch. We then compare the new bound against known bounds for this and other families of universal hashes and discover an important value of hash collision probability, which not only represents a *threshold* in the behaviour of bounds but also quantifies the *Wegman-Carter* effect.

1 Introduction and contribution

Universal hash function H with parameters (ϵ, r, K, b) was introduced by Carter and Wegman [8, 35]. Each family, which is indexed by a r-bit key k, consists of 2^r hash functions mapping a message representable by K bits into a b-bit hash output: $H\ (r, K, b) = \{h_k() : \{0,1\}^K \rightarrow \{0,1\}^b | k \in [0, 2^r)\}$.

In this paper we use combinatorial analysis to introduce a new bound, termed the *combinatorial* bound, for an *almost* universal hash function (AU). This result tells us the lower bound on the bitlength of the hash key with respect to a fixed amount of information we want to hash, the hash output bitlength and the hash collision probability ϵ.

Although there has been much work in this area, most researchers concentrate on *almost strongly* and *almost XOR* universal hash functions (ASU and AXU, more restrictive versions of AU as can be seen in their definitions below), because a much-used mechanism in practice, called MAC, use AXU or ASU [28, 29, 16, 17, 11, 15, 35, 8, 1]. We however believe that there is a similar potential for AU. For example, a new class of authentication schemes, based on new concepts of trust derived from human actions and interactions, has been recently proposed to replace PKI and passwords in pervasive computing environments [27, 34, 9, 22, 33, 20]. Some of these protocols make use of a new cryptographic *digest* function introduced in [22], with similar security properties and purposes to an AU. In these protocols, digest or hash keys are always random and fresh in each protocol session, and so a substitution attack, which relies of the reuse of a hash key for multiple messages, is irrelevant. Hence, what we require is a protection against hash collision attacks (AU) as opposed to substitution attacks (ASU).

Moreover, since universal hash keys in MAC are often large, one reuses a single secret key for multiple messages as mentioned above. This opens the way for key recovery and universal forgery attacks which exploit weak key properties or partial information on a secret key; such attacks have been recently reported by Handschuh and Preneel [12]. Avoiding reusing keys would render most key recovery attacks useless, and so it is desirable to construct universal hashes with short keys, which in turn generate the need to calculate the lower bound of universal hash key length.

We are aware of an equivalence between *almost* universal hash functions and error-correcting codes discovered by Johansson et al. [14], which implies that every bound of coding theory potentially corresponds to another bound for universal hashes, and vice versa. Consequently, we will show how to use the *Singleton* bound to derive a different *AU*-bound which is, however, not as tight as our combinatorial one. In particular, there exists a subclass of an *AU* which cannot be tranformed into an equivalent code that satisfies Singleton bound with equality, thus Singleton bound does not give tight result for the subclass of universal hash. To the best of our knowledge, this is the first time when combinatorial analysis yields a better universal hash bound than the use of the relation. However, the combinatorial *AU*-bound when being converted into parameters in coding theory becomes no better than Singleton bound.

In comparing the combinatorial *AU*-bound to Stinson's *AU*-bound [28, 29], we discover the significance of the value $(1+\frac{b}{K-b})2^{-b}$: as ϵ increases beyond the threshold, our bound is tighter than Stinson's *AU*-bound. Subsequently this threshold value will be shown to have the same theoretical significance in relationships between known *AXU*- and *ASU*-bounds. What this illustrates is a behaviour of any universal hash functions, known as the "Wegman-Carter effect" in the literature [6, 19], previously reported in [14, 15] by Johansson, Kabatianskii and Smeets: if ϵ exceeds 2^{-b} (the theoretical minimum[1]) by an arbitrarily small positive value, then the total number of messages, that can be authenticated, grows exponentially with the number of keys provided, but if $\epsilon = 2^{-b}$ it only grows linearly. However, while these authors only demonstrate this behaviour asymptotically, we are able to *quantify* it using the threshold value.

We end this paper by proving the *optimality* of polynomial hashing over finite field [7, 14, 32] in building an *AU*, i.e. the construction meets the combinatorial *AU*-bound with equality. This therefore extends the work of Johansson, Kabatianskii and Smeets [14, 15], where the authors proved the *asymptotic optimality* of polynomial hashing as an *ASU*.

In our work, we also introduce a new bound for an *AXU*. The bound is derived from Kabatianskii's *ASU*-bound [15] and a connection between *ASU* and *AXU* [35, 10]. The bound is then rigorously analysed in relation to other known bounds and will be shown to be met with equality in the second version of polynomial hashing.

2 Notations and definitions of universal hash functions

In this paper, all formulas are expressed in terms of bitlengths[2] of hash keys, input messages and hash output instead of the cardinalities of the sets of these parameters (2^r, 2^K and 2^b) as in other papers. The advantage of the notation will become clear when we explain why combinatorial analysis yields better bounds than the use of coding theory bounds in Section 3.2.

Let us recall the definitions of a number of families of universal hash functions. Here ϵ, which is sometimes written as $2^{\theta-b} = \gamma 2^{-b}$, is referred to as the collision, differential or interpolation probability associated with ϵ-*AU*, ϵ-*AXU* or ϵ-*ASU*, respectively.[3] In all following definitions, we look at the probability of some condition being met, e.g. hash collision, as the key k varies uniformly over its domain: $\mathbf{Pr}_k[]$.

[1]In practice, the minimum collision probability of an *AU* is $\frac{2^K-2^b}{2^{K+b}-2^b}$, which is less than 2^{-b}. This occurs in an *optimally universal* hash scheme introduced by Sarwate [25].

[2]Although r, K and b are often integers in practice, our result applies to both integer and noninteger bitlengths.

[3]The terms collision, differential and interpolation probabilities were introduced by Bernstein in the appendix of [3] to distinguish the differences between these families of universal hash functions.

An ϵ-*almost* universal hash function, ϵ-*AU* (r, K, b) [8, 28]
H is an ϵ-AU iff for all different messages m and $\hat{m}$: $\mathbf{Pr}_k[h_k(m) = h_k(\hat{m})] \leq \epsilon$

An ϵ-*almost XOR* universal hash function, ϵ-*AXU* (r, K, b) [16, 17, 28]
H is an ϵ-AXU iff for every pair of distinct messages $(m, \hat{m})$ and any $\omega \in \{0,1\}^b$: $\mathbf{Pr}_k[h_k(m) \oplus h_k(\hat{m}) = \omega] \leq \epsilon$

An ϵ-*almost strongly* universal hash function, ϵ-*ASU* (r, K, b) [35, 28]
(a) For every message m and hash output y: $\mathbf{Pr}_k[h_k(m) = y] \leq 2^{-b}$. (b) For every pair of distinct messages $(m, \hat{m})$ and for every pair of hash outputs $(y, \hat{y})$: $\mathbf{Pr}_k[h_k(m) = y, h_k(\hat{m}) = \hat{y}] \leq \epsilon 2^{-b}$

All *universal* hash functions discussed to date are pairwise, since we look at their properties in relation to two different messages. We will see that the combinatorial bound, and its proof, can be easily adapted to a more general version of *AU*, termed a l-wise AU_l, and therefore we give the definition below. We argue that not only is this of theoretical interest to study AU_l, but also useful in many applications, such as in the new family of authentication protocols discussed in the introduction, where the intruder attempts to fool multiple parties into accepting different versions of a piece of data that the protocol seeks to ensure they agree on. It is therefore desirable that we consider the possibility of a hash collision w.r.t more than two different input messages. However, unless indicated, our work presented in this paper always refers to pairwise universal hash functions.

A l-wise ϵ-*almost* universal hash function, ϵ-AU_l (r, K, b)
H is an ϵ-AU_l iff for any l different messages $\{m_1, \ldots, m_l\}$: $\mathbf{Pr}_k[h_k(m_1) = \cdots = h_k(m_l)] \leq \epsilon$

We assume the input message bitlength K is significantly greater than the hash bitlength b. Whenever we use the term $\log X$, we refer to the logarithm of base 2 to simplify the notation.

3 Bounds for almost universal hash functions

We first present a new bound of an *AU* using combinatorial analysis. Subsequently, *Singleton* bound [23] in coding theory will be used to derive another *AU*-bound,[4] which is not as tight as we had expected. In particular, when K is *not* a multiple of b, the combinatorial *AU*-bound is tighter (greater) than the latter. This result applies to both integer and non-intger values of K and b.

3.1 Combinatorial *AU*-bound

Theorem 1 *If there exists an ϵ-AU (r, K, b) then*

- *when K is a multiple of b:* $r \geq \log\left(\epsilon^{-1}\left(\frac{K}{b} - 1\right)\right)$
- *when K is not a multiple of b:* $r \geq \log(\epsilon^{-1}\left\lfloor\frac{K}{b}\right\rfloor)$

[4] Although several bounds in coding theory have been transformed into equivalent bounds for universal hashes, e.g. Plotkin bound [29] or Johnson bound [15], to the best of our knowledge, Singleton bound has never been used.

The proof makes use of the *pigeon-hole* principle: given two positive integers n and m, if n items are put into m holes then at least one hole must contain more than or equal to $\lceil n/m \rceil$ items.

Proof Let assume $K = bt + b'$, where t is an integer and $0 \leq b' < b$.

For any key k_1, there exists a hash value h_1 such that there are at least $\lceil 2^{K-b} \rceil$ different messages all hashing to h_1 under the same key k_1, thanks to the pigeon-hole principle. For any choice of k_2 other than k_1, there will also be a collection of at least $\lceil 2^{K-2b} \rceil$ of these messages mapping to some hash value h_2 under k_2. Repeating this process $(t-1)$ times will result in at least $\lceil 2^{K-(t-1)b} \rceil = \left\lceil 2^{b+b'} \right\rceil$ distinct messages that hash to the same values under $(t-1)$ keys, leading to two possibilities.

- When $b' = 0$ or K is a multiple of b, we *cannot* repeat this process any further because at least 2 distinct messages must be left after these interations. Thus a family of hash functions is ϵ-*almost universal* when $(t-1)$ is smaller than or equal to ϵ portion of the key space: $\epsilon 2^r \geq t - 1 = K/b - 1$, which means that $r \geq \log(\epsilon^{-1}(K/b - 1))$

- When $b' > 0$ or K is *not* a multiple of b, we have $\left\lceil 2^{b+b'} \right\rceil \geq 2^b + 1$. Repeating this process for one more key will end up with at least 2 distinct messages that map to the same values under $t = \lfloor K/b \rfloor$ keys. As a result, we have $\epsilon 2^r \geq \lfloor K/b \rfloor$, which means that $r \geq \log(\epsilon^{-1} \lfloor K/b \rfloor)$ ∎

Fortunately, the proof of the combinatorial bound for a pairwise AU_2 can be generalised to derive the corresponding bound for a l-wise AU_l, for any integer $l \geq 2$. Instead of leaving at least 2 different messages after these iterations as shown in the proof of Theorem 1, we need to leave l distinct messages. Similar analysis leads to the following theorem.

Theorem 2 *If there exists a l-wise ϵ-AU_l (r, K, b) and $K = bt + b'$, where t is an integer, then*

- *when* $0 \leq b' \leq \log(l-1)$*:* $r \geq \log\left(\epsilon^{-1}\left(\left\lfloor \frac{K}{b} \right\rfloor - 1\right)\right)$
- *when* $\log(l-1) < b' < b$*:* $r \geq \log\left(\epsilon^{-1}\left\lfloor \frac{K}{b} \right\rfloor\right)$

Although there has been some study of l-wise *almost strongly* universal hash functions by Stinson [30] and Kurosawa et al. [18], as far as we are aware, this is the first result on l-wise *almost* universal hash functions.

We end this section with another observation: there is no restriction on any of the parameters, i.e. bitlengths K, b and r, in both our pairwise and l-wise combinatorial AU-bounds, which makes them more attractive than a similar ASU-bound introduced by Kabatianskii et al. [15], as will be discussed in the sections to come.

3.2 Error-correcting codes and almost universal hash functions

While the connection between almost universal hashes and error-correcting codes (i.e. see Theorem 3), which was first observed by Johansson et al. [14], has often been used by reseachers to derive tight bounds for universal hashes [28, 29, 14, 15], the following comparative analysis will demonstrate that the strategy does not always give the best answer.

Let (n, T, d, q) be a q-ary error-correcting code, where n is the number of symbols in each codeword, T is the total number of codewords, and the minimum Hamming distance is d.

Theorem 3 **[14, 5, 29]**. *If there exists an* ϵ*-AU* (r, K, b)*, then there exists an* $(n = 2^r, T = 2^K, d = n - n\epsilon, q = 2^b)$ *code. Conversely, if there exists an* (n, T, d, q) *code, then there exists an* $(\epsilon = 1 - d/n)$*-AU* $(r = \log n, K = \log T, b = \log q)$*.*

Using the connection, we can derive another *AU*-bound from Singleton bound.

Singleton bound [23]: given an (n, T, d, q) code then $q^{n-d+1} \geq T$.

Theorem 4 *Another bound for an* ϵ*-AU* (r, K, b) *is:* $r \geq \log\left(\epsilon^{-1}\left(K/b - 1\right)\right)$

Proof Using Theorem 3, construct an $(n = 2^r, T = 2^K, d = n - n\epsilon, q = 2^b)$ code from the universal hash funcation ϵ-*AU* (r, K, b). This code must satisfy Singleton bound, so we obtain:

$$\begin{aligned} q^{n-d+1} &\geq T \\ 2^{b(\epsilon 2^r + 1)} &\geq 2^K \\ r &\geq \log\left(\epsilon^{-1}\left(K/b - 1\right)\right) \quad \blacksquare \end{aligned}$$

When K is a multiple of b, this is equivalent to the combinatorial *AU*-bound in Theorem 1.

In contrast, when K is not a multiple of b, the combinatorial bound is clearly tighter (or greater) than the one derived in Theorem 4. This is because, when K is not a multiple of b, any set of parameters (ϵ, r, K, b) which achieves equality in the bound derived in Theorem 4 *cannot* be converted into an (n, T, d, q) code whose values of both n (the number of symbols in a codeword) and d (the minimum Hamming distance) are integers.[5] Hence, it is impossible to construct an *AU* with the set of parameters as shown in the following example.

Let $K = 3$, $b = 2$ and $\epsilon = 1/2$, the *AU*-bound defined in Theorem 4 gives $2^r \geq \epsilon^{-1}\left(K/b - 1\right) = 1$, which is not tight because it is impossible to construct such an *AU* with a single key. The combinatorial *AU*-bound, on the other hand, gives $2^r \geq \epsilon^{-1}\lfloor K/b \rfloor = 2$ corresponding to an $(\epsilon = 1/2)$-*AU* $(r = 1, K = 3, b = 2)$ (i.e. see Table 1) or an $(n = 2, T = 8, d = 1, q = 4)$ code.

Since any *AU*-bound is also a bound for error correcting codes, one might question: does the combinatorial *AU*-bound give rise to a new bound in coding theory which is tighter than Singleton bound? It is however perhaphs suprising to discover when we convert the combinatorial bound into parameters in coding, it is no better than Singleton bound as demonstrated below.

- When K is a multiple of b, the two bounds are equivalent thanks to the above analysis.
- When $K = tb + b'$ and $0 < b' < b$, where t is an integer. The combinatorial *AU*-bound is equivalent to: $t = \lfloor K/b \rfloor \leq \epsilon 2^r = n - d$, and so $T = 2^{tb+b'} \leq q^{n-d} 2^{b'} < q^{n-d+1}$.

4 The significance of the threshold value of ϵ

We are going to compare the combinatorial *AU*-bound with other bounds for not only *AU* but also *AXU* and *ASU* to understand the accuracy and significance of our result. This comparative

[5] Assume $K = tb + b'$ where $0 < b' < b$ and t is an integer. If equality in the bound derived in Theorem 4 is achieved, then $2^r\epsilon = t - 1 + b'/b$. Using the equivalence between *AU* and error-correcting code in Theorem 3, we further have $n - d = 2^r\epsilon = t - 1 + b'/b$. Since b'/b is not integer, n and d cannot be integer at the same time.

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8
k_1	1	2	3	4	1	2	3	4
k_2	2	3	4	1	3	4	1	2

Table 1: A construction of an ($\epsilon = 1/2$)-*AU* (1, 3, 2), in which there are $2^r = 2$ hash keys $\{k_1, k_2\}$ and $2^K = 8$ input messages $\{m_1, \ldots, m_8\}$. The range of a hash output is $[1,2^b = 4]$.

analysis also uncovers the importance of the value $\epsilon = (1 + \frac{b}{K-b})2^{-b}$ which represents a *threshold* in the behaviour of bounds, and therefore, for the first time, quantifies the *Wegman-Carter* effect.

We introduce a new *AXU*-bound derived from the *ASU*-bound of Kabatianskii et al. [15] and a connection between *AXU* and *ASU* due to Wegman and Carter [35].

4.1 Comparison between the combinatorial and other *AU*-bounds

Stinson's *AU*-bound, which can be derived from Plotkin bound in coding theory [29], is as follows: $2^r \geq \frac{2^K(2^b-1)}{2^K(\epsilon 2^b-1)+2^{2b}(1-\epsilon)}$. When $\epsilon = 2^{-b}$, this is much tighter than ours for then it gives $r \geq K - b$, which means that the key bitlength grows at least linearly with the message bitlength. In contrast, as we increase ϵ to 2^{1-b} then setting $r = b$ satisfies the bound, i.e. the key needs be no longer than the bitlength of the hash.

To understand the dramatic collapse, we present a different way to interpret the formula when $\epsilon = \gamma 2^{-b} > 2^{-b}$, which is the same as $\gamma > 1$.

$$2^r \geq \frac{2^K(2^b-1)}{2^K(\gamma-1)+2^{2b}(1-\gamma 2^{-b})} = \frac{2^b-1}{(\gamma-1)+2^{2b-K}(1-\gamma 2^{-b})}$$

Note that since both terms in the denominator of the right-hand form are positive for $\gamma > 1$, with the second one converging to 0 as K increases, no matter how big K gets it can never prove a stronger lower abound on r than

$$r > \log \frac{2^b}{\gamma-1} = b + \log \frac{1}{\gamma-1}$$

In other words, while the combinatorial bound grows in proportion to $log\, K$, this bound is essentially constant as K increases. Hence there comes a point as K and ϵ increase where Stinson's bound becomes weaker than the combinatorial one. In order to locate that point, we find the value of ϵ above which ours is greater than Stinson's. To simplify the calculation, we will round up our combinatorial *AU*-bound to ($2^r \geq \frac{K}{\epsilon b}$). This gives a very good approximation to the crucial value.

$$\begin{aligned} \frac{K}{\epsilon b} &> \frac{2^K\left(2^b-1\right)}{2^K\left(\epsilon 2^b-1\right)+2^{2b}(1-\epsilon)} \\ \epsilon &> \frac{K2^K - K2^{2b}}{K2^{K+b} - K2^{2b} - b2^{K+b} + b2^K} \end{aligned}$$

Since $2^{2b} \ll 2^K \ll 2^{K+b}$, the above can be approximated as follows:

$$\epsilon > \frac{K2^K}{K2^{K+b} - b2^{K+b}} = \frac{K}{(K-b)2^b} = \left(1 + \frac{b}{K-b}\right)2^{-b}$$

From now on, we will refer to this value of ϵ as the *threshold* value. The result demonstrates that Stinson's AU-bound can only be tight within a very short range of ϵ, since K is always assumed to be significantly bigger than b. Moreover, the difference between the threshold value and 2^{-b}, i.e. $\frac{b}{(K-b)2^b}$, can be made as small positively as we want. This implies that if ϵ exceeds 2^{-b} by an arbitrarily small positive value, then the message bitlength grows at most exponentially with the key bitlength as demonstrated in the combinatorial AU-bound, but if $\epsilon = 2^{-b}$ it will grow at most linearly as shown in Stinson's AU-bound.

While the same asymptotic behaviour has also been derived from a relation between ASU and codes correcting independent errors by Johansson et al. [14, 15], it is not clear to us how we can derive the same threshold value of ϵ from the strategy used by Johansson et al. As a consequence, our approach of deriving the result quantitatively demonstrates three further important points:

- If we fix the bitlengths of an input message and a hash output, then Stinson's AU-bound is still useful when $2^{-b} < \epsilon < \left(1 + \frac{b}{K-b}\right) 2^{-b}$, more information can be found in Table 2.
- Given any value of ϵ which exceeds 2^{-b} by an arbitrarily small positive value, we can determine the threshold of input messages' bitlength ($K \geq b + \frac{b}{2^b\epsilon - 1}$) above which the message bitlength can apparently start to grow exponentially with the key bitlength, i.e. the combinatorial AU-bound gives a better estimate than Stinson's AU-bound.
- The threshold value of ϵ, perhaps surprisingly, has the same theoretical importance when we visit different ASU- and AXU-bounds in Appendix B. See Table 2 for more information.

4.2 Comparison between the combinatorial AU-bound and known ASU- and AXU-bounds

Since ASU is more restrictive than AU, intuitively we would expect that the number of bits required for the key in AU should be smaller than in ASU w.r.t the same set of parameters (ϵ, K, b). This analysis is reflected by the following comparisons:

- When $\epsilon = 2^{-b}$, Stinson's AU-bound ($r \geq K - b$) is smaller than Stinson's ASU-bound [28, 29][6] ($r > K + b - 1$) by $2b - 1$ bits. But when $\epsilon > 2^{-b}$, the gap gets closer as follows:
- Our AU-bound is smaller than Kabatianskii's ASU-bound [15], $r \geq b + \log(\epsilon^{-1}\lfloor K/b \rfloor)$, by at least b bits.[7]
- The difference between our AU-bound and Gemmell-Naor's ASU-bound,[8] $r \geq \log K + 2\log\epsilon^{-1} - \log\log\epsilon^{-1}$, gets very near to b when $\theta \ll b$: $\log\epsilon^{-1} + \log\frac{b}{\log\epsilon^{-1}} = b - \theta + \log\frac{b}{b-\theta}$

Coincidentally, it is known that if there exists an ϵ-AXU (r, K, b) then it can be used to construct an ϵ-ASU $(r + b, K, b)$, thanks to the work of Wegman and Carter [35], i.e. see Theorem 5.

[6] Stinson's ASU-bound can be derived from the second Johnson bound for constant weight binary codes [29].

[7] Kabatianskii's ASU-bound, which is derived from Johnson bound in Theorem 15 of [15], is valid when $K/b < \sqrt{2^{r-b+1}(1-2^{-b})} - 1/2$.

[8] We note that the bound was reported in the paper of Gemmell and Noar [11] (Section 5.1). However, it was noted there that the bound was actually introduced by Noga Alon through private communication.

Theorem 5 [35, 10]. *Let $H = \{h_k() \mid k \in [0, 2^r)\}$ be an ϵ-AXU (r, K, b),*[9] *then $\hat{H} = \{\hat{h}_{k,s}() \mid k \in [0, 2^r),\ s \in [0, 2^b),\ and\ \hat{h}_{k,s}() = h_k() \oplus s\}$ is an ϵ-ASU $(r + b, K, b)$.*

Proof of this theorem can be found in Appendix A. Applying Theorem 5 to Kabatianskii's *ASU*-bound, $r \geq b + \log(\epsilon^{-1}\lfloor K/b \rfloor)$, we can derive its *AXU*-variant as in the following theorem.

Theorem 6 *For any ϵ-AXU (r, K, b): $r \geq \log(\epsilon^{-1}\lfloor K/b \rfloor)$, provided*[10] $K/b < \sqrt{2^{r+1}(1 - 2^{-b})} - 1/2$

The theorem shows that *AU*-bound is strictly shorter than *AXU*-bound for some set of parameters (ϵ, K, b), i.e. when K is a multiple of b. This argument is consistent with the formal definitions, since *AXU* is a stronger definition of *AU*.

For example, when we set $\epsilon = 2^{-b}$, Stinson's *AU*-bound yields $K - b$ bits compared to K, derived from Stinson's *AXU*-bound ($2^r \geq \frac{2^K(2^b-1)}{2^b\epsilon(2^K-1)+2^b-2^K}$) [29].[11] We will see again that this comparative analysis is justified for larger values of ϵ when we visit constructions based on *polynomial hashing* over finite fields in Section 5.

5 The optimality of *polynomial hashing* as an *AU*

Polynomial hashing over finite fields was independently introduced by Boer [7], Johansson et al. [14], and Taylor [32]. To the best of our knowledge, polynomial hashing as an authentication code (*ASU*) has only been proved to be *asymptotically optimal* by Johansson et al. [14].[12]

Extending this result, we will show a different version of polynomial hashing which is designed as an *AU* is *optimal*, because it meets the combinatorial *AU*-bound with equality.

Fix some positive integer t. Let the set of all messages be $\{m = \langle m_1, \ldots, m_t \rangle; m_i \in \mathbb{F}_q\}$, here $b = \log q$ and the message bitlength is $K = tb = t \log q$.

In the first version of polynomial hashing as an *AU*, each message m will form a polynomial $m(x)$ of degree less than t over $\mathbb{F}_q$. For any key $k \in \mathbb{F}_q$, the hash of the message m with respect to the key k is equivalent to $m(k)$ over $\mathbb{F}_q$.

$$h_k(m) = m(k) = m_1 + m_2 k + m_3 k^2 + \cdots + m_t k^{t-1}$$

If we fix two different messages A and $B = A + m$, then a hash collision is equivalent to: $0 = h_k(A) + h_k(B) = A(k) + B(k) = m(k)$. Since the polynomial $m(k)$ is of degree up to $(t-1)$, we have $\epsilon = (t-1)q^{-1} = (K/b - 1)2^{-r}$, and so $r = \log\left(\epsilon^{-1}(K/b - 1)\right)$. The equality in the combinatorial *AU*-bound implies optimality of polynomial hashing as an *AU* for any $K/b = t \in [2, q]$.

Unfortunately, the construction above is not an *AXU* because if we set $\omega = A_1 + B_1$ and for all $i \in (1, t]$: $A_i = B_i = 0$, then for all $k \in \mathbb{F}_q$ we have $h_k(A) + h_k(B) = A_1 + B_1 = \omega$. In constrast,

[9] We note that the *AXU* in this theorem does not need to be uniformly distributed as argued by Etzel et al. [10].

[10] As pointed out in footnote 7 and [15], there is a condition for the validity of Kabatianskii's *ASU*-bound, and therefore the same condition should apply to the *AXU*-variant of Kabatianskii's bound.

[11] Stinson's *AXU*-bound is derived from the second Johnson bound for constant weight binary codes.

[12] Since Kabatianskii's *ASU*-bound has only been proved to be valid in a partial range of parameters (see footnote 7 or [15]), the optimality of polynomial hashing as an *ASU* remains to be proved. On the other hand, polynomial hashing as an *ASU* is known to be asymptotically optimal due to Johansson et al. [14], i.e. the authors used polynomials to construct an $(\epsilon = \frac{t}{2^b})$-$ASU(r = 2b, K = tb, b)$, where t is an integer, and they proved that for t fixed and $b \to \infty$ then $2^K = 2^{tb}$ is *asymptotically* the upper bound on the number of messages that can be securely authenticated.

letting message m form a polynomial of degree up to t can get around this problem completely:

$$h_k(m) = m(k) = m_1k + m_2k^2 + \cdots + m_tk^t$$

Similar calculations show that this is an $(\epsilon = t/q)$-AXU, which meets the AXU-variant of Kabatianskii's bound with equality: $\log(\epsilon^{-1}\lfloor K/b \rfloor) = \log q = r$. This, therefore, justifies the difference between our combinatorial AU-bound and the AXU-variant of Kabatianskii's bound, i.e. when K is a multiple of b, AXU-bound is greater than AU-bound w.r.t the same set of parameters (ϵ, K, b).

Using Theorem 5 and the above construction, we can build an $(\epsilon = \frac{t}{2^b})$-$ASU$ $(r = 2b, K = tb, b)$, which was originally introduced by Johansson et al. [14]. For any pair of keys $(k, s) \in \mathbb{F}_q^2$:

$$h_{k,s}(m) = s + m(k) = s + m_1k + m_2k^2 + \cdots + m_tk^t$$

This meets Kabatianskii's ASU-bound $(r \geq b + \log(\epsilon^{-1}\lfloor K/b \rfloor))$ with equality. However, Kabatianskii's ASU-bound and its AXU-variant have only been proved to be valid when $t < \sqrt{2^{b+1}(1-2^{-b})} - 1/2 = \sqrt{2q(1-1/q)} - 1/2$, and so the second and third polynomial hashings as AXU and ASU are only optimal under the condition as was also pointed out by Kabatianskii et al. [15].

6 Conclusions and future research

In this paper, we have demonstrated that the use of the connection between universal hash functions and error-correcting codes does not always give tight bounds for universal hashes. This work will open the way for re-examming existing bounds for universal hashes which have been derived from theoretical bounds of error-correcting codes (ECC-bounds) [29, 30, 11, 15] or other combinatorial objects such as difference matrices [29], orthogonal or perpendicular arrays [18, 29, 30, 31], and balanced incomplete block designs [18, 24, 28, 29, 31].

Intuitively, there exist subclasses of some universal hashes which cannot be transformed into equivalent codes that achieve equality in the ECC-bounds from which the corresponding AU or ASU-bounds are derived. Under the circumstance, equality in the AU or ASU-bounds are not achievable, and therefore combinatorial analysis might produce better bounds, for example, the combinatorial AU-bound introduced in Section 3.

In addition, we have also quantified the (asymptotic) Wegman-Carter effect using an important value of the hash collision probability ϵ that represents a threshold in behaviours of bounds for AU, AXU, and ASU; the behaviour is summarised in Table 2.

References

[1] *Bibliography on Authentication Codes.* (Up to 1998) Maintained by D.R. Stinson and R. Wei. See: `http://www.cacr.math.uwaterloo.ca/ dstinson/acbib.html`

[2] S. Bakhtiari, R. Safavi-Naini and J. Pieprzyk. *A message authentication code based on latin squares.* Australasian Conference on Information Security and Privacy, ACISP 1997, LNCS vol. 1270, 194-203.

[3] D.J. Bernstein. *Stronger security bounds for Wegman-Carter-Shoup authenticators.* Advances in Cryptology, EUROCRYPT 2005, LNCS vol. 3497, 164-180.

	$\epsilon < \left(1 + \frac{b}{K-b}\right) 2^{-b}$	$\epsilon > \left(1 + \frac{b}{K-b}\right) 2^{-b}$
ϵ-*AU*	Stinson's bound [28, 29] $\log \left(\frac{2^K (2^b - 1)}{2^K (\epsilon 2^b - 1) + 2^{2b}(1-\epsilon)} \right)$	K is a multiple of b *New*, Theorems 1 and 4 $\log \frac{K-b}{\epsilon b}$ K is *not* a multiple of b *New*, Theorem 1 $\log \left(\epsilon^{-1} \lfloor K/b \rfloor \right)$
ϵ-*AXU*	Stinson's bound [29] $\log \left(\frac{2^K (2^b - 1)}{2^b \epsilon (2^K - 1) + 2^b - 2^K} \right)$	*AXU*-variant of Kabatianskii's bound *New*, Theorem 6, Section 4.1 $\log \left(\epsilon^{-1} \lfloor K/b \rfloor \right)$ (provided $K/b < \sqrt{2^{r+1}(1 - 2^{-b})} - 1/2$)
ϵ-*ASU*	Stinson's bound [28, 29] $\log \left(1 + \frac{2^K (2^b - 1)^2}{2^b \epsilon (2^K - 1) + 2^b - 2^K} \right)$	Kabatianskii's bound [15] $b + \log \left(\epsilon^{-1} \lfloor K/b \rfloor \right)$ (provided $K/b < \sqrt{2^{r-b+1}(1 - 2^{-b})} - 1/2$) Gemmell and Noar's bound [11] $\log K + 2 \log \epsilon^{-1} - \log \log \epsilon^{-1}$

Table 2: Classification of different lower bounds on the key length r for *AU*, *AXU* and *ASU* with respect to the threshold value of ϵ: $\left(1 + \frac{b}{K-b}\right) 2^{-b}$.

[4] D.J. Bernstein. *The Poly1305-AES message-authentication code.* Fast software encryption, FSE 2005, LNCS vol. 3557, pp. 32-49.

[5] J. Bierbrauer, T. Johansson, G.A. Kabatianskii, and B.J.M. Smeets. *On Families of Hash Functions via Geometric Codes and Concatenation.* Advances in Cryptology, CRYPTO 1993, LNCS vol. 773, 331-342.

[6] J. Bierbrauer. *Introduction to Coding Theory.* (pages 240-241). Published by CRC Press, 2004. ISBN 1584884215, 9781584884217.

[7] B. den Boer. *A simple and key-economical unconditional authentication scheme.* Journal of Computer Security 2 (1993), 65-71.

[8] J.L. Carter and M.N. Wegman. *Universal Classes of Hash Functions.* Journal of Computer and System Sciences, 18 (1979), 143-154.

[9] S.J. Creese, M.H. Goldsmith, A.W. Roscoe, and I. Zakiuddin. *The attacker in ubiquitous computing environments: Formalising the threat model.* Workshop on Formal Aspects in Security and Trust, Pisa, Italy, 2003.

[10] M. Etzel, S. Patel, and Z. Ramzan. *SQUARE HASH : Fast message authentication via optimized universal hash functions.* Advances in Cryptology, CRYPTO 99, LNCS vol. 1666, 234-251.

[11] P. Gemmell and M. Naor. *Codes for Interactive Authentication.* Advances in Cryptology, CRYPTO 93, LNCS vol. 773, 355-367.

[12] H. Handschuh and B. Preneel. *Key-Recovery Attacks on Universal Hash Function Based MAC Algorithms.* Advances in Cryptology, CRYPTO 2008, LNCS vol. 5157, 144-161.

[13] S.-H. Heng and K. Kurosawa. *Square hash with a small key size.* Australasian Conference on Information Security and Privacy, ACISP 2003, LNCS vol. 2727, 522-531.

[14] T. Johansson, G.A. Kabatianskii, and B. Smeets. *On the relation between A-Codes and Codes correcting independent errors.* Advances in Cryptology, EUROCRYPT 1993, LNCS vol. 765, 1-11.

[15] G.A. Kabatianskii, B. Smeets, and T. Johansson. *On the cardinality of systematic authentication codes via error-correcting codes.* IEEE Transactions on Information Theory, IT-42 (1996), 566-578.

[16] H. Krawczyk. *LFSR-based Hashing and Authentication.* Advances in Cryptology, CRYPTO 1994, LNCS vol. 839, 129-139.

[17] H. Krawczyk. *New Hash Functions For Message Authentication.* Advances in Cryptology, EUROCRYPT 1995, LNCS vol. 921, 301-310.

[18] K. Kurosawa, K. Okada, H. Saido, and D.R. Stinson. *New combinatorial bounds for authentication codes and key predistribution schemes.* Designs, Codes and Cryptography, 15 (1998), 87-100.

[19] K. Kurosawa, S. Obana. *Combinatorial Bounds on Authentication Codes with Arbitration.* Design, Codes Cryptography 22 (3): 265-281 (2001).

[20] S. Laur and S. Pasini. *SAS-Based Group Authentication and Key Agreement Protocols.* Public Key Cryptography, PKC, 197-213 (2008).

[21] W. Nevelsteen and B. Preneel. *Software performance of universal hash functions.* Advances in cryptology, EUROCRYPT 1999, LNCS vol. 1592, pp. 24-41.

[22] L.H. Nguyen and A.W. Roscoe. *Authenticating ad hoc networks by comparison of short digests.* Information and Computation 206 (2008), 250-271.

[23] V.S. Pless and W. Huffman. *Handbook of Coding Theory* (Chapter 4, Sec 2.2), published by Elsevier (1998). ISBN 0444500871. Or see: `http://en.wikipedia.org/wiki/Singleton_bound`

[24] R.S. Rees and D.R. Stinson. *Combinatorial characterizations of authentication codes II.* Designs, Codes and Cryptography 7 (1996), 239-259.

[25] D.V. Sarwate. *A note on universal classes of hash functions.* Information Processing Letter, 10 (1): 41-45 (1980).

[26] V. Shoup. *On Fast and Provably Secure Message Authentication Based on Universal Hashing.* Advances in Cryptology, CRYPTO 1996, LNCS vol. 1109, 313-328.

[27] F. Stajano and R. Anderson. *The resurrecting duckling: Security issues for ad-hoc wireless networks.* Security Protocols 1999, LNCS vol. 1976, 172-194.

[28] D.R. Stinson. *Universal Hashing and Authentication Codes.* Advances in Cryptology, CRYPTO 1991, LNCS vol. 576, 74-85.

[29] D.R. Stinson. *On the Connections Between Universal Hashing, Combinatorial Designs and Error-Correcting Codes.* Congressus Numerantium, vol. 114 (1996), 7-27.

[30] D.R. Stinson. *The combinatorics of authentication and secrecy codes.* Journal of Cryptology 2 (1990), 23-49.

[31] D.R. Stinson. *Combinatorial techniques for universal hashing.* Journal of Computer and System Sciences 48 (1994), 337-346.

[32] R. Taylor. *An Integrity Check Value Algorithm for Stream Ciphers.* Advances in Cryptology, CRYPTO 1993. LNCS vol. 773, Springer-Verlag, pp. 40-48, 1994.

[33] J. Valkonen, N. Asokan, and K. Nyberg. *Ad Hoc Security Associations for Groups.* European Workshop on Security and Privacy in Ad hoc and Sensor Networks, 2006. LNCS vol. 4357, 150-164.

[34] S. Vaudenay. *Secure Communications over Insecure Channels Based on Short Authenticated Strings.* Advances in Cryptology, CRYPTO 2005, LNCS vol. 3621, 309-326.

[35] M.N. Wegman and J.L. Carter. *New Hash Functions and Their Use in Authentication and Set Equality.* Journal of Computer and System Sciences, 22 (1981), 265-279.

A Proof of a connection between AXU and ASU: Theorem 5

Proof For any message m and hash output y, we have

$$P_I = \mathbf{Pr}_{k,s}\left[\hat{h}_{k,s}(m) = y\right] = \mathbf{Pr}_{k,s}[h_k(m) \oplus s = y]$$

For any value of k, s is uniquely determined by $s = h_k(m) \oplus y$, and thus $P_I = \frac{2^r}{2^{r+b}} = 2^{-b}$.

For every pair of distinct messages $(m, \hat{m})$ and for every pair of hash outputs $(y, \hat{y})$, we have

$$P_S = \mathbf{Pr}_{k,s}\left[\hat{h}_{k,s}(m) = y, \hat{h}_{k,s}(\hat{m}) = \hat{y}\right] = \mathbf{Pr}_{k,s}\left[h_k(m) \oplus s = y, h_k(m) \oplus h_k(\hat{m}) = y \oplus \hat{y}\right]$$

For any value of k, s is uniquely determined by $s = h_k(m) \oplus y$. Since $h_k()$ is an ϵ-AXU there are at most $\epsilon 2^r$ different keys that satisfy $h_k(m) \oplus h_k(\hat{m}) = y \oplus \hat{y}$, and thus $P_S \le \frac{\epsilon 2^r}{2^{r+b}} = \epsilon 2^{-b}$. ∎

B The threshold value in relation to AXU and ASU

We note that Stinson's bounds for AXU and ASU have similar forms to his *AU*-bound. Furthermore, the same similarity in form holds between Kabatianskii's *ASU*-bound, the AXU-variant of Kabatianskii's bound (see Theorem 6) and our combinatorial *AU*-bound. Owing to this symmetry, we assert that the threshold value of ϵ has the same significance in the relationships between the two versions of *ASU*-bound, and of AXU-bound respectively.

The following calculation locates the value of ϵ above which Kabatianskii's *ASU*-bound becomes better than Stinson's *ASU*-bound.[13]

$$\begin{aligned} \frac{K2^b}{\epsilon b} &\ge \frac{2^K(2^b-1)^2}{2^b\epsilon(2^K-1)+2^b-2^K} \\ \epsilon &\ge \frac{K2^{b+K} - K2^{2b}}{K2^{2b+K} - K2^{2b} - b2^{2b+K} + b2^{b+K+1} - b2^K} \end{aligned}$$

Since $2^{2b} \ll 2^K \ll 2^{K+b}$ the above can be approximated as follows:

$$\epsilon > \frac{K2^{b+K}}{K2^{2b+K} - b2^{2b+K}} = \frac{K}{(K-b)2^b} = \left(1 + \frac{b}{K-b}\right)2^{-b}$$

A similar calculation also leads us to conclude that Stinson's AXU-bound is overtaken by the AXU-variant of Kabatianskii's bound at the threshold value of ϵ.

A summary of the relation between all these different bounds for *AU*, AXU and *ASU* w.r.t the threshold value of ϵ is given in Table 2.

[13]Since the constant 1 in Stinson's *ASU*-bound ($2^r \ge 1 + \frac{2^K(2^b-1)^2}{2^b\epsilon(2^K-1)+2^b-2^K}$) is very small compared to 2^r, we will ignore it in subsequent analysis to simplify the calculation. In addition, we will round up Kabatianskii's *ASU*-bound from $2^r \ge \frac{2^b}{\epsilon}\lfloor K/b \rfloor$ to $2^r \ge \frac{2^b K}{\epsilon b}$.