A Tight Bound for EMAC

Krzysztof Pietrzak*

Département d'Informatique, École Normale Supérieure, Paris
pietrzak@di.ens.fr

Abstract. We prove a new upper bound on the advantage of any adversary for distinguishing the encrypted CBC-MAC (EMAC) based on random permutations from a random function. Our proof uses techniques recently introduced in [BPR05], which again were inspired by [DGH+04]. The bound we prove is tight — in the sense that it matches the advantage of known attacks up to a constant factor — for a wide range of the parameters: let n denote the block-size, q the number of queries the adversary is allowed to make and ℓ an upper bound on the length (i.e. number of blocks) of the messages, then for $\ell \leq 2^{n/8}$ and $q \geq \ell^2$ the advantage is in the order of $q^2/2^n$ (and in particular independent of ℓ). This improves on the previous bound of $q^2\ell^{\Theta(1/\ln\ln\ell)}/2^n$ from [BPR05] and matches the trivial attack (which thus is basically optimal) where one simply asks random queries until a collision is found.

1 Introduction

Cipher Block Chaining (CBC) is a popular mode of operation for block ciphers which is used (in some variations) for encryption and message authentication, i.e. as a Message Authentication Code (MAC).

Some Definitions. The CBC function with key $\pi : \{0,1\}^n \to \{0,1\}^n$, denoted CBC_π, takes as input a message (whose length must be a multiple of n) $M = M_1 \cdots M_m \in (\{0,1\}^n)^m$ and outputs C_m which is inductively computed as

$$\mathsf{CBC}_\pi(M) = C_m \text{ where } C_0 = 0^n \text{ and } C_i = \pi(C_{i-1} \oplus M_i) \text{ for } i = 1, \ldots, m$$

The ECBC function (E for encrypted) is derived from the CBC function by additionally encrypting the output with an independent permutation[1]

$$\mathsf{ECBC}_{\pi_1,\pi_2}(M) \stackrel{\text{def}}{=} \pi_2(\mathsf{CBC}_{\pi_1}(M))$$

CBC based MACs. The CBC and ECBC function, with the π's instantiated by a block-cipher, are popular MACs called CBC-MAC and EMAC respectively.

As for the CBC-MAC, two parties sharing a secret key $K \in \mathcal{K}$ for a block-cipher $E : \mathcal{K} \times \{0,1\}^n \to \{0,1\}^n$ can authenticate their communication by

* Part of this work is supported by the Commission of the European Communities through the IST program under contract IST-2002-507932 ECRYPT.

[1] The ECBC function must not be confused with the ECBC-MAC from [BR00].

sending, together with their message M, the authentication tag $\mathsf{CBC}_{E(K,.)}(M)$.

On the Security of CBC based MACs. The CBC-MAC as just described is well known to be completely insecure in general,[2] but has been proven secure (under the assumption that the underlying block-cipher is a secure pseudorandom permutation) under the restriction that all messages have the same length in [BKR00], which then has been relaxed to the condition that no message is the prefix of another [PR00]. This means that the CBC-MAC can be safely used for messages of different length, if some prefix free encoding is applied.

The EMAC is a popular variant of the CBC-MAC which was developed by the RACE project [BP95], unlike the "plain" CBC-MAC it is secure without any restriction on the message space [PR00]. The EMAC, along with the UMAC, TTMAC and HMAC, is one of the message authentication codes recommended by NESSIE [NES].

The Model. As nowadays usual, we analyse the security of the construction we are interested in (which is $\mathsf{ECBC}_{\pi_1,\pi_2}$) in a setting where the underlying primitive (here π_1, π_2) are realized by their ideal functionality (here uniformly random permutations), thus separating the analysis of the security of the construction from the security of the underlying primitive.[3] More precisely, we prove an upper bound on $\mathbf{Adv}_{\mathsf{ECBC}}(q, n, \ell)$, by which we denote probability of any adversary making q queries of length at most ℓ blocks, in (existentially) forging $\mathsf{ECBC}_{\pi_1,\pi_2}$.

Following [BR00,BPR05], we view the EMAC as a Carter-Wegman MAC [CW79]. This reduces the task of bounding $\mathbf{Adv}_{\mathsf{ECBC}}(q, n, \ell)$ to the task of bounding the probability that there is a collision amongst the CBC-MACs of q messages of length at most ℓ blocks, we denote this probability by $\mathbf{CP}_{q,n,\ell}$ (see (5)). In practice one would instantiate the π_i's by a block-cipher (and not with uniform random permutations). If this block-cipher is secure in the sense of being a good pseudorandom permutation, then the security of the EMAC is basically $\mathbf{CP}_{q,n,\ell}$, thus proving a good bound on this probability translates into improved security guarantees for the EMAC.

Known Lower Bounds. There is a trivial lower bound $\mathbf{CP}_{q,n,\ell} \in \Omega(q^2/2^n)$ for any q, n and $\ell > 1$ as by the birthday bound we can find a collision with probability $\Omega(q^2/2^n)$ for any input shrinking function by asking random queries.[4]

For $q = 2$ [BPR05] show a lower bound of $\mathbf{CP}_{2,n,\ell} \in \Omega(d(\ell)/2^n)$ where $d(\ell) \stackrel{\text{def}}{=} max_{t \leq \ell} |\{x; 1 \leq x \leq 2^n, x|t\}|$ denotes the maximum number of divisors between 1 and 2^n of any number $\leq \ell$. It is known (Theorem 317 in [HW80]) that $D(\ell) \stackrel{\text{def}}{=} max_{t \leq \ell} |\{x; x|t\}| \in \ell^{\Theta(1/\ln \ln \ell)}$, so the same bound applies for $d(\ell)$ if $\ell \leq 2^n$ as then $d(\ell) = D(\ell)$.

---

[2] In particular, it is not existentially unforgeable as shown by the following simple attack: for any $X \in \{0,1\}^n$, request the MAC $C = CBC_\pi(X) = \pi(X)$, and output a message $X \| X \oplus C$ with tag C. This is a successful forgery as $CBC_\pi(X \| X \oplus C) = \pi(\pi(X) \oplus X \oplus C) = \pi(X) = C$.

[3] See e.g. [Mau02] for more detailed discussion of this concept.

[4] For $\ell = 1$ we have $\mathbf{CP}_{q,n,\ell} = 0$ as a permutation does not have collisions.

Where	Upper Bound, $O(.)$ of	Range restriction	Other restrictions
[PR00]	$\ell^2 q^2/2^n$	-	-
[BPR05]	$d(\ell)q^2/2^n$	$\ell \in O(2^{n/4})$	-
[DGH$^+$04]	$q^2/2^n$	$\ell \in O(2^{n/3})$	Equal length messages
Here	$q^2/2^n$	$\ell \in O(2^{n/8}), q \in \Omega(\ell^2)$	-

Where	Lower Bound, $\Omega(.)$ of
Folklore (birthday bound)	$q^2/2^n$
[BPR05]	$d(\ell)q/2^n$

Fig. 1. Upper and lower bounds for $\mathbf{CP}_{q,n,\ell}$ (which then imply basically the same bounds for $\mathbf{Adv}_{\mathsf{ECBC}}(q, n, \ell)$).

KNOWN UPPER BOUNDS. Until now the best known upper bound was a $\mathbf{CP}_{n,q,\ell} \in O(d(\ell)q^2/2^n)$ (for $\ell \leq 2^{n/4}$) due to Bellare et al. [BPR05], this bound improved on the $O(\ell^2 q^2/2^n)$ bound of Petrank and Rackoff [PR00]).

TIGHT BOUND FOR EQUAL LENGTH. Dodis et al. [DGH$^+$04] investigated a restricted case where the messages have same length (which is uninteresting for the EMAC construction, but this was not their goal), they state a tight collision probability of $\mathbf{CP}_{2,n,\ell} \in O(q^2/2^n)$ (for $\ell \leq 2^{n/3}$) for the CBC-MAC of two messages, which immediately gives an optimal $\mathbf{CP}_{q,n,\ell} \in \Theta(q^2/2^n)$ bound for the collision probability of q *equal length* messages.

OUR CONTRIBUTION. In this paper we prove the optimal bound $\mathbf{CP}_{q,n,\ell} \in \Theta(q^2/2^n)$ for $q \geq \ell^2$ and $\ell \leq 2^{n/8}$. So for this range the security of ECBC (and thus the EMAC) matches the security of an ideal MAC (i.e. the birthday bound) up to constant factors.

THE TECHNIQUE FROM [BPR05]. Both, the "classical" $O(q^2\ell^2/2^n)$ [PR00] and the $O(d(\ell)q^2/2^n)$ upper bound [BPR05] are achieved by first proving an upper bound on $\mathbf{CP}_{2,n,\ell}$, the collision probability of two messages, and then applying the union bound

$$\mathbf{CP}_{q,n,\ell} \leq \frac{q(q-1)}{2} \cdot \mathbf{CP}_{2,n,\ell} \tag{1}$$

to get a bound for $\mathbf{CP}_{q,n,\ell}$. In particular [BPR05] prove that

$$\mathbf{CP}_{2,n,\ell} \leq 2d(\ell)/2^n + 64\ell^4/2^{2n}. \tag{2}$$

This bound is tight up to the higher order term and a factor 2:

$$\mathbf{CP}_{2,n,\ell} \geq d(\ell)/2^n \tag{3}$$

The proof of (2) uses ideas from [DGH$^+$04,Dod05] and goes roughly as follows: For any two messages M_1, M_2 and a permutation π one maps the computation of $\mathsf{CBC}_\pi(M_1)$ and $\mathsf{CBC}_\pi(M_2)$ to a graph (called structure graph) consisting of two paths associated with the message M_1 and M_2 respectively. In this graph the vertices correspond to the outputs of π during this computation.

Each such graph contains zero or more *accidents*, by which one denotes the "unexpected" collisions in the graph. The main technical lemma (Lemma 2 in this paper) now states that the probability (over the choice of π) that some particular structure graph G will appear is exponentially small in the number of accidents of G. From this lemma one gets that the probability that a random structure graph has at least one accident is in $O(\ell^2/2^n)$. We can now use that $\mathsf{CBC}_\pi(M_1) = \mathsf{CBC}_\pi(M_2)$ implies that there must be at least one accident to get a $O(\ell^2/2^n)$ upper bound on $\mathbf{CP}_{2,n,\ell}$, and further with (1) the "classical" $\mathbf{CP}_{q,n,\ell} \in O(q^2\ell^2/2^n)$ bound. But this bound is not tight as having an accident is only necessary, but not sufficient to have $\mathsf{CBC}_\pi(M_1) = \mathsf{CBC}_\pi(M_2)$. By more carefully upper bounding the number of graphs for which $\mathsf{CBC}_\pi(M_1) = \mathsf{CBC}_\pi(M_2)$ by $O(d(\ell)/2^n + \ell^4/2^{2n})$ one gets the (2) bound. Here the $O(d(\ell)/2^n)$ term bounds the graphs which have *exactly* one accident and $\mathsf{CBC}_\pi(M_1) = \mathsf{CBC}_\pi(M_2)$, whereas all graphs with two or more accidents are "generously" bounded by the "higher order" term $O(\ell^4/2^{2n})$, whech will be dominated by the leading $d(\ell)/2^n$ while ℓ is not too large, $\ell \in O(2^{n/4})$ is small enough.

Unfortunately the bound (3) implies that bounding the collision probability for two messages and then using (1) one cannot prove $\mathbf{CP}_{q,n,\ell} \in o(d(\ell)q^2/2^n)$.

Proof Idea. The obvious idea to overcome this barrier is to upper bound the number of structure graphs built by many (and not just two) messages. We prove a lemma (Lemma 4) which states that the number of structure graphs built from any k messages of length at most ℓ blocks, having exactly one accident and a collision on the output for some pair of messages, is at most $k(k+\ell^2)$, this then gives the claimed $\mathbf{CP}_{q,n,\ell} \in O(q^2/\ell^2)$ bound. Unfortunately now the graph is so big (i.e. $q\ell$ vertices) that the higher order term which bounds the cases where we have two or more accidents is in the order $q^4\ell^4/2^{2n}$ (so unless we assume some bound $o(2^{n/2})$ on q, we only achieve a tight $O(q^2/2^n)$ for constant ℓ, but this is already achieved by the classical $q^2\ell^2/2^n$ of [PR00]).

Fortunately one can get out of this apparent cul-de-sac using an approach "between" the one just described and the one given by (1). The q messages are divided into q/ℓ^2 sets of size $r = \ell^2$. Now, if there's a collision, then this collision occurs in the union of two (or maybe just one) such sets. For such a union of two sets (of size $2r$) we can now upper bound the probability that there's a collision amongst any two of the $2r$ messages by $O(r^2/2^n)$ as the sets are sufficiently large (such that applying the before-mentioned Lemma 4 gives a $2r(2r+\ell^2) = \Theta(r^2)$ upper bound on the number of structure graphs), but still small enough for the higher order term to be ignored for a reasonable range of ℓ. Finally we get our $\mathbf{CP}_{q,n,\ell} \in O(q^2/2^n)$ bound (for $\ell \leq 2^{n/8}$) from the union bound applied over all pairs of sets.

About the Range. The tight upper bound $\mathbf{CP}_{q,n,\ell} \in O(q^2/2^n)$ we prove holds for $q \in \Omega(\ell^2)$ and $\ell \in O(2^{n/8})$. In the next two paragraphs we'll shortly discuss those two bounds.

LOWER BOUND ON q. The $\mathbf{CP}_{2,n,\ell} = \Theta(d(\ell)/2^n)$ bound implies (under a reasonable assumption[5]) $\mathbf{CP}_{q,n,\ell} = \Omega(d(\ell)q/2^n)$. Thus $\mathbf{CP}_{q,n,\ell} \in O(q^2/2^n)$ can only hold if we have a lower bound for q of at least $\Omega(d(\ell))$, the bound we actually require is $q \in \Omega(\ell^2)$.[6] But this lower bound on q is not really relevant as long as there's a upper bound $\ll 2^{n/2}$ on ℓ, as it only means that we don't match the birthday bound $O(q^2/2^n)$ for a range of parameters, where the collision probability given by the classical $q^2\ell^2/2^n$ bound is extremely small anyway.
UPPER BOUND ON ℓ. Wlog. we can assume an upper bound $\ell \leq 2^n!$ as considering longer messages makes no sense: note that every $x, 1 \leq x \leq 2^n$ divides $2^n!$ and thus $\mathbf{CP}_{2,n,2^n!} \geq d(2^n!)/2^n = 1$, i.e. we can find a collision with probability one with only two queries.[7] This (doubly exponential) bound is far from the $\ell \leq 2^{n/8}$ we require, and can probably be relaxed already with the techniques used in this paper. One possibility would be via a better counting argument, which means improving on Lemma 4 from this paper (in particular, Claim 2 from the proof of this lemma seems quite loose). Lowering the $O(q(q+\ell^2))$ bound on the number of graphs given by the lemma to $q(q+o(\ell))$ would already allow a range of $\ell \leq 2^{n/(4+o(1))}$. Further, counting graphs with more than just one (but still constantly many) accidents could have the potential to get the bound to $\ell \leq 2^{n/(2+\epsilon)}$ for any $\epsilon > 0$. Such a bound might still be far from the necessary one, but would be sufficient for any practical application as a length of $2^{n/2}$ is quite big already for small block lengths (say $n = 128$ which is the smallest block-length provided by AES).

2 Definitions and the Main Technical Lemma

NOTATION If x is a string then $|x|$ denotes its length. We let $B_n \stackrel{\text{def}}{=} \{0,1\}^n$. If $X \subseteq \{0,1\}^*$ then $X^{\leq m}$ denotes the set of all non-empty strings formed by concatenating m or fewer strings from X. If S is a set equipped with some probability distribution then $s \stackrel{\$}{\leftarrow} S$ denotes the operation of picking s from S according to this distribution. If no distribution is explicitly specified, it is understood to be uniform. We denote by $\mathrm{Perm}(n)$ the set of all permutations over $\{0,1\}^n$ and with $\mathrm{Func}(n)$ the set of all functions $\{0,1\}^* \to \{0,1\}^n$.
SECURITY. An *adversary* is a computationally unbounded, randomised oracle-algorithm which finally outputs a bit. $\mathcal{A}_{q,n,\ell}$ denotes the class of adversaries that make at most q oracle queries, each of length at most ℓ n-bit blocks. For a family of functions $F\colon B_n^* \to \{0,1\}^n$, the distinguishing advantage of $\mathcal{A}_{q,n,\ell}$ for F is

$$\mathbf{Adv}_F(q,n,\ell) = \max_{A \in \mathcal{A}_{q,n,\ell}} \{\, \mathbf{Adv}_F(A) \,\} \quad \text{where}$$

[5] We must assume that one can generate $q/2$ pairs of messages where each pair achieves the "worst case" collision probability $\Omega(d(\ell)/2^n)$, and moreover the events that any pair of messages collides are sufficiently independent.
[6] As both, the lower $d(\ell)$ and the upper ℓ^2 bound follow by rather loose arguments, the truth is probably strictly in-between, i.e. in $\omega(d(\ell))$ and $o(\ell^2)$.
[7] In fact, with $\ell = 2^n!$ we can forge a message in a no-query attack as for any $X \in B_n$ and $\pi \in \mathrm{Perm}(n)$ one has $\mathsf{CBC}_\pi(X^{2^n!}) = X$.

$$\mathbf{Adv}_F(A) = \Pr[f \xleftarrow{\$} F:\ A^f \Rightarrow 1] - \Pr[f \xleftarrow{\$} \mathrm{Func}(n):\ A^f \Rightarrow 1]$$

CBC AND ECBC. Fix $n \geq 1$. Recall that for $M = M^1 \cdots M^m \in B_n^m$ and $\pi\colon B_n \to B_n$ we defined in the introduction

$$\mathsf{CBC}_\pi(M) = C_m \text{ where } C_0 = 0^n \text{ and } C_i = \pi(C_{i-1} \oplus M_i) \text{ for } i = 1, \ldots, m$$

Let $\mathsf{CBC} = \{\mathsf{CBC}_\pi\colon \pi \in \mathrm{Perm}(n)\}$, this set of functions has the distribution induced by picking π uniformly from $\mathrm{Perm}(n)$. The encrypted CBC MAC is

$$\mathsf{ECBC}_{\pi_1,\pi_2}(M) \stackrel{\text{def}}{=} \pi_2(\mathsf{CBC}_{\pi_1}(M))$$

Let $\mathsf{ECBC} = \{\mathsf{ECBC}_{\pi_1,\pi_2}\colon \pi_1, \pi_2 \in \mathrm{Perm}(n)\}$, with the distribution induced by picking π_1, π_2 independently and uniformly at random from $\mathrm{Perm}(n)$.
COLLISIONS. For q distinct messages $M_1, \ldots, M_q \in B_n^*$ we denote by

$$\mathbf{CP}_n(M_1, \ldots, M_q) \Pr_{\pi \xleftarrow{\$} \mathrm{Perm}(n)}[\exists i, j, M_i \neq M_j : \mathsf{CBC}_\pi(M_i) = \mathsf{CBC}_\pi(M_j)]$$

the probability that the CBC-MACs (based on a uniform random permutation) of any two messages collide. The maximum collision probability for any q messages of length at most ℓ n-bit blocks is denoted by

$$\mathbf{CP}_{q,n,\ell} = \max_{M_1,\ldots,M_q \in B_n^{\leq \ell}} \mathbf{CP}_n(M_1, \ldots, M_q) \tag{4}$$

Following [BR00], we view ECBC as an instance of the Carter-Wegman paradigm [CW79]. This enables us to reduce the problem of bounding $\mathbf{Adv}_{\mathsf{ECBC}}(q, n, \ell)$ to bounding the collision probability $\mathbf{CP}_{q,n,\ell}$ as

$$\mathbf{Adv}_{\mathsf{ECBC}}(q, n, \ell) \leq \mathbf{CP}_{q,n,\ell} + q(q-1)/2^{n+1} \tag{5}$$

We prove the following bound on $\mathbf{CP}_{q,n,\ell}$.

Lemma 1. *For any $q \geq \ell^2$:* $\mathbf{CP}_{q,n,\ell} \leq 16 \cdot q^2/2^n + 128 \cdot q^2\ell^8/2^{2n}$

From this lemma and (5) we get that $\mathbf{Adv}_{\mathsf{ECBC}}(q, n, \ell) \in O(q^2/2^n)$ whenever $q \in \Omega(\ell^2)$ and $\ell \in O(2^{n/8})$, for example

Corollary 1 *For any $q \geq \ell^2$ and $\ell \leq 2^{n/8-1}$:* $\mathbf{Adv}_{\mathsf{ECBC}}(q, n, \ell) \leq 18 \cdot q^2/2^n$

3 A Graph-Based Representation of CBC

In this section we review the graph-based approach to bound collision probabilities from (the full version of) [BPR05]. In this approach the collision probability is related to the number of graphs satisfying some property.

We fix for the rest of this section a blocklength $n \geq 1$, the number of messages $t \geq 1$ and t distinct messages $\mathcal{M} \stackrel{\text{def}}{=} \{M_1, \ldots, M_t\}$, where for $1 \leq i \leq t$ we denote with $m_i \geq 1$ the length (in blocks) of the i'th message $M_i \stackrel{\text{def}}{=} M_i^1 \cdots M_i^{m_i} \in B_n^{m_i}$.

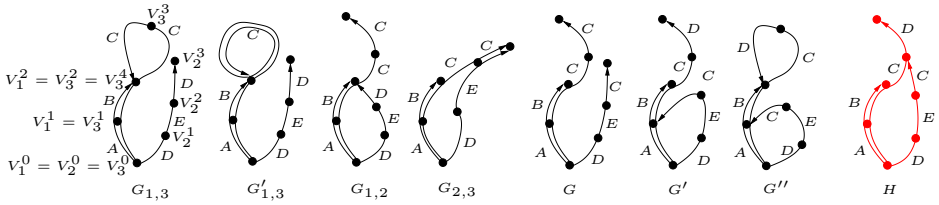

Fig. 2. $\mathcal{G}_{col}(\mathcal{M}) = \{G_{1,3}, G'_{1,3}, G_{1,2}, G_{2,3}\}$ are all structure graphs for $\mathcal{M} = \{AB, DEC, ABCC\}$ which have exactly one accident and a collision on the outputs. Further $\{G, G', G''\} \in \mathcal{G}(\mathcal{M}) \setminus \mathcal{G}_{col}(\mathcal{M})$ are valid structure graphs but not in $\mathcal{G}_{col}(\mathcal{M})$ as: G has 0 and G'' has 2 accidents. G' has exactly one accident but no collision on the outputs. H is not a structure graph as there's a vertex which has two ingoing edges, both labelled C but not being parallel.

For $1 \leq j \leq t$ let $m^j = \sum_{i=1}^{j} m_i$ be the length of the first j messages. It is convenient to set $m^0 \stackrel{\text{def}}{=} 0$ and $m \stackrel{\text{def}}{=} m^t$ to be the total length. Let $M \stackrel{\text{def}}{=} M_1 \| M_2 \| \cdots \| M_t$ denote the concatenation of all messages and M^i the i'th block of M, i.e. $M \stackrel{\text{def}}{=} M^1 \cdots M^m$.

Structure graphs. To $\mathcal{M}$ and any $\pi \in \text{Perm}(n)$ we associate the *structure graph* $G_\pi^{\mathcal{M}}$, which is a directed graph (V, E) where $V \subseteq [0, \ldots, m]$.

> The structure graph $G_\pi^{\mathcal{M}} = G = (V, E)$ is defined as follows: We set $C_0 = 0^n$ and for $i = 1, \ldots, m$ we define
>
> $$C_i = \begin{cases} \pi(C_{i-1} \oplus M^i) & \text{if } i \notin [m_0 + 1, \ldots, m_{t-1} + 1] \\ \pi(M^i) & \text{otherwise} \end{cases}$$
>
> From this C_i's we define the mapping $[.]_G : [0, \ldots, m] \to [0, \ldots, m]$ as $[i]_G = \min\{j : C_j = C_i\}$. It is convenient to define a mapping $[.]'_G$ as $[i]'_G = [i]_G$ if $i \notin [m^0, \ldots, m^{t-1}]$ and $[i]'_G = 0$ otherwise. Now the structure graph $G_\pi^{\mathcal{M}} = G = (V, E)$ is given by
>
> $$V = \{[i]_G : 0 \leq i \leq m\} \qquad E = \{([i-1]'_G, [i]_G) : 1 \leq i \leq m\}$$

From this definition it is clear that the mapping $[.]_G$ defines G uniquely and vice versa. Throughout the "i'th edge of G" refers to the edge $([i-1]'_G, [i]_G)$ (note that this not injective) and the "label" of the i'th edge is M^i.

If the C_i's are all distinct, then G is simply a star like tree with t paths leaving the root 0, the i'th path being $0 \to m^{i-1} + 1 \to \ldots \to m^{i-1} + m_i = m^i$. In general G is the graph one gets by starting with the tree just described and doing the following while possible: if there are two vertices i, j where $i \neq j$ and $C_i = C_j$ then collapse i and j into one vertex and label it $\min\{i, j\}$.

For a structure graph G we will denote the vertices on the path built by the i'th message by $V_i^0(G), V_i^1(G), \ldots, V_i^{m_i}(G)$, we call this path the i-path, we write V_i^j for $V_i^j(G)$ if G is understood (cf. Figure 2).

Let $\mathcal{G}(\mathcal{M}) = \{G_\pi : \pi \in \text{Perm}(n)\}$ denote the set of all structure graphs associated to messages $\mathcal{M}$. This set has the probability distribution induced by picking π at random from $\text{Perm}(n)$.

COLLISIONS. Suppose a structure graph $G = G_\pi^{\mathcal{M}} \in \mathcal{G}(\mathcal{M})$ is exposed edge by edge (i.e. in step i the value $[i]_G$ is shown to us). We say that G has a *collision* in step i if the edge exposed in step i points to a vertex which is already in the graph. With $\mathsf{Col}(G)$ we denote all collisions, i.e. all pairs (i, j) where in step i there was a collision which hit the vertex computed in step $j < i$:

$$\mathsf{Col}(G) = \{(i, [i]_G) : [i]_G \neq i\}$$

We distinguish between *induced collisions* IndCol and *accidents* Acc where

$$\mathsf{Col}(G) = \mathsf{Acc}(G) \cup \mathsf{IndCol}(G) \qquad \mathsf{Acc}(G) \cap \mathsf{IndCol}(G) = \emptyset$$

Informally, an induced collision in step i is a collision which is implied by the collisions in the first $i - 1$ steps, whereas an accident is a "surprising" collision.

The following lemma is the heart of the whole approach, it states that the probability that a randomly sampled structure graph will be some particular graph H is exponentially small in $\mathsf{Acc}(H)$.

Lemma 2. *Let $n \geq 1, t \geq 1, \mathcal{M} = \{M_1, \ldots, M_t\}$ where $M_i \in B_n^{m_i}$ and $m = m_1 + \ldots + m_t$. Then for any structure graph $H \in \mathcal{G}(\mathcal{M})$:*

$$\Pr[G \xleftarrow{\$} \mathcal{G}(\mathcal{M}) : \ G = H] \leq (2^n - m)^{-|\mathsf{Acc}(H)|}$$

Form this lemma we get the following bound on the probability that a random structure graph has two or more accidents:

Lemma 3. *With $\mathcal{M}, m$ as in the previous lemma*

$$\Pr[G \xleftarrow{\$} \mathcal{G}(\mathcal{M}) : \ |\mathsf{Acc}(G)| \geq 2] \leq 4m^4/2^{2n}$$

The proofs of Lemma 2 and 3 can be found in the full version of [BPR05].

SOME USEFUL FACTS. In [BPR05] accidents are formally defined to be exactly those collisions which do *not close a (even length) cycle with alternating edge directions*. It is shown that this are exactly those collisions which are "surprising" in the sense that they are not induced by the already exposed edges. We will not need to work with this formal definition of accidents here, it will be sufficient to consider the more intuitive concept of *true collisions*, which are all collisions except those where no edge is added, or equivalently, we have a true collision in some step i if in this step we add a new edge, but no new vertex (from this definition we see that in a structure graph $G = (V, E)$ the number of true collisions is $|E| - |V| + 1$). Also, it's not hard to see that if G has k accidents, then it has at least k true collisions.[8] Although the converse is not true in general, there are implications in the other direction which will be sufficient for us. In particular, it's not hard to see that the first true collision that occurs must always

[8] This follows from the definitions, recall that accidents are those collisions which do *not close a cycle with alternating edge directions*, and true collisions are those which do *not close a cycle with alternating edge directions of length 2* (as such a cycle is given by two parallel edges). So true collisions are just a subset of the accidents.

be an accident. And if we only consider structure graphs built by at most two paths, then also the second true collision is necessarily an accident (see Lemma 10 in the full version of [BPR05], fact (i) below follows from this).

For $G \in \mathcal{G}(\mathcal{M})$ and i, j, $1 \le i < j \le q$ let $G_{[i,j]}$ denote the subgraph of G built by the i-path and the j-path. We will need the following facts:

(i) If G has at most one accident, then for any i, j the $G_{[i,j]}$ has at most one true collision.
(ii) If G has exactly one accident, then G is uniquely determined by $\mathcal{M}$ and any subgraph of G which contains a true collision.

Informally, fact (ii) holds as given the single accident, we know the only "surprising" collision, and thus can deterministically extend the subgraph to G.

4 Bounding $\mathbf{CP}_{q,n,\ell}$

For $i = 1, \ldots, q$, let $M_i \in B_n^{\le \ell}$ be such that the collision probability is maximised, i.e. with $\mathcal{M} = \{M_1, \ldots, M_q\}$ we have $\mathbf{CP}_{q,n,\ell} = \mathbf{CP}(\mathcal{M})$. To bound $\mathbf{CP}(\mathcal{M})$ we now consider the random experiment where a permutation π is chosen at random and $\mathsf{CBC}_\pi(M_i)$ is computed for $i = 1, \ldots, q$. We can decide whether there was a collision $\mathsf{CBC}_\pi(M_i) = \mathsf{CBC}_\pi(M_j)$ given the structure graph $G_\pi^{\mathcal{M}}$ of this computation. Thus we see $\mathbf{CP}_{q,n,\ell}$ as the probability that $G_\pi^{\mathcal{M}}$ (for a random π) contains such a collision on the outputs of two messages. Let $\mathcal{G}_{col}(\mathcal{M}) \subset \mathcal{G}(\mathcal{M})$ denote the subset of structure graphs where there's a collision on the outputs:

$$\mathcal{G}_{col}(\mathcal{M}) \stackrel{\text{def}}{=} \{G \in \mathcal{G}(\mathcal{M}) \ ; \ \exists i, j, 1 \le i < j \le q : V_i^{m_i}(G) = V_j^{m_j}(G)\}$$

As just said, with this definition $\mathbf{CP}_{q,n,\ell} = \Pr_{G \stackrel{\$}{\leftarrow} \mathcal{G}(\mathcal{M})}[G \in \mathcal{G}_{col}(\mathcal{M})]$.
We split this probability into the "single accident" and the "two or more accidents" case. For this let $\mathcal{G}_{col}^i \stackrel{\text{def}}{=} \{G \in \mathcal{G}_{col}(\mathcal{M}) \ ; \ |\mathsf{Acc}(G)| = i\}$, now

$$\mathbf{CP}_{q,n,\ell} = \Pr_{G \stackrel{\$}{\leftarrow} \mathcal{G}(\mathcal{M})}[G \in \mathcal{G}_{col}^1(\mathcal{M})] + \Pr_{G \stackrel{\$}{\leftarrow} \mathcal{G}(\mathcal{M})}[G \in \mathcal{G}_{col}^i(\mathcal{M}) \text{ for some } i \ge 2]. \quad (6)$$

To bound the second term on the rhs. of (6) we can use Lemma 3 and "generously" upper bound the probability that there are two or more accidents.

$$\Pr_{G \stackrel{\$}{\leftarrow} \mathcal{G}(\mathcal{M})}[G \in \mathcal{G}_{col}^i(\mathcal{M}) \text{ for some } i \ge 2] \quad \le \Pr_{G \stackrel{\$}{\leftarrow} \mathcal{G}(\mathcal{M})}[|\mathsf{Acc}(G)| \ge 2] \le \frac{4q^4\ell^4}{2^{2n}}. \quad (7)$$

To bound the first term on the rhs. of (6) we can't be so generous any more and simply upper bound the probability of $|\mathsf{Acc}(G)| = 1$ as this would only give a $O(q^2\ell^2/2^n)$ bound. We will more carefully upper bound $|\mathcal{G}_{col}^1(\mathcal{M})|$ (by Lemma 4 below), and then apply Lemma 2 which in our case states that $G \in \mathcal{G}_{col}^1(\mathcal{M})$ appears with

$$\Pr_{G \stackrel{\$}{\leftarrow} \mathcal{G}(\mathcal{M})}[G \in \mathcal{G}_{col}^1(\mathcal{M})] \le \frac{|\mathcal{G}_{col}^1(\mathcal{M})|}{2^n - \ell q}. \quad (8)$$

Lemma 4. *Let $n, q \geq 1$ and $1 \leq m_1, \ldots, m_q \leq \ell$ and $\mathcal{M} = \{M_1, \ldots, M_q\}$ with $M_i \in B_n^{m_i}$ be distinct messages, then*

$$|\mathcal{G}_{col}^1(\mathcal{M})| \leq q(q + \ell + \ell^2)/2$$

Now combining (6)-(8) and the above Lemma we get:

Lemma 5. $\mathbf{CP}_{q,n,\ell} \leq \frac{q(q+\ell+\ell^2)}{2(2^n - \ell q)} + \frac{4q^4\ell^4}{2^{2n}}$

This already gives $\mathbf{CP}_{q,n,\ell} \in O(q^2/2^n)$ for $q^2\ell^4 \in O(2^n)$ and $q \in \Omega(\ell^2)$. But we can do better. The reason why this bound is not so great is that the term which bounds the "two or more" accident case is of rather large order $q^4\ell^4/2^n$ as we consider a graph (i.e. a total message length) of size $q\ell$. We achieve the bound claimed by Lemma 1 by splitting the messages in chunks of size ℓ^2 (with foresight) and then applying the following lemma which is a generalisation of (1).

Lemma 6. *If r divides q then* $\mathbf{CP}_{q,n,\ell} \leq \mathbf{CP}_{2r,n,\ell} \cdot \frac{q(q-r)}{2 \cdot r^2}$

Proof. Consider q messages $M_1, \ldots, M_q$ where $\mathbf{CP}_{q,n,\ell} = \mathbf{CP}_n(M_1, \ldots, M_q)$. We split the q messages into q/r sets $S_1, \ldots, S_{q/r}$, each containing r messages. If two messages collide, then there are two sets containing this two messages, so using the union bound $\mathbf{CP}_n(M_1, \ldots, M_q) \leq \sum_{i,j,1 \leq i < j \leq q/r} \mathbf{CP}_n(S_i, S_j)$.
The lemma follows as by definition $\mathbf{CP}_n(S_i, S_j) \leq \mathbf{CP}_{2r,n,\ell}$ and the sum has $q(q-r)/2r^2$ terms. □

We now have all ingredients to prove our main result

Proof (of Lemma 1). Let $\widetilde{q}$ be minimal satisfying $\widetilde{q} \geq q$ and $\ell^2 | \widetilde{q}$. Now using Lemma 6 (with $r = \ell^2$) in the second, and Lemma 5 in the third step

$$\mathbf{CP}_{q,n,\ell} \leq \mathbf{CP}_{\widetilde{q},n,\ell} \leq \mathbf{CP}_{2\ell^2,n,\ell} \cdot \frac{\widetilde{q}^2}{2\ell^4} \leq \left(\frac{\ell^2(3\ell^2 + \ell)}{2^n - 2\ell^3} + \frac{4(2\ell^2)^4\ell^4}{2^{2n}} \right) \frac{\widetilde{q}^2}{2\ell^4} \quad (9)$$

We can assume that $2\ell^3 \leq 2^{n-1}$ and $n > 1$ as otherwise the above is ≥ 1 which is a trivial upper bound for $\mathbf{CP}_{q,n,\ell}$. We also have $\widetilde{q} < 2q$ by the $q \geq \ell^2$ precondition an can further simplify (9) to $\mathbf{CP}_{q,n,\ell} \leq \frac{16 \cdot q^2}{2^n} + \frac{128 \cdot q^2 \cdot \ell^8}{2^{2n}}$. □

Proof (of Lemma 4). Wlog. we assume that $m_j \leq m_{j+1}$ for $1 \leq j \leq q-1$. Let

$$\mathcal{G}_{i,j} = \{G \in \mathcal{G}_{col}(\mathcal{M}) \; ; \; V_i^{m_i}(G) = V_j^{m_j}(G) \wedge |\mathsf{Acc}(G) = 1|\}$$

denote the structure graphs with exactly one accident, and where there's a collision on the outputs of the i'th and j'th message. Let $\mathcal{P}_j \subseteq [j-1]$ denote the indices of the messages which are prefixes of M_j after the common suffix has been removed, more formally

$$\mathcal{P}_j = \{i \in [1, \ldots, j-1] \; ; \exists S, X \in B_n^* \; : \; M_j = M_j' \| S, M_i = M_i' \| S, M_j' = M_i' \| X\}$$

Let $\overline{\mathcal{P}} = [1, \ldots, j-1] \setminus \mathcal{P}$. For example if $\mathcal{M} = \{M_1 = A, M_2 = AB, M_3 = ABC, M_4 = ACDB\}$ then $\mathcal{P}_4 = \{1, 2\}$ and $\overline{\mathcal{P}}_4 = \{3\}$.

We will prove two claims, which then will imply the statement of the lemma. The first claim — which is basically Lemma 19 from [BPR05] — states that if $i \in \overline{\mathcal{P}}_j$, then there's at most one structure graph with exactly one accident where M_i and M_j collide.

The second claim bounds the number of structure graphs having one accident and a collision between M_j and any other message M_i where $i \in \mathcal{P}_j$ by $\ell(\ell+1)/2$ (note that this bound only depends on the length, but not on the number of messages considered). To prove this claim we use the simple observation that if there's a collision between M_j and any M_i where $i \in \mathcal{P}_j$, then it must be the case that the j-path makes a loop. So we can upper bound the number of structure graphs having such a collision by the number of structure graphs having where the j-path loops.

Fig. 3. Figure for proof of Claim 1

Claim 1 *For each* $i \in \overline{\mathcal{P}}_j$, $|\mathcal{G}_{i,j}| \leq 1$.

Proof (of Claim). Let P denote the common prefix and S the common suffix of M_i and M_j. So $M_i = P\|M_i'\|S$ and $M_j = P\|M_j'\|S$ where the M_i' and M_j' are nonempty as $i \in \overline{\mathcal{P}}_j$. Let $p = |P|/n, s = |S|/n$.

By definition $G \in \mathcal{G}_{i,j}$ means $V_i^{m_i} = V_j^{m_j}$, this implies that also $V_i^{m_i-s} = V_j^{m_j-s}$ (as for the last s steps the i and j path must go in parallel). Now as $M_i^{m_i-s-1} \neq M_j^{m_j-s-1}$ (otherwise we could extend the suffix) we have $V_i^{m_i-s-1} \neq V_j^{m_j-s-1}$ (because in a structure graph two edges with distinct labels cannot be parallel). So there's a true collision in $G_{[i,j]}$ which hits the vertex $V_i^{m_i-s}$.

As by fact (i)[9] there can be only one true collision in $G_{[i,j]}$ this means that the "suffix path" $V_i^{m_i-s} = V_j^{m_j-s} \to \ldots \to V_i^{m_i} = V_j^{m_j}$ has no loops. For the same reason the "prefix path" $V_i^1 = V_j^1 \to \ldots \to V_i^p = V_j^p$ makes no loop and also the prefix and suffix paths must be disjoint. So the subgraph of $G_{[i,j]}$ built by the first $p+1$ and the last $s+1$ edges of the i and j path looks like shown on the left in Figure 3. There's only way to extend this subgraph to the full $G_{[i,j]}$ without introducing more true collisions, this is the second graph in Figure 3.

So there's only one possible $G_{[i,j]}$, and by fact (ii) it uniquely determines the whole structure graph, thus there's just one $G \in \mathcal{G}_{i,j}$. △

Claim 2 $\left|\bigcup_{i\in\mathcal{P}} \mathcal{G}_{i,j}\right| \leq \ell(\ell+1)/2.$

[9] We refer to the facts stated at the end of Section 3.

Proof (of Claim). Consider any $i \in \mathcal{P}$, and let S denote the common suffix of M_i and M_j. Now, as $i \in \mathcal{P}$, for some P we can write $M_j = P\|M'_j\|S$ and $M_i = P\|S$. Let $p = |P|/n$ and $s = |S|/n$.

Consider any $G \in \mathcal{G}_{i,j}$, by definition $V_i^{m_i} = V_j^{m_j}$, which implies $V_i^{m_i-s} = V_j^{m_j-s}$ as the last s blocks are equal. And as the first p blocks are equal we have $V_i^p = V_j^p$. Now $V_i^p = V_i^{m_i-s}$ and thus also $V_j^p = V_j^{m_j-s}$, as $p < m_j - s$ there's a true collision on the j-path (i.e. it contains a loop). As there are at most $m_j(m_j+1)/2$ possibilities for the j-path to make a loop[10] and as by fact (ii) the shape of the j path determines G completely, there can be at most $m_j(m_j+1)/2 \le \ell(\ell+1)/2$ different G's in $\bigcup_{i\in\mathcal{P}} \mathcal{G}_{i,j}$. △

The lemma follows by the two claims as

$$
\begin{aligned}
|\mathcal{G}^1_{col}(\mathcal{M})| &\le \sum_{1\le i<j\le q} |\mathcal{G}_{i,j}| \le \sum_{j=1}^{q} \left(\Big| \bigcup_{i\in\mathcal{P}} \mathcal{G}_{i,j} \Big| + \sum_{i\in\overline{\mathcal{P}}} |\mathcal{G}_{i,j}| \right) \\
&\le \sum_{j=1}^{q} \left(\frac{\ell(\ell+1)}{2} + j - 1 \right) \le \frac{q(q-1+\ell(\ell+1))}{2}. \qquad \square
\end{aligned}
$$

References

[BKR00] Mihir Bellare, Joe Kilian, and Phillip Rogaway. The security of the cipher block chaining message authentication code. *Journal of Computer and System Sciences*, 61(3):362–399, 2000. Earlier version in *Crypto '94*.

[BP95] Antoon Bosselaers and Bart Preneel, editors. *Integrity Primitives for Secure Information Systems, Final Report of RACE Integrity Primitives Evaluation RIPE-RACE 1040*, volume 1007 of *LNCS* Springer, 1995.

[BPR05] Mihir Bellare, Krzysztof Pietrzak, and Phillip Rogaway. Improved security analyses for CBC MACs. In *Proc. Crypto '05*. Full Version on `www.crypto.ethz.ch/~pietrzak/publications.html`.

[BR00] John Black and Phillip Rogaway. CBC MACs for arbitrary-length messages: The three-key constructions. In*Proc. Crypto '00*.

[CW79] Larry Carter and Mark N. Wegman. Universal classes of hash functions. *Journal of Computer and System Sciences (JCSS)*, 18:143–154, 1979.

[DGH+04] Yevgeniy Dodis, Rosario Gennaro, Johan Håstad, Hugo Krawczyk, and Tal Rabin. Randomness Extraction and Key Derivation Using the CBC, Cascade and HMAC Modes. In *Proc. Crypto '04*.

[Dod05] Yevgeniy Dodis, 2005. Personal Communication.

[HW80] G. Hardy and E. Wright. *An Introduction to the Theory of Numbers.* Oxford University Press, 1980.

[Mau02] Ueli Maurer. Indistinguishability of random systems. In *Proc. Eurocrypt '02*.

[NES] NESSIE. European project ist-1999-12324 on new european schemes for signature, integrity and encryption. `http://www.cryptonessie.org.`

[PR00] Erez Petrank and Charles Rackoff. Cbc mac for real-time data sources. *Journal of Computer and System Sciences*, pages 315–338, 2000.

[10] This observation is trivial, it also follows from the (more general) equation (6) in [BPR05].