From Weak to Strong Watermarking

Nicholas Hopper[1], David Molnar[2], and David Wagner[2]

[1] University of Minnesota, Minneapolis MN 55455, USA
hopper@cs.umn.edu
[2] University of California - Berkeley, Berkeley CA 94720 USA
{dmolnar, daw}@eecs.berkeley.edu

Abstract. The informal goal of a watermarking scheme is to "mark" a digital object, such as a picture or video, in such a way that it is difficult for an adversary to remove the mark without destroying the content of the object. Although there has been considerable work proposing and breaking watermarking schemes, there has been little attention given to the formal security goals of such a scheme. In this work, we provide a new complexity-theoretic definition of security for watermarking schemes. We describe some shortcomings of previous attempts at defining watermarking security, and show that security under our definition also implies security under previous definitions. We also propose two weaker security conditions that seem to capture the security goals of practice-oriented work on watermarking and show how schemes satisfying these weaker goals can be strengthened to satisfy our definition.

1 Introduction

Informally, a digital watermarking scheme is a procedure which embeds a "mark" in an object so that it is hard to remove the mark without "damaging" the object. These procedures have a wide variety of applications to digital rights management, including detection of unauthorized copies, limitations on media copying, tracing of information leaks, and resolution of ownership disputes over digital content; for further exposition on various applications see, for example [1, ch. 20]. As a result, watermarking schemes have seen intense research efforts; for example, see [2] and the references therein, or the proceedings [3–16]. Most of this work is focused on the construction of schemes for various digital media and attacks on these schemes, where there is a long history of schemes being broken almost immediately after they are proposed.

Given this history, it is not surprising that in the security community, there is a perception that secure watermarking is "theoretically impossible," as expressed, for instance, in [1, 17, 18]. While this idea is intuitively appealing, it is difficult to prove something is (im)possible without first formally defining the notion. Consider for instance, the related notions of program obfuscation and steganography, which were both widely believed to be impossible. Program obfuscation was formalized and shown to be impossible in general [19], but subsequently some progress has been made in limited cases [20, 21]. Steganography, on the other hand, was formalized and shown to be possible, but at limited rates [22–24].

Surprisingly, formal definitions for watermarking security have only recently appeared in the literature. The state of the art focuses on defining schemes secure against specific "protocol attacks," which attack the protocols that use a watermark rather than removing a mark from an object [25]; these very powerful attacks changed researchers' understanding of what it means for a watermark to be "secure." For example, Kutter *et al.* [26] introduced the *copy attack,* in which a watermark is copied from an object O_1 into an object O_2 to form an object O_2' that appears marked even though it was never legitimately watermarked. This makes it impossible to use the attacked watermarking scheme for various applications, such as resolving ownership disputes.

Later Adelsbach, Katzenbeisser, and Veith formalized copy attacks and a different protocol attack known as an ambiguity attack. They then showed protocols intended to be provably secure against these attacks [27]. Several other authors have also produced schemes claimed to be provably resistant to copy attacks or other protocol attacks [28, 29].[3] While this line of work has led to interesting results, there are some limitations,

[3] We stress that these constructions, similarly to our own, do not attempt to construct a provably secure watermark "from scratch" but rather try to build something "secure against X" from a watermark that is not assumed to be secure in this sense.

which we summarize in Appendix B. Additionally, this approach leads to an "arms race," in which, as new protocol attacks are discovered, new watermarking schemes must be designed and proven secure.

The primary contribution of this work is to initiate the systematic study of watermarking security definitions. We define a "strong watermarking" security condition with respect to a metric space on objects, which compares a watermark to an *ideal functionality* in which an object is marked if and only if it is similar to some object previously marked by the functionality. We show that this definition implies security against previously known protocol attacks, and explore the question of proving impossibility. We also explore weaker security conditions and show how, under some conditions, schemes satisfying these weaker definitions can be strengthened or amplified to produce strong watermarks.

We stress that in these latter results, we explicitly do not construct "secure" watermarking schemes from scratch. Instead, we show that watermark designers can achieve a strong notion of security from weak constructions that are not secure against protocol attacks. These results have two implications. First, impossibility results for strong watermarking in a metric space will also imply impossibility of these weaker goals. Second, this means that watermark designers need not complicate their schemes by attempting to rule out protocol attacks. Instead, they need only achieve the weaker notion and then apply our results; put another way, it is enough to build schemes that heuristically satisfy these goals and apply our constructions to build (heuristically) strong watermarking schemes, similar to results that say we can build (heuristically) strong secret-key encryption schemes from (heuristically) strong block ciphers.

Overview of our results. In Section 3 we propose a new definition of secure watermarking schemes, that we call *strong watermarking*, in the case that the marking and detecting procedures share a secret key. Our definition allows the adversary to make adaptive queries to oracles for both marking an object and detecting whether an object is marked. The main idea of the definition is that a strong watermarking scheme (in which there is *no communication between the marking and detection procedures*) should simulate an "ideal watermarking functionality," which we define. We show that strong watermarking implies security against all known protocol attacks, and argue that the definition will imply security against future protocol attacks. Furthermore, we show that security in our model depends critically on both the notion of similarity and the distribution on objects to be marked; specifically, we show an example of these settings under which strong watermarking is impossible, and an example where strong watermarking exists, relative to an oracle.

In Section 4 we introduce a "weaker" notion of watermark, which we call a *non-removable embedding*. This is a weak notion because it only requires that the watermark cannot be removed; we explicitly allow copy and ambiguity attacks to succeed against non-removable embeddings. We formalize this notion, prove a separation between the notion and our proposed strong definition, and point out that many watermarking schemes in the literature use a security metric closely related to this notion. We also introduce a notion of "limited" adversaries, who only create new objects based on some limited set of transformations. This notion is interesting since there are some techniques in the watermarking literature which seem to imply provable security against "limited" attacks such as Gaussian noise. Additionally, some applications of watermarking only require watermarks to be "robust" against distortions caused by physical processes; these can be modeled by limited adversaries. We note that all of our results on amplification can be easily extended to the limited adversarial setting. We then show how schemes that are provably secure under the strong watermarking definition can be constructed from non-removable embeddings plus a semi-offline trusted third party, a standard digital signature scheme, and a semantically secure symmetric encryption scheme. This shows that our notion of strong watermarking can be built on the "weak" primitive of non-removable embeddings. While we do require a third party, this party is not required during watermark detection.

In Section 5 we study an alternative method for producing a strong watermarking scheme. Specifically, we consider the question of *security amplification* of watermarking schemes. We formally specify two new notions that correspond to a weaker version of strong watermarking and show how schemes which satisfy these natural conditions can be efficiently composed to produce strong watermarking schemes. Note that this construction can be seen as an heuristic method to create strong watermarking schemes as well as a way to extend impossibility results for a given notion of similarity.

2 Preliminaries

We will work with discrete metric spaces. A *discrete metric space* $\mathcal{M}$ is a finite space equipped with a distance function $d : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{Z}^+ \cup \{0\}$. The distance function is symmetric, obeys the triangle inequality and has the property that if $d(x, y) = 0$ then $x = y$. We will associate with a metric space a similarity relation $\sim$ defined by $x \sim_\delta y \equiv d(x, y) \leq \delta$ for some fixed δ. When the meaning is clear from context, we will drop the δ and simply write $\sim$. For simplicity, we will assume that all parties can efficiently evaluate $\sim$. Finally, we denote by $\mathcal{D}$ a distribution on $\mathcal{M}$. Unless otherwise specified, we assume that all parties can efficiently sample from $\mathcal{D}$ and we denote by $O \leftarrow_R \mathcal{D}$ an object $O \in \mathcal{M}$ sampled according to the distribution $\mathcal{D}$.

We will also make use of a digital signature scheme $\mathcal{S} = \{\mathsf{SGen}, \mathsf{Sig}, \mathsf{Ver}\}$. We say that a signature scheme is (t, q, ϵ)-existentially unforgeable under adaptive chosen message attack [30] if all adversaries running in time at most t making at most q queries to a signature oracle have chance at most ϵ of obtaining a signature on a message not previously queried.

We will use a symmetric encryption scheme $\mathcal{SE} = \{\mathsf{Encrypt}, \mathsf{Decrypt}\}$. We say that a symmetric encryption scheme is (t, q, ϵ)-secure in the left-or-right sense [31] if every time t adversary, given q queries to a "left-or-right" oracle $LOR_K(b, x_0, x_1) = \mathsf{Encrypt}(K, x_b)$ cannot distinguish between the case that $b = 0$ and $b = 1$ with advantage better than ϵ.

Finally, we will need a pseudorandom function ensemble $\left\{F : \{0,1\}^k \times \{0,1\}^{L(k)} \rightarrow \{0,1\}^{\ell(k)}\right\}_{k\in\mathbb{N}}$ [32]. We say that a function ensemble is (t, q, ϵ)-pseudorandom if any adversary running in time at most t and making at most q queries to a function oracle can distinguish an oracle for $F(U_k, \cdot)$ from an oracle for a random function $f : \{0,1\}^{L(k)} \rightarrow \{0,1\}^{\ell(k)}$ with advantage at most ϵ.

3 Strong Watermarking

As previously mentioned, the informal notion of a watermarking scheme requires the ability to somehow "mark" digital objects, such as pictures, sound, video, or text. The scheme should also satisfy several additional requirements:

- The result, O', of marking an object, O, should be "similar" to O.
- An adversary, given O', should not be able to find an object O'' that is similar to O' but unmarked; this prevents removal of the mark except by "damaging" the object.
- Most objects O must not be marked. If this is not the case, then certain desirable uses of watermarks, such as searching for copies of O' and proving ownership of O', are not possible.
- There should be no communication required between the marking procedure and the detecting procedure; or this communication should be minimized. This is necessary for many applications, for example, a media player that may not have a network connection.

We will model the notion of similarity or damage by postulating the existence of a "perceptual metric" that measures the distance between objects of a given type. Thus such a metric would assign a small distance between two pictures that look alike and a large distance between two very different pictures. In practice it is difficult to characterize such a metric space, so researchers typically focus on Euclidean or weighted L_1 distance in some "perceptually significant" space such as the Fourier [33], Wavelet [34], or Fourier Mellin [35] transforms. Once we fix a metric d, the natural notion of similarity is the relation $\sim_\delta$ defined previously, that is, we will say that objects O_1 and O_2 are similar if $d(O_1, O_2) \leq \delta$.

Given this formalization of similarity, we can construct a perfectly secure watermarking scheme that optimally satisfies the above requirements. To mark an object O with key K, the ideal scheme simply adds O to its list of objects marked with K; to test whether an object O' is marked with K, the ideal scheme simply searches the appropriate list of marked objects and returns `true` if it finds an object similar to O' and `false` otherwise. This "ideal" scheme does not allow an adversary to succeed in "unmarking" a marked object but leaves the largest possible set of objects unmarked subject to this constraint. The ideal scheme is undesirable in that it requires unbounded, online communication between the marking and detection algorithms; our intent is to compare a real-world watermarking scheme (which does not allow any online communication between the marking and detection procedures) to this ideal.

Oracle $\mathsf{Mark}^*(O)$:	**Oracle** $\mathsf{Detect}^*(O)$:	**Oracle** $\mathsf{Challenge}^*_{\mathcal{D}}()$
1. $O' \leftarrow \mathsf{Mark}(K, O)$	1. $b \leftarrow \mathsf{Detect}(K, O)$	1. $O \leftarrow_R \mathcal{D}$
2. $\mathsf{Marked} \leftarrow \mathsf{Marked} \cup \{O'\}$	2. $B' \leftarrow \mathsf{IdealDetect}(O)$	2. $O' \leftarrow \mathsf{Mark}(K, O)$
3. **return**(O')	3. **if** $b \notin B'$	3. $\mathsf{chalns} \leftarrow \mathsf{chalns} \cup \{O'\}$
	4. **then** $\mathsf{bad} \leftarrow \mathsf{true}$	4. $\mathsf{Marked} \leftarrow \mathsf{Marked} \cup \{O'\}$
	5. **return**(b)	5. **return**(O')

Fig. 1. Definition of Mark^*, $\mathsf{Challenge}^*$, and Detect^* oracles for strong watermarking. The global variables K, Marked, chalns, and bad are initialized in figure 2

An informal statement of our definition allows an adversary access to a marking oracle and a detection oracle for a watermarking scheme. The adversary then attempts to attack the scheme by finding an object such that the results of the actual detection algorithm and the ideal detection procedure differ: either the object is marked and should not be, or it is unmarked and should be. Unfortunately, any watermarking scheme that produces objects that are similar to its input and has a static detection scheme would be insecure under this definition. The intuition is that the following attack would succeed with very high probability:

1. The adversary samples an object $O \in \mathcal{M}$. Since it has not been queried to the marking procedure, it is not yet marked under the ideal scheme.
2. Next the adversary queries $\mathsf{Mark}(O)$, to get an object O' similar to O.
3. Finally, the adversary queries $\mathsf{Detect}(O)$. In the watermarking scheme under attack, O should not be marked (since it was not marked in step 1, and there is no communication between marking and detection schemes). But in the ideal scheme, it is close to O', which *is* marked. Thus the adversary has succeeded in finding an object on which the real and ideal schemes differ.

We give a formal proof of this in Appendix C, where we also show that a cryptographically natural alternative definition also rules out secure schemes that distort originals by less than half the similarity radius. Our solution is to introduce a third, *challenge* oracle that selects objects to watermark from some probability distribution; the performance of the watermarking scheme is only compared to that of the ideal scheme on these challenge objects.

3.1 Definition of Strong Watermarking Schemes

A *secret-key watermarking scheme* $\mathcal{W} = \{\mathsf{WMGen}, \mathsf{Mark}, \mathsf{Detect}\}$ consists of three algorithms: $\mathsf{WMGen} : 1^* \rightarrow \mathsf{Keys}$ generates a secret key to be used in marking and detection; $\mathsf{Mark} : \mathsf{Keys} \times \mathcal{M} \rightarrow \mathcal{M}$ takes a key and an object to mark and returns a new object; and $\mathsf{Detect} : \mathsf{Keys} \times \mathcal{M} \rightarrow \{\mathsf{true}, \mathsf{false}\}$. Notice that *we do not explicitly allow any online communication between the* Detect *and* Mark *procedures*, since in many applications the devices detecting and marking objects may not have any means by which to communicate.

We can now define *strong watermark security.* Our definition formalizes the informal discussion above. An adversary is given access to oracles for Mark and Detect, and a special $\mathsf{Challenge}^*$ oracle that samples and marks objects from an efficiently sampleable distribution $\mathcal{D}$ over $\mathcal{M}$. The adversary wins if he calls Detect^* on an object that is either marked, but not similar to the result of a Mark^* or $\mathsf{Challenge}^*$ query, or unmarked, but similar to the result of some $\mathsf{Challenge}^*$ query. Notice that unlike in the hypothetical discussion above, we only require the objects near the *result* of Mark (rather than the input) to be marked, since these are (presumably) the ones that the adversary will be able to access. The formal security experiment has four global variables: Marked and chalns, sets of objects; bad, a boolean flag; and K, a key. In Figures 1 and 2 we show pseudocode for initializing the security experiment and the ideal detection functionality, as well as for oracles Mark^*, $\mathsf{Challenge}^*$, and Detect^*. We note that some of our reductions require the ability to sample from a distribution $\mathcal{D}'$ on $\mathcal{M}$.

We say that a watermark is *ρ-preserving for $\mathcal{D}$* if $\Pr[K \leftarrow \mathsf{WMGen}(1^k); O \leftarrow \mathcal{D}; O' \leftarrow \mathsf{Mark}(K, O) : d(O, O') > \rho]$ is negligible in k; that is, if the marked version of an object is almost always within distance ρ of the original. This "bounded distortion" requirement is not strictly necessary for security in all applications, but is typically vital to the utility of a watermarking scheme.

Experiment $\mathbf{Exp}^{strong}_{\mathcal{D},W}(A)$:
$K \leftarrow \mathsf{WMGen}(1^k)$
$\mathsf{bad} \leftarrow \mathsf{false}$
$\mathsf{Marked} \leftarrow \emptyset$
$\mathsf{chalns} \leftarrow \emptyset$
$A^{\mathsf{Mark}^*,\mathsf{Challenge}^*,\mathsf{Detect}^*}()$
return (bad)

Procedure $\mathsf{IdealDetect}(O)$:
if $(\exists O' \in \mathsf{chalns}\ :\ O \sim O')$
  then return $\{\mathsf{true}\}$
else if $(\exists O' \in \mathsf{Marked}\ :\ O \sim O')$
  then return $\{\mathsf{true}, \mathsf{false}\}$
else
  return $\{\mathsf{false}\}$

$$\mathbf{Adv}^{strong}_{\mathcal{D},\mathcal{W}}(A) = \Pr[\mathsf{bad} = \mathsf{true}]$$

Fig. 2. Definition of security experiment for strong watermarking.

The advantage of an adversary A_{Strong} is $\mathbf{Adv}^{strong}_{\mathcal{D},W}(A_{Strong})$ as defined in Figure 2. The scheme is a $(\mathcal{D}, t, q_M, q_D, q_C, \epsilon, \delta)$-*strong watermarking scheme* if for all adversaries A_{Strong} running in time at most t, making at most q_M queries to Mark^*, at most q_D queries to Detect^*, and at most q_C queries to $\mathsf{Challenge}^*$, the advantage of A_{Strong} is at most ϵ with respect to similarity relation $\sim_\delta$.

Philosophically, one may think of the above experiment as a game between, say, a "hacker" and a "studio." The hacker can "give" movies to the studio to see how they look when marked, and he can check, using his personal DVD player, whether any particular object is marked. Meanwhile, the studio will release other videos not created by the hacker; it is the hacker's goal to "unmark" one of these movies, or alternatively, to create a movie that appears to be marked but was never marked by the studio. If the hacker cannot do this, the studio can have good confidence that a movie will appear marked iff it was produced by them.

Dependence on $\sim$ and $\mathcal{D}$. It should be clear that the existence of strong watermarks depends critically on both the similarity relation $\sim$ and the distribution on challenge objects, $\mathcal{D}$. For instance, if an attacker can deduce, given the result of a query to $\mathsf{Challenge}^*_{\mathcal{D}}$ the object $O \leftarrow D$ from line 1 of Figure 1, then as pointed out in our earlier discussion, the scheme cannot be secure for $\mathcal{D}$ and $\sim$. Thus $\mathcal{D}$ must have high entropy, and be "one-way" for most keys. Likewise, if for any given O, enumerating the set $N_\delta(O) = \{O' : O' \sim O\}$ is feasible, then a watermarking scheme cannot be secure. In this work, we do not explore all the necessary conditions on $\sim$ and $\mathcal{D}$; it seems to be a difficult challenge to even identify the correct similarity metric and distribution for many of the applications of watermarking. Here we briefly give two results that show that even when the previous two conditions are satisfied, there cannot be a "generic" argument for the existence or impossibility of strong watermarks.

Proposition 1. *Let $\mathcal{D}$ be the uniform distribution on k-bit strings and let $d(x, y)$ be the hamming distance metric on k-bit strings. Then there is no δ-preserving, $(\mathcal{D}, O(k), 1, 1, 1, 1/2^{\delta+1}, \delta)$-strong watermarking scheme.*

Notice that for $\delta(k) = O(\log k)$, the neighbor set has size superpolynomial in k, and $\mathcal{D}$ has k bits of entropy, yet no watermarking scheme can have security better than $1/2k$. The proposition can be seen to be true as follows. Suppose we uniformly pick a point $x \in \{0, 1\}^k$; consider the point y returned by $\mathsf{Mark}^*(x)$, and let z and w be uniformly chosen points in $N_\delta(y)$ and $N_\delta(x)$, respectively. Now we know that if a watermarking scheme is to be ε-secure, it must be that $\Pr[\mathsf{Detect}^*(z) = \mathsf{false}] \leq \varepsilon$, since otherwise an adversary can remove a mark with probability greater than ϵ by sampling a random point in the neighborhood of a marked object. It can also be shown that $\Pr[z \in N_\delta(x)] \geq 1/2^\delta$. This gives us that $\Pr[z \in N_\delta(x) \wedge \mathsf{Detect}^*(z) = \mathsf{true}] \geq 1 - (\Pr[z \notin N_\delta(x)] + \Pr[\mathsf{Detect}^*(z) = \mathsf{false}]) \geq 2^{-\delta} - \varepsilon$. Note that ε security also requires that $\Pr[\mathsf{Detect}^*(w) = \mathsf{true}] \leq \varepsilon$, since otherwise we can easily find a marked point – by randomly sampling an object in the neighborhood of a random point – breaking the watermark. Thus we also have that $\varepsilon \geq \Pr[\mathsf{Detect}^*(w) = \mathsf{true} \wedge w \in N_\delta(y)]$. But by symmetry, for any fixed choice of K, x, y, we have $\Pr[\mathsf{Detect}^*(w) = \mathsf{true} \wedge w \in N_\delta(y)] = \Pr[\mathsf{Detect}^*(z) = \mathsf{true} \wedge z \in N_\delta(x)]$. This gives $\varepsilon \geq 2^{-\delta} - \varepsilon$, or $\varepsilon \geq 2^{-\delta-1}$.

Notice that a similar argument applies to any metric space, distribution and marking function such that (i) the neighborhood of an object and its marked version are symmetric, (ii) these neighborhoods have noticeable intersection, and (iii) it is possible to uniformly sample from the neighborhood set of an object. Thus to rule out an impossibility result, we seek to violate these properties.

Proposition 2. *There exists an oracle Π, relative to which there exists a distribution $\mathcal{D}_\Pi$, a metric d_Π, and a 1-preserving watermarking scheme W^Π such that W^Π is $(\mathcal{D}_\Pi, t, t, t, t, t^2/2^k, 1)$-strong.*

Experiment $\mathbf{Exp}^{cp}_{\mathcal{D},W}(B)$:
$K \leftarrow \mathsf{WMGen}(1^k)$
$O_1 \leftarrow_R \mathcal{D}$
$O_1' \leftarrow \mathsf{Mark}(K, O_1)$
$O_2 \leftarrow_R \mathcal{D}$
$O_2' \leftarrow B(O_1', O_2)$
if $\mathsf{Detect}(K, O_2')$
  and $O_2 \sim O_2' \not\sim O_1'$
  then $b = \mathsf{true}$
  else $b = \mathsf{false}$
return(b)

$\mathbf{Adv}^{cp}_{\mathcal{D},W}(B) = \Pr[b = \mathsf{true}]$

Experiment $\mathbf{Exp}^{amb}_{\mathcal{D},W}(B)$:
$K \leftarrow \mathsf{WMGen}$
repeat
  $O_1 \leftarrow_R \mathcal{D}$
until $\mathsf{Detect}(K, O_1) = \mathsf{false}$
$O_1' \leftarrow B(O_1)$
if $\mathsf{Detect}(K, O_1')$ **and** $O_1 \sim O_1'$
  then $b = \mathsf{true}$
  else $b = \mathsf{false}$
return(b)

$\mathbf{Adv}^{amb}_{\mathcal{D},W}(B) = \Pr[b = \mathsf{true}]$

Adversary $A^B_{cp}()$:
1. $O_1 \leftarrow \mathsf{Mark}^*(O \leftarrow \mathcal{D})$
2. $O_2 \leftarrow_R \mathcal{D}$
3. $O_2' \leftarrow B(O_1', O_2)$
4. $\mathsf{Detect}^*(O_2')$

Adversary A^B_{amb}:
1. $O_1 \leftarrow_R \mathcal{D}$
2. $\mathsf{Detect}^*(O_1)$
3. $O_1' \leftarrow B(O_1)$
4. $\mathsf{Detect}^*(O_1')$

Fig. 3. Experiments for copy and ambiguity attacks and the corresponding strong watermark adversary.

Intuitively, we will choose Π, d_Π and $\mathcal{D}_\Pi$ so that for most strings x it will be very hard to even find a string y such that $d_\Pi(x, y) = 1$, but the oracle gives us a way to sample from a set of "special" strings x' that violate this property. Once we mark an object x' it is no longer in this special set, so it is hard for the adversary to remove the mark. Formally, the oracle Π "knows" a uniformly chosen bijection $\pi : \{0,1\}^{2k} \rightarrow \{0,1\}^k \times \{0,1\}^k$ for each k and answers three types of queries: sample, dist, and move. $\Pi(\mathsf{sample}, y)$ returns $\pi^{-1}(y, 0^k)$. $\Pi(\mathsf{dist}, x_0, x_1)$ computes $(y_b, z_b) = \pi(x_b)$, and then returns 0 if $x_0 = x_1$, 1 if $y_0 = y_1$ and some $z_b = 0^k$, 2 if $y_0 = y_1$, and 3 otherwise. $\Pi(\mathsf{move}, x, z')$ computes $(y, z) = \pi(x)$; if $z = 0^k$ then it returns $\pi^{-1}(y, z')$; if $z = z'$ it returns $\pi^{-1}(y, 0^k)$, and otherwise it returns x. The distribution $\mathcal{D}_\Pi$ is defined as $\Pi(\mathsf{sample}, U_k)$ and the metric $d_\Pi(x, y) = \Pi(\mathsf{dist}, x, y)$, so that for most $2k$-bit strings x, there is only one string at distance 1 from x. The marking scheme W^Π uses k-bit keys, and computes $\mathsf{Mark}^\Pi(K, x) = \Pi(\mathsf{move}, x, K)$, while $\mathsf{Detect}^\Pi(K, x)$ returns true iff $\Pi(\mathsf{move}, x, K) \neq x$.

We remark that, obviously, the oracle distribution Π does not prove that strong watermarks exist. It merely shows that there cannot be a "black-box" proof that rules out all possible strong watermarking schemes without considering the details of $\mathcal{D}$ and $\sim$. We believe it is an interesting open question to find any $\mathcal{D}$ and $\sim$, even if they are contrived, that provably admit a strong watermarking scheme without reference to an oracle, or even with small values (q_M, q_C, q_D).

3.2 Strong Watermarks Are Secure Against Protocol Attacks

Adelsbach et al. provided the first formal definition of copy attacks and ambiguity attacks [27]. We adapt their definitions to our setting, in which we consider only the presence of a mark rather than its content. We show that strong watermarks are secure against copy and ambiguity attacks.

First we consider copy attacks. Informally, a copy attack occurs when an adversary can "copy" a watermark from a marked object O_1' to a second object O_2. In our watermarking model, "copy" means that the adversary, given a marked object O_1', can cause an object O_2 to return true for Detect^* despite never having been queried to Mark. More formally, we say a watermarking scheme is $(\mathcal{D}, t, \epsilon_{cp}, \delta_{cp})$-secure against copy attacks if all adversaries B running in time at most t have advantage $\mathbf{Adv}^{cp}_{\mathcal{D},\mathcal{W}}(B) \leq \epsilon_{cp}$ with respect to similarity relation $\sim_{\delta_{cp}}$. Notice that in this definition (and in the original definition of Adelsbach et al. [27]) the copy adversary is not afforded access to a Mark^* or Detect^* oracle. We can prove that a $\mathcal{D}$-strong watermarking scheme is not vulnerable to copy attacks for any sampleable distribution $\mathcal{D}'$: if there exists an adversary B that successfully carries out a copy attack, then the adversary A^B_{cp} in Figure 3 succeeds at breaking the strong watermark. A formal theorem statement and proof are in Appendix A.

Next, we consider ambiguity attacks. A classical ambiguity attack takes an unmarked object O_1, and produces a new "original" object O_2 such that O_1 appears to be marked with O_2 as the original. In our model, we can recast ambiguity attacks as, given an unmarked object O_1, find an object O_2 such that $O_2 \sim O_1$ and O_2 appears to be marked, without legitimately marking O_2. Strong watermarking implies security against ambiguity attacks: if B succeeds at carrying out an ambiguity attack, then the adversary A^B_{amb} shown in Figure 3 breaks the strong watermark. Details are in Appendix A.

Remark. We note that some works on protocol attacks describe attacks where the adversary is allowed to choose the key to the watermarking scheme. While it is important to eventually address such *chosen-key attacks*, we believe it is an interesting and important first step to concentrate on getting the definitions right

Experiment $\mathbf{Exp}_{\mathcal{D}}^{NRE}(A)$:
1. $(z, z') \leftarrow \mathsf{EMGen}(1^k)$
2. $\mathsf{Embedded} \leftarrow \emptyset$
3. $O_A \leftarrow A^{\mathsf{Embed}(z,\cdot,\cdot),\mathsf{Challenge}^*}(z')$

Oracle $\mathsf{Challenge}^*(m)$:
1. $O \leftarrow_R \mathcal{D}$
2. $O' \leftarrow \mathsf{Embed}(z, O, m)$
3. $\mathsf{Embedded} \leftarrow \mathsf{Embedded} \cup \{(O', m)\}$
4. **return** O'

$$\mathbf{Adv}_{\mathcal{D}}^{NRE}(A) = \Pr[\exists (O_i, m_i) \in \mathsf{Embedded} : O_A \sim O_i \wedge \mathsf{Extract}(z', O_A) \neq m_i]$$

Fig. 4. Security experiment and Embed^* oracle for non-removable embeddings.

for the more basic scenario. Thus in this paper we do not consider attacks that involve manipulating the keys of the marking and detection procedures.

4 Non-Removable Embeddings and Strong Watermarks

Many watermarking schemes in the literature actually provide a somewhat different interface from the watermarking primitive described in the previous section. Instead, these schemes focus on embedding a short string within an object so that if the adversary does not distort the object too much, the embedded string can be recovered. Typical schemes do not attempt to prevent "insertion" of strings into an object, which is the reason that many protocol attacks succeed. In this section, we give a formal notion of a primitive, the *non-removable embedding* (NRE), that seems to capture this design goal. We will demonstrate that NREs are provably weaker objects than strong watermarks: if NREs exist at all, then there are NREs that allow copy attacks. After separating the notions of NREs and strong watermarks, we give a construction which makes limited use of a semitrusted third party to construct a strong watermarking scheme from a NRE.

The notion of an NRE is closely related to a security notion widespread in the watermarking literature. Many schemes presented in the watermarking literature, for example [36–40], take as their evaluation metric the bit error rate for a watermarked message given a specified constraint on the distortion allowed the adversary, or "watermark to noise ratio." Essentially, these schemes attempt to bound the rate of bit errors in the embedded string for a given amount of distortion induced by the adversary. One of the interesting properties of the NRE notion is that we can easily build an NRE from such schemes. Because we deal with probabilistic polynomial time adversaries, we can assume that the bit errors follow a computationally bounded distribution. Therefore, we can use the coding methods of Micali et al. to obtain an NRE from up to a bit error rate of one half: we simply encode the message before embedding and decode on extraction [41].

To begin, an embedding scheme $(\mathsf{Embed}, \mathsf{Extract}, \mathsf{EMGen})$ is a triple of algorithms with the following signatures: $\mathsf{Embed} : \mathsf{Aux} \times \mathcal{M} \times \{0,1\}^k \rightarrow \mathcal{M}$, $\mathsf{Extract} : \mathsf{Aux}' \times \mathcal{M} \rightarrow \{0,1\}^k \cup \perp$, and $\mathsf{EMGen} : 1^* \rightarrow \mathsf{Aux} \times \mathsf{Aux}'$ for some fixed k. Here $\mathcal{M}$ is a metric space, and Aux and Aux' are sets of possible auxiliary inputs. For example, Aux might be a set of secret keys, while Aux' might be a set of public keys. k is the length of strings to be embedded in objects.

We further require that embedded messages can be extracted, i.e. for $(z, z') \leftarrow \mathsf{EMGen}(1^k)$, we have $\mathsf{Extract}(z', \mathsf{Embed}(z, O, x)) = x$ with high probability. An embedding scheme is ρ-preserving for $\mathcal{D}$ if for all $m \in \{0,1\}^k$, $d(\mathsf{Embed}(O, m), O) \leq \rho$ with high probability over $O \leftarrow \mathcal{D}$. Together, these give a correctness and a bounded distortion requirement for a non-removable embedding.

We define security of embedding scheme NRE by saying it is $(\mathcal{D}, t, q_E, q_C, \epsilon, \delta)$ *non-removable for distribution* D if for all A running in time at most t, that make at most q_E queries to an Embed oracle and at most q_C queries to the $\mathsf{Challenge}^*$ oracle, the advantage $\mathbf{Adv}_{\mathcal{D}}^{NRE}(A)$ defined in Figure 4 is at most ϵ.

Remarks. This definition does not rule out the protocol attacks we have discussed: in particular, if there is a ρ-preserving non-removable embedding for the metric space $\mathcal{M}$ with metric d, we can construct a 2ρ-preserving non-removable embedding for the metric space $\mathcal{M} \times \{0,1\}^k$ with metric d', that allows copy attacks to succeed, as follows. We define the metric $d'((O_1, y_1), (O_2, y_2))$ to be $d(O_1, O_2)$ if $y_1 = y_2$ and $d(O_1, O_2) + \rho$ otherwise; define $\mathsf{Embed}'(z, (O, y), x) = (\mathsf{Embed}(z, O, x), x)$, and $\mathsf{Extract}'(z', (O, x)) = \mathsf{Extract}(z', O)$ if $\mathsf{Extract}(z', O) \neq \perp$ and $\mathsf{Extract}'(z', (O, x)) = x$ otherwise. Then it is easy to see that, as long as $\rho < \delta$, given a marked object $O = (O_1, x)$ and an unmarked object $O' = (O_2, y)$ we can "copy" the mark from O onto O' by setting $O'' = (O_2, x)$; yet it is still hard to remove x from O.

Although we do not explicitly require it, we note that typical applications will require that $\rho < \delta$ and in many cases, $\rho \ll \delta$. We also note that it is trivial to construct a ρ-preserving non-removable embedding for the case that $\rho = \sup_{(x,y) \in \mathcal{M} \times \mathcal{M}} d(x, y)$, using an error correcting code with minimum distance 2δ, if one exists for the metric space $\mathcal{M}$.[4] Thus the interesting question, for a given metric space, becomes "for what values of (ρ, δ) is a NRE possible?"

Barak et al. [42] defined watermarking for circuits, showing there are families of circuits for which such watermarking is impossible, and that the notion is incompatible with obfuscation even for watermarks that only succeed on some circuits. They briefly discuss how allowing "approximate implementations" may change their results. Our definition, in contrast, place these decisions in the choice of $\sim$ and the distribution $\mathcal{D}$.

We also note that many "public-key" watermarking schemes in the literature seem to target $(\mathcal{D}, t, q_E, 1, \epsilon, \delta)$ non-removability, expressed in terms of bit error rate for the watermarked message as noted above. A simple hybrid argument implies such schemes also have $(\mathcal{D}, t, q_E, q_C, q_C\epsilon, \delta)$ non-removability [43, 40]. Thus while we are not aware of any strong candidate NREs, the existence of such a scheme seems to be a natural assumption if watermarking can be feasible at all.

We note that Moulin and Wang have shown that quantization index modulation (QIM) techniques provide provably good watermarks against an adversarial *memoryless channel*. The restriction to memoryless channels, together with an assumption that the host signal is Gaussian, allows them to analytically derive the "worst possible" channel and evaluate the bit error rate for a watermark signal under a specified bound on the mean squared error introduced by the adversary. Therefore, we can view their result as showing that QIM techniques yield a non-removable embedding for the class of memoryless adversary channels. While this is a severely limited class of adversaries, it shows that our notion is realizable at least under "toy" circumstances.

Finally, the StirMark benchmark [44, 45] performs transformations such as resampling, resizing, and "jitter" in images; this benchmark is widely used to evaluate watermarks. We can capture both Moulin and Wang's result and the StirMark benchmark in our framework. If $\mathcal{C}$ is a set of object transformations, we define an attacker from class $\mathcal{C}$ to be an adversary who can only create objects via sampling from $\mathcal{D}$, queries to oracles, and applying transformations from $\mathcal{C}$ to objects he has already created. Then it is a straightforward extension of our results to show that if there is an NRE that is secure against all attackers from class $\mathcal{C}$, there is a strong watermarking scheme that is secure against all attackers from $\mathcal{C}$.

4.1 Building Strong Watermarks from Embeddings

We now show how to build ideal watermarking schemes from non-removable embeddings, digital signature schemes, and a trusted third party (TTP). The main benefit of our scheme is that the TTP need not be present during watermark detection; anyone can check whether an object is marked without needing to contact the TTP in a wide variety of cases. Our scheme requires digital signatures in addition to a TTP because the underlying embeddings are not assumed secure against insertion of watermarks or copy attacks. The nonremovable embedding is necessary to allow offline detection, because otherwise an adversary could remove any metadata that might be attached to an object as a mark.

The TTP has well-known public keys and provides two services over authenticated channels: $\mathsf{Register}(O, K, x)$ picks a unique identifier i, checks that $x = \mathsf{Encrypt}(K, O)$, and returns $(i, \mathsf{Sig}_{TTP}(i, x))$; $\mathsf{Retrieve}(i)$ returns the x associated with i if any exists, or $\perp$ otherwise; we assume that neither call returns unless a correctly signed response is received. We require that parties who execute Mark can communicate with the TTP as necessary. Note, however, that $\mathsf{Retrieve}$ is implemented in a semi-offline manner. Because unique identifiers are assigned in ascending order, the TTP publishes a signed list, $\mathsf{TTPList}$, of all (i, x) pairs each day, $\mathsf{Retrieve}(i; \mathsf{TTPList})$ only needs to contact the TTP if $i > \mathsf{TTPList.length}$. Standard measures (such as substituting a zero-knowledge proof of knowledge of (O, K) for (O, K); maintaining an ordered, signed $\mathsf{TTPList}$; checking for consistency of TTP lists between updates; *et cetera*) can be taken to reduce the level of trust required in the TTP; we omit them for clarity of presentation, and because they do not affect the security proof.

Now let $\mathcal{E} = (\mathsf{Embed}, \mathsf{Extract}, \mathsf{EMGen})$ be an embedding; and let $\mathcal{SE} = (\mathsf{Encrypt}, \mathsf{Decrypt})$ be a symmetric encryption scheme. We then define a new watermarking scheme $\mathcal{W}_{\mathcal{E}} = (\mathsf{WMGen}_{\mathcal{E},\mathcal{SE}}, \mathsf{Mark}_{\mathcal{E},\mathcal{SE}}, \mathsf{Detect}_{\mathcal{E},\mathcal{SE}})$

[4] We let $\mathsf{Embed}(O, x) = \mathtt{encode}(x)$ and $\mathsf{Extract}(O) = \mathtt{decode}(O)$. If the code's minimum distance is 2δ then clearly any distortion by distance δ or less will result in extraction of the "embedded" message, but the worst-case distortion of this procedure is the maximum possible distance between two objects in $\mathcal{M}$.

```
Algorithm Mark_E((z, z', K), O)
x ← Encrypt(K, O)
(i, σ) ← Register(O, K, x)
O' ← Embed(z, O, (i, σ))
return O'

Algorithm WMGen_E(1^k)
(z, z') ← EMGen(1^k)
K ←_R {0, 1}^k
return (z, z', K)
```

```
Algorithm Detect_E((z, z', K), O*; TTPList):
if (Extract(z', O*) =⊥) then return false
(i*, σ*) ← Extract(z', O*)
x* ← Retrieve(i*; TTPList)
O ← Decrypt(K, x*)
if (x* =⊥ or O =⊥ or Ver_TTP((i*, x*), σ*) = false)
      then return false
if Embed(z, O, (i*, σ*)) ~ O*
      then return true
else return false
```

Fig. 5. Pseudocode for $\mathsf{WMGen}_{\mathcal{E}}$, $\mathsf{Mark}_{\mathcal{E}}$, and $\mathsf{Detect}_{\mathcal{E}}$

.

as shown in Figure 5. $\mathsf{Mark}(O)$ encrypts O, registers the ciphertext with the TTP, and embeds the TTP's identifier and signature in O. $\mathsf{Detect}(O; \mathsf{TTPList})$ extracts the TTP identifier and signature, retrieves the associated ciphertext, and checks that O is close to the result of Embed applied to the plaintext.

The main result of this section is that if the underlying embedding is non-removable, then the scheme $\mathcal{W}_{\mathcal{E}}$ satisfies our notion of strong watermarking. Formally, we can state the following theorem, whose proof is in Appendix A.2.

Theorem 1. *Suppose $\mathcal{E}$ is a $(\mathcal{D}, t_E, q_{EM}, q_{EC}, \epsilon_E, \delta)$-secure non-removable embedding, $S = (\mathsf{SGen}, \mathsf{Sig}, \mathsf{Ver})$ is (t_S, q_S, ϵ_S)-existentially unforgeable under chosen message attack, and $\mathcal{SE} = (\mathsf{Encrypt}, \mathsf{Decrypt})$ is $(t, q_{en}, \epsilon_{en})$ left-or-right secure under chosen plaintext attack. Then $\mathcal{W}_{\mathcal{E}}$ is a $(t', q_M, q_D, q_C, \epsilon', \delta)$-strong watermarking scheme, where $\epsilon' = 2\epsilon_S + \epsilon_{en} + \epsilon_E$, $q_M + q_C \leq \min(q_{en}, q_S)$, $q_M \leq q_{EM}$, and $q_C \leq q_{EC}$.*

Remarks. We note that the scheme as written requires the Embed procedure to be deterministic; this is without loss of generality because the shared symmetric key between Mark and Detect can include a seed for a pseudorandom generator that is used to generate the random bits used by Embed in a deterministic way without changing the security properties of the scheme.

We also note that if the distribution $\mathcal{D}$ has Shannon entropy less than k – the length of strings embedded by $\mathcal{E}$ – then in principle the TTP can be removed from this scheme. In this case, the marking scheme first losslessly compresses the object O into a short string x of length less than k; the string x is then encrypted and authenticated using standard cryptographic techniques to get a ciphertext c which is embedded into O. The detection scheme recovers c, checks it for authenticity and if it passes, decrypts c to obtain x, then expands x to the original object O before comparing it to the input object. Thus our TTP can be seen as implementing a compression algorithm for unknown or incompressible distributions $\mathcal{D}$.

5 Strengthening Watermarks by Composition

Suppose we are given a watermarking scheme with known attacks that succeed at insertion or removal of a watermark with high probability, for example 90%, but retains some weak sense of security, in that it is not known how to defeat it with probability 1. In this section, we show that this sense of security is essentially enough for strong watermarking. Given an offline watermarking scheme W that satisfies two weak properties, we can construct an (offline) strong watermarking scheme in the sense of Section 3. The first property is that the scheme is secure in this weak sense – every adversary fails to defeat the scheme with some constant probability. The second property is that marking an object many times preserves some similarity to the original.

As mentioned previously, we believe this results has both positive and negative applications. Many of the heuristic watermarking schemes in the literature are broken, but frequently the known attacks do not succeed with probability 1. Thus applying our amplification scheme could heuristically create schemes which are, in some sense, secure "against known attacks." On the other hand, our results show that in order to rule out even weakly secure watermarking schemes for a given metric and distribution, it is sufficient to concentrate on showing the impossibility of a strong watermarking scheme.

5.1 Weakly secure watermarking schemes

Our scheme will work by applying the Mark function to its own output several times. Because our security notions depend on the probability distribution on the inputs to Mark, we will need some assumption on the distribution of the outputs of Mark. The strongest assumption is that these distributions are identical, but in general this amounts to assuming that Mark is the identity function. Thus, instead, we assume that the (weak) security of a watermark holds even if we make some small distortions to an object before marking it. Formally, we say that a randomized algorithm D is a (t, r)-*perturbation of* $\mathcal{D}$ if D runs in time t and $\Pr[O \leftarrow \mathcal{D}; O' \leftarrow D(O) : d_{\mathcal{M}}(O, O') > r]$ is negligible. We will say that our watermarking schemes are weakly secure for $\mathcal{D}$ if they are weakly secure for any (t, r)-perturbation of $\mathcal{D}$.

(Weak) security against removal. We define the removal advantage of an adversary against a watermarking scheme to be the probability that an adversary can produce, given a watermarked object drawn from a (t, r)-perturbation of $\mathcal{D}$, a similar object that is not marked. Formally, define

$$\mathbf{Adv}^{rm}_{W,D}(\mathcal{A}) = \Pr[K \leftarrow W.\mathsf{WMGen}(1^k); O \leftarrow \mathcal{D}; O' \leftarrow W.\mathsf{Mark}_K(D(O)); \\ O'' \leftarrow \mathcal{A}(O') : \quad W.\mathsf{Detect}_K(O'') = \mathsf{false} \wedge O'' \sim_\delta O'] \ .$$

Then, we say that a watermark $W$is $(t, \epsilon_{rm}, \delta, \mathcal{D}, r)$-*secure against removal* if for every time-t adversary $\mathcal{A}$, and every (t, r)-perturbation D of $\mathcal{D}$, $\mathbf{Adv}^{rm}_{W,D}(\mathcal{A}) \le \epsilon_{rm}$. Informally, this definition says that every adversary who runs in time at most t fails to remove the watermark of an object drawn from a (t, r)-perturbation of $\mathcal{D}$ with probability at least $1 - \epsilon_{rm}$.

We remark that this experiment captures the intuitive notion of trying to remove a watermark without damaging some challenge object, a common goal of attacks on watermarking schemes found in the literature. We also note that the goal of our scheme is to strengthen a watermark with only *constant* security against removal – meaning that we explicitly allow a watermarking scheme that can be removed, say, 99% of the time.

(Weak) Security against insertion. We informally define the insertion advantage of an adversary against a watermarking scheme to be the probability that an adversary can produce, given a single watermarked object, another watermarked object. Formally, define

$$\mathbf{Adv}^{ins}_{W,D}(\mathcal{A}) = \Pr[K \leftarrow \mathsf{WMGen}(1^k); O \leftarrow \mathcal{A}(1^k); O' \leftarrow W.\mathsf{Mark}_K(O); \\ O'' \leftarrow \mathcal{A}(O') : \quad W.\mathsf{Detect}_K(O'') = \mathsf{true} \wedge O'' \not\sim_\delta O'] \ .$$

Then, we say that a watermark W is $(t, \epsilon_{ins}, \delta)$-*secure against insertion* if for every time-t adversary $\mathcal{A}$, $\mathbf{Adv}^{ins}_{W,D}(\mathcal{A}) \le \epsilon_{ins}$. Informally, this definition says that every adversary who runs in time t must fail to produce a (new) watermarked object with probability at least $1 - \epsilon_{ins}$. We remark that security against insertion is essentially an adversarial notion of the "false positive rate" of a watermark [2, 27]. We can now state the main result of this section; the proof depends on several additional results proved in the remainder of the section:

Theorem 2. *Suppose there exists a watermarking scheme W such that:*

- *W is ρ-preserving;*
- *W is both $(t, \epsilon_{rm}, \delta, \mathcal{D}, k^{O(1)}\rho)$-secure against removal and $(t, \epsilon_{ins}, \delta)$-secure against insertion; and*
- *$\epsilon_{rm}, \epsilon_{ins}$ are constants such that $4\epsilon_{ins} \lg \frac{1}{\epsilon_{rm}} < 1$; and $t = k^{\omega(1)}$*

Then there exists a $(\mathcal{D}, t', q_M, 1, q_D, \nu, \delta)$-strong watermarking scheme W', where $t' = k^{\omega(1)}$ and $\nu = 1/k^{\omega(1)}$. The scheme W' is $k^{O(1)}\rho$-preserving.

Proof. The new watermark W' is constructed from W using the techniques developed in the remainder of this section: first the "alternating" composition ALT_ℓ with $\ell = O(\lg k)$ levels, from Section 5.3 is applied to W. By repeated application of Theorem 3 the resulting scheme $\mathsf{S}(W)$ is ν-secure against removal and insertion, for negligible ν. Lemma 1 implies that this scheme is also a $(\mathcal{D}, t', q_M, 1, q_D, \nu, \delta)$-strong watermark, for $q_M + q_D = 1$. To achieve arbitrary q_M and q_D, we construct the scheme $\mathsf{S}'(W)$ described in Section 5.4 with $m = q_M + q_D$. By Theorem 4 the resulting scheme is a $(\mathcal{D}, t', q_M, 1, q_D, \nu, \delta)$-strong watermark.

5.2 Single-Property Amplification.

Let $\mathbb{K} = (K_1, K_2, \ldots, K_m)$ be a set of independently chosen secret keys. We define

$$\mathsf{Mark}^W_{\mathbb{K}}(O) := W.\mathsf{Mark}_{K_m}(W.\mathsf{Mark}_{K_{m-1}}(\ldots W.\mathsf{Mark}_{K_1}(O)\ldots)) \ ,$$

i.e. $\mathsf{Mark}^W_{\mathbb{K}}$ is the sequential marking of an object O with each secret key in the vector $\mathbb{K}$. We now have two choices for defining the $\mathsf{Detect}^W_{\mathbb{K}}(O')$ algorithm, each resulting in a different watermarking scheme. Define the schemes as follows:

$$\mathsf{AND}(m, W).\mathsf{Detect}_{\mathbb{K}}(O') = \bigwedge_{1 \le i \le m} W.\mathsf{Detect}_{K_i}(O')$$
$$\mathsf{OR}(m, W).\mathsf{Detect}_{\mathbb{K}}(O') = \bigvee_{1 \le i \le m} W.\mathsf{Detect}_{K_i}(O')$$

Intuitively, we expect that $\mathsf{AND}(m, W)$ will improve the insertion security of watermark W while impeding the removal security. This is because to insert a watermark one must insert m copies of W, while to delete a watermark one need only delete 1 out of m. Likewise, we intuitively would expect that $\mathsf{OR}(m, W)$ will decrease the insertion security while increasing the removal security. We can write this formally in the following theorem, whose proof is in Appendix A.3.

Theorem 3. *Let W be ρ-preserving, $(t, \epsilon_{ins}, \delta)$-secure against insertion, and $(t, \epsilon_{rm}, \delta, \mathcal{D}, r)$-secure against removal. Then:*

(a) $\mathsf{OR}(m, W)$ *is* $(t', m\epsilon_{ins}, \delta - m\rho)$ *secure against insertion.*
(b) $\mathsf{AND}(m, W)$ *is* $(t', m\epsilon_{rm}, \delta - m\rho, \mathcal{D}, r - m\rho)$ *secure against removal.*

Where $t' = t - mT_M - O(1)$ *if* T_M *is the time to mark an object. Furthermore, for any* $q(k) \in k^{O(1)}$*,*

(c) $\mathsf{AND}(m, W)$ *is* $(t', \epsilon^m_{ins} + 1/q, \delta - m\rho)$ *secure against insertion.*
(d) $\mathsf{OR}(m, W)$ *is* $(t', \epsilon^m_{rm} + 1/q, \delta - m\rho, \mathcal{D}, r - m\rho)$ *secure against removal.*

Where $t' = t/poly(q, m)$*.*

5.3 Simultaneous Amplification

Let W be a watermarking scheme with key space K and define the scheme $\mathsf{ALT}(W)$ with key space K^4 by $\mathsf{ALT}(W) = \mathsf{AND}(2, \mathsf{OR}(2, W))$. Then by the previous theorem, if W is $(k^{\omega(1)}, c/2, \delta, \mathcal{D}, r)$ secure against removal and $(k^{\omega(1)}, d/4, \delta)$ secure against insertion, then $\mathsf{ALT}(W)$ is $(k^{\omega(1)},\ c^2/2, \delta - 4\rho, \mathcal{D}, r - 4\rho)$-secure against removal and $(k^{\omega(1)},\ d^2/4, \delta - 4\rho)$-secure against insertion. If we define the scheme $\mathsf{ALT}_\ell(W)$ by $\mathsf{ALT}_1(W) = \mathsf{ALT}(W)$ and $\mathsf{ALT}_\ell(W) = \mathsf{ALT}(\mathsf{ALT}_{\ell-1}(W))$, we see that $\mathsf{ALT}_\ell(W)$ is $(k^{\omega(1)},\ d^{2^\ell}/4, \delta - 4^\ell\rho)$-secure against insertion and $(k^{\omega(1)},\ c^{2^\ell}/2, \delta - 4^\ell\rho, \mathcal{D}, r - 4^\ell\rho)$-secure against removal, for $\ell = O(\log k)$. By setting $\ell = \lceil \log k \rceil$ and letting $\mathsf{S}(W) = \mathsf{OR}(2, \mathsf{ALT}_\ell(W))$ we obtain a scheme that inserts $poly(k)$ marks such that any $poly(k)$-time adversary has negligible advantage for both removal and insertion, if the original scheme is weakly secure against (for example) subexponential time adversaries.

Intuitively, we can think of this scheme as building a tree of marking schemes over the object O to be marked. By building the tree appropriately, alternating AND and OR at each level, we can reduce both the insertion and deletion probabilities for the resulting detection scheme. Each leaf of the tree corresponds to an independently keyed insertion of a watermark. Suppose we have a depth t tree comprising 2^t independent keys. The top gate, an OR, will recursively compute $\mathsf{AND}.\mathsf{Detect}(O, k[1]...k[2^{t-1}])$ and $\mathsf{AND}.\mathsf{Detect}(O, k[2^{t-1}]...k[2^t])$ and return true if at least one recursive branch returns true. OR is defined analogously. Alternatively, from the bottom-up view, there is one object in which we may have embedded $n = 2^t$ marks; we check if each mark is present and then compute a formula based on these truth values to decide whether the composed mark is present.

We note that the full alternating binary tree only exponentially reduces the insertion and removal probabilities if we start with $\epsilon_{rm} < 1/2$ and $\epsilon_{ins} < 1/4$. For many watermarking schemes in the literature, however, we might expect that the insertion probability is low, say $\epsilon_{ins} < 1/100$, while the removal probability is high, say $\epsilon_{rm} = 0.9$. In this case, we can make the lowest level of the tree consist of an OR of 20 marks to get $\epsilon'_{rm} = 1/e^2 < 1/2$ and $\epsilon'_{ins} < 1/5$. We can then build a binary tree on top of the resulting watermark.

It remains to show that the scheme $\mathsf{S}(W)$ is *correct*, i.e. that $\mathsf{S.Detect}_{\mathbb{K}}$ $(\mathsf{S.Mark}_{\mathbb{K}}(\mathcal{D})) = \mathsf{true}$ except with negligible probability. Notice, however, that $\mathsf{S.Detect}$ returns true if either its left branch or its right branch return true. But the insertion of the marks in the right branch is just one particular instance of an adversary (against the left branch) that returns an output that is distorted by distance at most $4^\ell \rho$ from its input, so if $\delta > 4^\ell \rho$, the probability that this "adversary" succeeds in removing the mark inserted by the left branch is negligible.

5.4 Strong watermark security from insertion and removal security.

Notice that the definition of $(t, \epsilon_{ins}, \delta)$ security against insertion implies $(\mathcal{D},\ t,\ 1,\ 1,\ 0,\ \epsilon_{ins},\ \delta)$-strong watermark security: any strong watermark adversary $\mathcal{A}$ who makes one Mark^* query and one Detect^* query can be converted into a weak insertion adversary $\mathcal{B}$: $\mathcal{B}(1^k)$ simply runs $\mathcal{A}$ until $\mathcal{A}$ makes a query to Mark^*, say O, and outputs O; $\mathcal{B}(O')$ returns O' to $\mathcal{A}$ and outputs the object O'' that $\mathcal{A}$ queries to Detect^*. Since the list chalns is empty, submitting an unmarked O'' will give $b = \mathsf{false}$ and $b' = \mathsf{false}$, so $\mathcal{A}$ can only win by "inserting" a watermark. Additionally satisfying $(t, \epsilon_{rm}, \delta, \mathcal{D}, r)$-security against removal implies $(D(\mathcal{D}), t, 0, 1, 1, \epsilon, \delta)$ strong watermark security for any D that perturbs $\mathcal{D}$ by at most r, because an adversary who makes only a single query $O' \leftarrow \mathsf{Challenge}^*(D(\mathcal{D}))$ can only win by querying $\mathsf{Detect}^*(O'')$ such that:

- $O'' \sim O'$ and $\mathsf{Detect}_K(O'') = \mathsf{false}$; if this happens with probability greater than ϵ_{rm} then the removal security of the scheme is contradicted.
- $d(O'', O') > \delta$ and $\mathsf{Detect}_K(O'') = \mathsf{true}$; if this happens with probability greater than ϵ_{ins} then the insertion security is violated: an insertion adversary can always draw his challenge object $O' \leftarrow D(\mathcal{D})$.

This observation leads to the following lemma:

Lemma 1. *If W is $(t, \epsilon_{ins}, \delta)$-secure against insertion and $(t, \epsilon_{rm}, \delta, \mathcal{D}, r)$-secure against removal then W is a $(D(\mathcal{D}), t, q_M, 1, q_C, \epsilon_{ins} + \epsilon_{rm}, \delta)$-strong watermarking scheme, for any distortion function $D \in \mathrm{TIME}(t)$ that perturbs $\mathcal{D}$ by distance at most r, and any $q_C \leq 1 - q_M$.*

Suppose that we extend the definition of a strong watermark to allow Mark to maintain a local state. Then we can generically increase the number of (mark and challenge) queries we are secure against by a factor of n while also increasing the running time of Detect by a factor of n as follows. We require that Mark'_K keeps a count, i, of the number of objects it has marked (say modulo n). When $\mathsf{Mark}'_K(O)$ marks a new object, it computes the entire set of keys to use as $\mathbb{K}_i = F_K(i)$, where F is a pseudorandom function of the appropriate output size, and then calls $\mathsf{Mark}_{\mathbb{K}_i}(O)$. Then in $\mathsf{Detect}'_K(O)$ we try $\mathbb{K} = F_K(1), F_K(2) \ldots F_K(n)$ and output true if any of these watermarks is detected. This increases the insertion probability by at most a factor of n. We make this more formal in the following theorem, whose proof is in Appendix A.3.

Theorem 4. *Let $W = (\mathsf{Mark}, \mathsf{Detect})$ be a $(\mathcal{D}, t, q_M, 1, 1 - q_M, \epsilon_{wm}, \delta)$-strong watermarking scheme and let $W' = (\mathsf{Mark}', \mathsf{Detect}')$ be a watermarking scheme with the stateful Mark' algorithm described above, and let F be a (t, n, ϵ_{prf})-pseudorandom function. Then W' is a $(\mathcal{D}, t, q_M, 1, n - q_M, n\epsilon_{wm} + \epsilon_{prf}, \delta)$-strong watermarking scheme.*

6 Conclusions

In this paper we have initiated the scientific study of complexity-based security of watermarking schemes. We define a notion of watermarking security based on comparison to an ideal scheme, and give evidence that this is the right notion of security for watermarks in two ways. First, we show that security in our sense implies previous definitions of security, while the converse is not true. Second, we have shown how to construct a watermark which is secure in our sense from several weaker primitives, which seem to capture the goals of research in watermarking primitives. Our intent is not to introduce new watermarking protocols, but to suggest that security in the "strong watermark" sense is the "right definition" - if secure watermarks (in any sense) are feasible at all, then so are strong watermarking schemes. A key question left open by our work, therefore, is the construction of similarity-preserving strong watermarking schemes that are provably-secure under standard cryptographic assumptions; even a construction for a contrived metric space would be an interesting first step in this direction.

Acknowledgements

We thank Andre Adelsbach, Luis von Ahn, Hayley Iben, Pierre Moulin, and Ahmad-Reza Sadeghi for helpful discussions. We also thank Kannan Ramchandran, Dan Schonberg, and the Berkeley Information Hiding Group for feedback on early versions of this work. Nicholas Hopper was supported by the National Science Foundation, under grant CNS-0546162. David Molnar was supported by a graduate research fellowship from the National Science Foundation.

References

1. Anderson, R.J.: Security Engineering: A Guide to Building Dependable Distributed Systems. John Wiley & Sons, Inc., New York, NY, USA (2001)
2. Cox, I., Miller, M.L., Bloom, J.A.: Digital watermarking. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2002)
3. Wong, P.W., Delp, E.J., eds.: Security and Watermarking of Multimedia Contents, Proceedings. In Wong, P.W., Delp, E.J., eds.: Security and Watermarking of Multimedia Contents. Volume 3657 of Proceedings of SPIE., SPIE (1999)
4. Wong, P.W., Delp, E.J., eds.: Security and Watermarking of Multimedia Contents II, 2000, Proceedings. In Wong, P.W., Delp, E.J., eds.: Security and Watermarking of Multimedia Contents. Volume 3971 of Proceedings of SPIE., SPIE (2000)
5. Wong, P.W., Delp, E.J., eds.: Security and Watermarking of Multimedia Contents III, 2001, Proceedings. In Wong, P.W., Delp, E.J., eds.: Security and Watermarking of Multimedia Contents. Volume 4314 of Proceedings of SPIE., SPIE (2001)
6. Delp, E.J., Wong, P.W., eds.: Security and Watermarking of Multimedia Contents IV, 2002, Proceedings. In Delp, E.J., Wong, P.W., eds.: Security and Watermarking of Multimedia Contents. Volume 4675 of Proceedings of SPIE., SPIE (2002)
7. Delp, E.J., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents V, 2003, Proceedings. In Delp, E.J., Wong, P.W., eds.: Security and Watermarking of Multimedia Contents. Volume 5020 of Proceedings of SPIE., SPIE (2003)
8. Delp, E.J., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents VI, San Jose, California, USA, January 18-22, 2004, Proceedings. In Delp, E.J., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents. Volume 5306 of Proceedings of SPIE., SPIE (2004)
9. Delp, E.J., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents VII, San Jose, California, USA, January 17-20, 2005, Proceedings. In Delp, E.J., Wong, P.W., eds.: Security, Steganography, and Watermarking of Multimedia Contents. Volume 5681 of Proceedings of SPIE., SPIE (2005)
10. Anderson, R.J., ed.: Information Hiding, First International Workshop, Cambridge, U.K., May 30 - June 1, 1996, Proceedings. In Anderson, R.J., ed.: Information Hiding. Volume 1174 of Lecture Notes in Computer Science., Springer (1996)
11. Aucsmith, D., ed.: Information Hiding, Second International Workshop, Portland, Oregon, USA, April 14-17, 1998, Proceedings. In Aucsmith, D., ed.: Information Hiding. Volume 1525 of Lecture Notes in Computer Science., Springer (1998)
12. Barni, M., Herrera-Joancomartí, J., Katzenbeisser, S., Pérez-González, F., eds.: Information Hiding, 7th International Workshop, IH 2005, Barcelona, Spain, June 6-8, 2005, Revised Selected Papers. In Barni, M., Herrera-Joancomartí, J., Katzenbeisser, S., Pérez-González, F., eds.: Information Hiding. Volume 3727 of Lecture Notes in Computer Science., Springer (2005)

13. Pfitzmann, A., ed.: Information Hiding, Third International Workshop, IH'99, Dresden, Germany, September 29 - October 1, 1999, Proceedings. In Pfitzmann, A., ed.: Information Hiding. Volume 1768 of Lecture Notes in Computer Science., Springer (2000)
14. Moskowitz, I.S., ed.: Information Hiding, 4th International Workshop, IHW 2001, Pittsburgh, PA, USA, April 25-27, 2001, Proceedings. In Moskowitz, I.S., ed.: Information Hiding. Volume 2137 of Lecture Notes in Computer Science., Springer (2001)
15. Petitcolas, F.A.P., ed.: Information Hiding, 5th International Workshop, IH 2002, Noordwijkerhout, The Netherlands, October 7-9, 2002, Revised Papers. In Petitcolas, F.A.P., ed.: Information Hiding. Volume 2578 of Lecture Notes in Computer Science., Springer (2003)
16. Fridrich, J.J., ed.: Information Hiding, 6th International Workshop, IH 2004, Toronto, Canada, May 23-25, 2004, Revised Selected Papers. In Fridrich, J.J., ed.: Information Hiding. Volume 3200 of Lecture Notes in Computer Science., Springer (2004)
17. Gollmann, D.: Computer security. John Wiley & Sons, Inc., New York, NY, USA (1999)
18. Wikipedia: Digital watermarking — wikipedia, the free encyclopedia (2006) [Online; accessed 31-July-2006].
19. Barak, B., Goldreich, O., Impagliazzo, R., Rudich, S., Sahai, A., Vadhan, S.P., Yang, K.: On the (im)possibility of obfuscating programs. In Kilian, J., ed.: CRYPTO. Volume 2139 of Lecture Notes in Computer Science., Springer (2001) 1–18
20. Lynn, B., Prabhakaran, M., Sahai, A.: Positive results and techniques for obfuscation. In Cachin, C., Camenisch, J., eds.: EUROCRYPT. Volume 3027 of Lecture Notes in Computer Science., Springer (2004) 20–39
21. Wee, H.: On obfuscating point functions. In Gabow, H.N., Fagin, R., eds.: STOC, ACM (2005) 523–532
22. Cachin, C.: An information-theoretic model for steganography. In Aucsmith, D., ed.: Information Hiding. Volume 1525 of Lecture Notes in Computer Science., Springer (1998) 306–318
23. Hopper, N.J., Langford, J., von Ahn, L.: Provably secure steganography. In Yung, M., ed.: CRYPTO. Volume 2442 of Lecture Notes in Computer Science., Springer (2002) 77–92
24. Dedic, N., Itkis, G., Reyzin, L., Russell, S.: Upper and lower bounds on black-box steganography. In Kilian, J., ed.: TCC. Volume 3378 of Lecture Notes in Computer Science., Springer (2005) 227–244
25. Craver, S., Memon, N., Yeo, B.L., Yeung, M.M.: Resolving rightful ownerships with invisible watermarking techniques: Limitations, attacks, and implications. IEEE J. SAC **16**(4) (May 1998) 573–586 Special issue on copyright & privacy protection.
26. Kutter, M., Voloshynovskiy, S., Herrigel, A.: The watermark copy attack. In: Proceedings of the SPIE vol. 3971, Security and Watermarking of Multimedia Contents II. (2000) 371–380
27. Adelsbach, A., Katzenbeisser, S., Veith, H.: Watermarking schemes provably secure against copy and ambiguity attacks. In: DRM '03: Proceedings of the 2003 ACM workshop on Digital rights management, ACM Press (2003) 111–119
28. Dittmann, J., Katzenbeisser, S., Schallhart, C., Veith, H.: Provably secure authentication of digital media through invertible watermarks. Cryptology ePrint Archive, Report 2004/293 (2004) `http://eprint.iacr.org/`.
29. Li, Q., Chang, E.C.: On the possibility of non-invertible watermarking schemes. In: Information Hiding Workshop, Springer-Verlag LNCS 3200. (2004)
30. Goldwasser, S., Micali, S., Rivest, R.L.: A digital signature scheme secure against adaptive chosen-message attacks. SIAM J. Comput. **17**(2) (1988) 281–308
31. Bellare, M., Desai, A., Jokipii, E., Rogaway, P.: A concrete security treatment of symmetric encryption. In: FOCS '97: Proceedings of the 38th Annual Symposium on Foundations of Computer Science (FOCS '97), Washington, DC, USA, IEEE Computer Society (1997) 394
32. Goldreich, O., Goldwasser, S., Micali, S.: How to construct random functions. J. ACM **33**(4) (1986) 792–807
33. Li, L., Pan, Z., Zhang, M., Ye, K.: Watermarking subdivision surfaces based on addition property of fourier transform. In: GRAPHITE '04: Proceedings of the 2nd international conference on Computer graphics and interactive techniques in Australasia and South East Asia, New York, NY, USA, ACM Press (2004) 46–49
34. Meerwald, P., Uhl, A.: A survey of wavelet-domain watermarking algorithms. In Wong, P.W., Delp, E.J., eds.: Proceedings of SPIE, Electronic Imaging, Security and Watermarking of Multimedia Contents III. Volume 4314., San Jose, CA, USA, SPIE (January 2001)
35. Ruanaidh, J.J.O., Pereira, S.: A secure robust digital image watermark. In: International Symposium on Advanced Imaging and Network Technologies - Conference on Electronic Imaging: Processing, Printing, and Publishing in Colour. (1998)
36. Moulin, P., Wang, Y.: Improved QIM strategies for gaussian watermarking. In: International Workshop on Digital Watermarking (IWDW '05). (2005)
37. Bas, P.: A quantization watermarking technique robust to linear and non-linear valumetric distortions using a fractal set of floating quantifiers. In: Information Hiding Workshop (IHW) '05. (2005)

38. Martin, V., Chabert, M., Lacaze, B.: A spread spectrum watermarking scheme based on periodic clock changes for digital images. In: Information Hiding Workshop (IHW) '05. (2005)
39. Doerr, G., Dugelay, J.L.: A quantization watermarking technique robust to linear and non-linear valumetric distortions using a fractal set of floating quantifiers. In: Information Hiding Workshop (IHW) '05. (2005)
40. Hartung, F., Girod, B.: Fast public-key watermarking of compressed video. In: International Conference on Image Processing (ICIP'97). Volume I., Santa Barbara, California, U.S.A. (1997) 528–531
41. Micali, S., Peikert, C., Sudan, M., Wilson, D.: Optimal error correction against computationally bounded noise. In: Theory of Cryptography Conference (TCC) '05. (2005) `http://theory.lcs.mit.edu/~cpeikert/pubs/mpsw.ps`.
42. Barak, B., Goldreich, O., Impagliazzo, R., Rudich, S., Sahai, A., Vadhan, S., Yang, K.: On the (im)possibility of obfuscating programs. In: CRYPTO. (2001)
43. Wong, P.W., Memon, N.: Secret and public key image watermarking schemes for image authentication and ownership verification. IEEE Trans. Image Processing **10**(10) (October 2001) 1593–1601
44. Petitcolas, F.A., Anderson, R.J., Kuhn, M.G.: Attacks on copyright marking systems. In: Information Hiding Workshop (IHW) 1998, Springer-Verlag Lecture Notes in Computer Science 1525. (1998) 219–239
45. Petitcolas, F.A.: Watermarking schemes evaluation. IEEE Signal Processing **17**(4) (September 2000) 58–64
46. Canetti, R., Halevi, S., Steiner, M.: Hardness amplification of weakly verifiable puzzles. In: TCC 2005 : Proceedings of the 2nd Theory of Cryptography Conference. (2005)
47. Comesana, P., Prez-Freire, L., Prez-Gonzalez, F.: Fundamentals of data-hiding security and their application to spread-spectrum analysis. In: Information Hiding Workshop. (2005)
48. Pointcheval, D., Stern, J.: Security proofs for signature schemes. In: EUROCRYPT. (1996) 387–398

A Proofs

A.1 Copy and Ambiguity Attacks

We now formally state a theorem showing that strong watermarks are not vulnerable to copy attacks.

Theorem. *Suppose $\mathcal{W}$ is a $(\mathcal{D}, t, 1, 1, 0, \epsilon, \delta)$-strong watermarking scheme. Let $\mathcal{D}'$ be any distribution on $\mathcal{M}$ that can be sampled in time t_{sample}. Then $\mathcal{W}$ is $(\mathcal{D}', t', \epsilon', \delta')$-secure against copy attacks, where $t' = t - t_{sample} - O(1)$, $\epsilon' = \epsilon$, and $\delta' = \delta$.*

Proof. Suppose there exists an adversary B that runs in time at most t such that $\mathbf{Adv}^{cp}_{\mathcal{D}',\mathcal{W}}(B) > \epsilon'$. We construct an adversary A^B_{cp} that uses B as an oracle and show that $\mathbf{Adv}^{strong-wm}_{\mathcal{D},\mathcal{W}}(A^B_{cp}) > \epsilon$. We will see that this contradicts our assumption that $\mathcal{W}$ is a $(\mathcal{D}, t, 0, 1, 1, \epsilon, \delta)$-strong watermarking scheme.

We show the code for A^B_{cp} in Figure 3. By the definition of a copy attack, the call $\mathsf{Detect}^*(O_2')$ returns true with probability at least ϵ. Also by the definition of a copy attack, $O_2' \not\sim O_1$; note also that δ is the same for both the strong watermarking experiment and the copy attack experiment. At this point in the simulation, Marked contains only O_1, so therefore $O_2' \notin \mathsf{Marked}$. Therefore the A^B_{cp} query to Detect^* with O_2' causes bad to be set to true, and consequently A^B_{cp} wins the strong watermarking experiment if B wins the copy attack experiment. Finally, note that A^B_{cp} requires one query to Detect^* and one query to Mark^*.

We now state a theorem showing that strong watermarks are not vulnerable to ambiguity attacks.

Theorem. *Suppose $\mathcal{W}$ is a $(\mathcal{D}, t, 0, 2, 0, \epsilon, \delta)$- strong watermarking scheme. Suppose that $\mathcal{D}'$ can be sampled in time t_{sample}. Then $\mathcal{W}$ is $(\mathcal{D}', t', \epsilon', \delta)$-secure against ambiguity attacks, with $t' = t - t_{sample} - O(1)$ and $\epsilon' = \epsilon$.*

Proof. Suppose there exists an adversary B that runs in time at most t' such that $\mathbf{Adv}^{amb}_{\mathcal{D}',\mathcal{W}}(B) > \epsilon'$. We construct an adversary A^B_{amb} that uses B as an oracle and show that $\mathbf{Adv}^{strong-wm}_{\mathcal{D},\mathcal{W}}(A^B_{amb}) > \epsilon$ with two queries to Detect^*. This contradicts our assumption that $\mathcal{W}$ is a $(\mathcal{D}, t, 0, 2, 0, \epsilon)$-strong watermarking scheme.

We show the code for A^B_{amb} in Figure 3. At line 2, the lists chalns and Marked are both empty, so if $\mathsf{Detect}(K, O_1)$ returns true, we will have set $\mathsf{bad} \leftarrow \mathsf{true}$. Let $B2$ denote this event, i.e., the call to Detect^* in line 2 sets $\mathsf{bad} \leftarrow \mathsf{true}$. Note that conditioned on $\overline{B2}$, the object input to B in line 3 by A and the object input to B in $\mathbf{Exp}^{amb}_{\mathcal{D},\mathcal{W}}(B)$ have the same distribution. By the definition of the ambiguity experiment, when B

wins we have $O_1 \sim O'_1$ and $\mathsf{Detect}(K, O'_1) = \mathsf{true}$, yet the lists Marked and chaIns are both empty. Therefore, the call to $\mathsf{Detect}^*(O'_1)$ sets bad to true if B wins the ambiguity experiment. Thus we have

$$\begin{aligned}\mathbf{Adv}^{swm}_{\mathcal{D},\mathcal{W}}(A^B_{amb}) &= \Pr[B2] + \Pr[\mathsf{bad} = \mathsf{true}|\overline{B2}]\Pr[\overline{B2}] \\ &\geq \Pr[B2] + \epsilon(1 - \Pr[B2]) \\ &\geq \epsilon \Pr[B2] + \epsilon(1 - \Pr[B2]) = \epsilon\end{aligned}$$

To finish the proof, note that A^B_{amb} requires only two queries to Detect^*.

A.2 Strong Watermarks from Non-Removable Embeddings

We now give the proof for our theorem stating that our canonical scheme built on non-removable embeddings is a strong watermark scheme.

Theorem. *Suppose $\mathcal{E}$ is a $(\mathcal{D}, t_E, q_{EM}, q_{EC}, \epsilon_E, \delta)$-secure non-removable embedding, $S = (\mathsf{SGen}, \mathsf{Sig}, \mathsf{Ver})$ is (t_S, q_S, ϵ_S)-existentially unforgeable under chosen message attack, and $\mathcal{SE} = (\mathsf{Encrypt}, \mathsf{Decrypt})$ is $(t, q_{en}, \epsilon_{en})$ left-or-right secure under chosen plaintext attack. Then $\mathcal{W}_{\mathcal{E}}$ is a $(t', q_M, q_D, q_C, \epsilon', \delta)$-strong watermarking scheme, where $\epsilon' = 2\epsilon_S + \epsilon_{en} + \epsilon_E$, $q_M + q_C \leq \min(q_{en}, q_S)$, $q_M \leq q_{EM}$, and $q_C \leq q_{EC}$.*

Proof. (Sketch) Suppose not. Therefore there exists an adversary A_{strong} against the watermarking scheme such that the advantage of A_{strong} in the strong watermarking experiment is more than ϵ'. There are two cases.

1. There exists a j such that for the object O_j that was an argument of A_{strong}'s j'th query to Detect^*, we have $b = \mathsf{true}$ and $B' = \mathsf{false}$.
2. There exists a j such that for the object O_j that was an argument of A_{strong}'s j'th query to Detect^*, we have $b = \mathsf{false}$ and $B' = \mathsf{true}$, i.e. there is an object $O'_i \in$ chaIns such that $O'_i \sim O_j$.

Let E_1 be the event that case (1) occurs and E_2 be the event that case (2) occurs and case (1 does not. We see that $\mathbf{Adv}^{strong}_{\mathcal{W},\mathcal{D}}(A_{strong}) = \Pr[E_1] + \Pr[E_2]$. The main idea of the proof is that we construct a sequence of adversaries $A_{sig,1}$, $A_{sig,2}$, A_{enc}, and A_{nre}. We show that $A_{sig,1}$ forges a TTP signature with probability $\Pr[E_1]$. Then we show that $\Pr[E_2]$ is at most the sum of $A_{sig,2}$'s probability of forgery plus A_{enc}'s left-or-right advantage against the symmetric encryption scheme $\mathcal{SE}$ plus q_D times A_{nre}'s advantage against the non-removable embedding $\mathcal{E}$. We conclude that $\mathbf{Adv}^{strong-wm}_{\mathcal{W}_{\mathcal{E}},\mathcal{D}}(A_{strong}) \leq 2\epsilon_S + \epsilon_{en} + q_D\epsilon_E$.

We now describe the adversary $A_{sig,1}$ that mounts a chosen message attack on the signature scheme S. The adversary $A_{sig,1}$ runs $\mathsf{EMGen}(1^k)$ to obtain a key pair (z, z') for $\mathcal{E}$, picks an encryption key $K \leftarrow \{0,1\}^k$ then simulates $\mathcal{W}_{\mathcal{E}}$ for the adversary A_{strong}. Whenever A_{strong} makes a Mark^* query on an object O, A^1_{sig} creates the appropriate (i, O) pair, adds $(i, \mathsf{Encrypt}(K, O))$ to its TTP list, then uses its adaptive chosen message oracle to sign the pair and uses its embedding key to return an $(O'; \mathsf{TTPList})$ with the proper (i, σ) pair embedded. When A_{strong} queries Detect^* with an object $(O_j; \mathsf{TTPList}^*$ that causes bad to be set to true, $A_{sig,1}$ sets $(i, \sigma) = \mathsf{Extract}(z', O_i)$, retrieves (i, x) from $\mathsf{TTPList}^*$, and returns $(i, \mathsf{Decrypt}(K, x)), \sigma$ as its forgery against S.

Now suppose E_1 occurs. Therefore, there exists some object O_i that causes Detect to return true, but O_i is not similar to any object previously queried to Mark^*. By the definition of Detect , however, O_i is similar to object i on $\mathsf{TTPList}^*$, and furthermore (i^*, O^*) is properly signed by S. Therefore, we see that if E_1 occurs, (i^*, O^*) cannot have been previously queried to the signing oracle, yet σ^* passes signature verification with VK. Therefore $A_{sig,1}$ succeeds at producing a forgery if E_1 occurs.

Now suppose E_2 occurs. There are two further cases:

(a) The object $O'_i \sim O_j$ such that O_i was generated by a previous challenge is not the object associated to i^* in TTPList.
(b) $\mathsf{Extract}(z, O_j)$ returns the error value $\perp$ or a pair (i^*, σ^*) such that $\mathsf{Ver}_{TTP}((i^*, O_{i^*}), \sigma^*) = \mathsf{false}$.

Let us refer to the event that case (2a) occurs by the event E_{2a} and the event that case (2b) occurs and neither case (2a) nor case (1) occur by the event E_{2b}, with $\Pr[E_2] = \Pr[E_{2a}] + \Pr[E_{2b}]$. The adversary $A_{sig,2}$ is similar to the adversary $A_{sig,1}$. We see that if E_{2a} occurs, then $A_{sig,2}$ succeeds in outputting a forgery just as $A_{sig,1}$ does.

Next we describe a hybrid experiment H and show a left-or-right encryption adversary A_{enc} who makes $q_C + q_M$ queries and breaks $\mathcal{SE}$ with advantage $(\Pr_H[E_{2b}] - \Pr_{strong}[E_{2b}])$. The hybrid experiment H works exactly like the strong watermarking experiment with $\mathcal{W}$, *except* that when $\mathsf{Challenge}^*$ computes $\mathsf{Mark}_K(O)$ it draws a second object O' and submits $\mathsf{Encrypt}(K, O')$ to the $\mathsf{TTPList}$; and Detect^* modifies detect to substitute O for O' in line 4 of $\mathsf{Detect}_{\mathcal{E}}$. A_{enc} picks all of the necessary keys for signature and embedding, and emulates Mark^*,$\mathsf{Challenge}^*$,and Detect^*, except that every time $\mathsf{Challenge}^*$ calls $\mathsf{Mark}(O)$, A_{enc} chooses a second object $O' \leftarrow \mathcal{D}$ and replaces line 1 with $x \leftarrow LOR_K(O, O')$. A_{enc} then outputs 1 if the event E_{2b} occurs and 0 otherwise. Since the functionality of the Mark and Detect routines is maintained in this hybrid experiment, the only difference is whether LOR_K encrypts its first or second arguments; if the first argument is encrypted, $\Pr[A_{enc}^{LOR_K(0,\cdot,\cdot)} = 1] = \Pr_{strong}[E_{2b}]$, and otherwise $\Pr[A_{enc}^{LOR_K(1,\cdot,\cdot)} = 1] = \Pr_H[E_{2b}]$. Thus the advantage of A_{nre} is as claimed.

We now describe the adversary A_{nre} that breaks the non-removability of the underlying scheme if event E_{2b} occurs in the hybrid experiment H. $A_{nre}(z')$ runs the signature scheme key generator to obtain $(\mathsf{SK}, \mathsf{VK})$ for the public key signature scheme and picks a random $j \leftarrow \{1, \dots, q_D\}$. Then $A_{nre}(z')$ runs A_{strong} as a subroutine and uses its Embed^* oracle to simulate answers to Mark^* in the obvious way; A_{nre} uses its $\mathsf{Challenge}^*$ oracle to respond to A_{strong}'s challenge queries by drawing an object $O' \leftarrow \mathcal{D}$, computing $x = \mathsf{Encrypt}(K, O')$, and querying $\mathsf{Challenge}^*(x)$. $A_{nre}(z')$ uses z' to compute responses to $\mathsf{Detect}^*(O)$ queries as in the hybrid experiment H; when A_{strong} queries a $\mathsf{Detect}^*(O_j)$ query that causes event E_{2b}, $A_{nre}(z')$ returns the object O_j. Note that by the definition of event E_{2b}, O_j is similar to some object returned by A_{nre}'s $\mathsf{Challenge}^*$ oracle, but $\mathsf{Extract}(z', O_j) \neq (i^*, \sigma^*)$. Thus $\mathbf{Adv}_{\mathcal{D}}^{NRE}(A_{nre}) = \Pr_H[E_{2b}]$, which gives $\Pr_{strong}[E_{2b}] \leq \epsilon_{en} + \epsilon_E$.

A.3 Simultaneous Amplification

Theorem. *Let W be ρ-preserving, $(t, \epsilon_{ins}, \delta)$-secure against insertion, and $(t, \epsilon_{rm}, \delta, \mathcal{D}, r)$-secure against removal. Then:*

(a) $\mathsf{OR}(m, W)$ *is $(t', m\epsilon_{ins}, \delta - m\rho)$ secure against insertion.*
(b) $\mathsf{AND}(m, W)$ *is $(t', m\epsilon_{rm}, \delta - m\rho, \mathcal{D}, r - m\rho)$ secure against removal.*

Where $t' = t - mT_M - O(1)$ if T_M is the time to mark an object. Furthermore, for any $q(k) \in k^{O(1)}$,

(c) $\mathsf{AND}(m, W)$ *is $(t', \epsilon_{ins}^m + 1/q, \delta - m\rho)$ secure against insertion.*
(d) $\mathsf{OR}(m, W)$ *is $(t', \epsilon_{rm}^m + 1/q, \delta - m\rho, \mathcal{D}, r - m\rho)$ secure against removal.*

Where $t' = t/poly(q, m)$.

Proof. The proofs of statements (a) and (b) are essentially standard hybrid arguments: suppose, for example, that (b) does not hold. Then there must be some pair $\mathcal{A}, D \in \mathrm{TIME}(t - mT_M)$ such that $\mathcal{A}$ produces, given the result of $\mathsf{Mark}_{\mathbb{K}}^W(O = D(\mathcal{D}))$, an O' with $d(O', O) < \delta - m\rho$ and $Pr[(\bigwedge_i \mathsf{Detect}_{K_i}(O')) = \mathsf{false}] > m\epsilon_{rm}$. But in this case we have $E_i[\Pr[\mathsf{Detect}_{K_i}(O') = \mathsf{false}]] > \epsilon_{rm}$ and thus for some i we have a $D' = \mathsf{Mark}_{K_1,\dots,K_i}^W(D(\cdot))$ and $\mathcal{A}' = \mathcal{A}(\mathsf{Mark}_{K_{i+1},\dots,K_m}(\cdot))$ that succeed with probability at least ϵ_{rm}, while D' perturbs $\mathcal{D}$ at most $m\rho$ and $\mathcal{A}'(D'(O)) \sim O$. This gives a contradiction and thus (b) must hold. The proof of (a) is similar.

We briefly sketch the proof of (d), which closely follows the proof of [46, Lemma 1]; the proof of (c) is similar. The basic idea is that a sample $O \leftarrow D(\mathcal{D})$ together with a vector of independent keys $K_1, \dots, K_m$ form a "weakly verifiable puzzle" in that, given the keys and the adversary's input $O' = \mathsf{Mark}_{\mathbb{K}}^W(O)$, we can check that an adversary's output O'' is not marked by any of $K_1, \dots, K_m$ and is sufficiently close to O' to constitute a removal. As in [46] we can imagine a giant matrix M associated to each $(\mathcal{A}, D)$ where the columns are indexed by keys K_1 and the rows are indexed by $K_2, \dots, K_m, O$; the element indexed by

$(K_1, \ldots, K_m, O)$ has a 1 in the first position if $\mathsf{Detect}_{K'}(O'') = \mathsf{false}$ and $O'' \sim O'$, and a 1 in the second position if $\bigwedge_{2 \leq i \leq m}(\mathsf{Detect}_{K_i}(O'') = \mathsf{false})$ and $O'' \sim O'$. If (d) does not hold then for some $(D, \mathcal{A})$ the fraction of $(1, 1)$ entries is at least $\epsilon_{rm}^m + 1/q$, and this implies that either: (1) there exists some column k_1 such that at least a $\epsilon_{rm}^{m-1} + 1/q$ fraction of the entries have the form $(*, 1)$; or (2) the probability that an entry is $(1, 1)$ conditioned on $(*, 1)$ is at least ϵ. In the first case, we can find such a k_1 in time roughly $q^{O(1)}$ and inductively run the procedure with $D' = \mathsf{Mark}_{k_1}(\cdot)$. In the second case, we can randomly pick $q^O(1)$ random completions $(K_2, \ldots, K_m)$ for (K_1, O) and expect that for one of them $K_2, \ldots, K_M$ are removed; in this case, it is a "good bet" (probability ϵ) that K_1 is removed as well, giving an adversary for W. The complete proof that this strategy works is nearly identical to the proof in [46].

We now state a theorem regarding the security of the stateful construction shown in Section 5.4 and give its proof.

Theorem. *Let $W = (\mathsf{Mark}, \mathsf{Detect})$ be a $(\mathcal{D}, t, q_M, 1, 1 - q_M, \epsilon_{wm}, \delta)$-strong watermarking scheme and let $W' = (\mathsf{Mark}', \mathsf{Detect}')$ be a watermarking scheme with the stateful* Mark*' algorithm described above, where F is a (t, n, ϵ_{prf})-pseudorandom function. Then W' is a $(\mathcal{D}, t, q_M, 1, n - q_M, n\epsilon_{wm} + \epsilon_{prf}, \delta)$-strong watermarking scheme.*

Proof. The proof has two steps. The first is to consider a hybrid W' with access to a truly random function f with the same domain as F; it is easy to see that for any A, $\Pr[\mathsf{Exp}^{\mathsf{strong-wm}}_{\mathcal{D},W'(F_K)}(A) = 1] - \Pr[\mathsf{Exp}^{\mathsf{strong-wm}}_{\mathcal{D},W'(f)}(A) = 1] \leq \epsilon_{prf}$. Next we show how to convert an adversary A who makes n queries to $\mathsf{Challenge}^*$ and Mark^* against $W'(f)$ with advantage ε into an adversary B against W who makes 1 query to $\mathsf{Challenge}^*$ or Mark^* and has advantage at least ε/n. B guesses which key K_i A will succeed in removing or inserting and passes the i^{th} query made by A on to its $\mathsf{Challenge}^*$ or Mark^* oracle; for all other queries, B picks a fresh random key and responds appropriately. Whenever A wins the strong watermarking game, it is because either (1) A's query to Detect^* was unmarked and similar to some O_j returned by $\mathsf{Challenge}^*$; or (2) A's query to detect was marked by some K_j and never returned by $\mathsf{Challenge}^*$ or Mark^*. In either case, B will succeed with probability $1/n$ when A does.

B Limitations of Previous Work

Although the literature on watermarking includes several previous works on formal security definitions, these works tend to be too permissive or incomplete. We will later give a more complete discussion of issues with previous work. Here we give a short summary.

Adelsbach, Katzenbeisser, and Veith gave formal definitions of ambiguity and copy attacks, and constructions for watermarks provably secure against these attacks [27]. These definitions do not allow the adversary to mount "chosen-object" attacks, where the adversary may submit objects to be watermarked and observe their watermarked versions; in a copyright registration scenario, this attack is realistic. Further, their definitions do not formally describe what is required for the watermark to be non-removable under attack. Finally, in the Appendix we discuss issues with one of their proposed constructions.

Li and Chang give a construction of watermarks using a pseudo-random generator that are claimed secure against ambiguity attacks [29]. Their definition does not rule out attacks that remove the watermark. For example, a watermarking scheme that encrypts the low-order bits of a picture would satisfy their requirements, but the watermark can easily be removed by setting all low-order bits to 0. There is a further conceptual issue: their adversary must work with a specific challenge object O and is not allowed to return an object O' such that $O' \sim O$. As a result, their notion of security is too restrictive of the adversary.

Dittmann et al. propose definitions and constructions for secure authentication of digital media using invertible watermarks [28]. The scheme proposed there, however, appears to rely on assumptions about the watermarking scheme that are not stated in the proof of security, rendering the proof incomplete. For concreteness, we summarize the scheme and an attack that works under certain conditions in Appendix B.1. In addition to this difficulty, this particular work is a further example of the watermarking "arms race" we seek to avoid, in that the authors focus on a specific attack rather than trying to obtain a general security condition.

Comesana et al. propose a notion of watermarking security that focuses on information about the secret key leaked to the adversary, under several different notions of the adversary's view [47]. They consider a scheme secure if the mutual information between marked objects observed by the adversary and the secret key of the watermarking scheme is small. The authors themselves point out security in their sense does not necessarily mean the watermark is difficult to remove. For example, the identity map, which does not depend on the secret key, appears to be perfectly secure under their definition because the distribution of "marked" objects is independent of the secret key. Therefore, it is not clear whether security in this sense is useful for evaluating watermark schemes. The security notions we present here are closer to what Comesana et al. and others in the watermarking literature call "robustness" in that the focus is on whether the mark is detectable after an adversarial transformation of the marked object.

B.1 Dittmann et al.'s scheme

Dittmann *et al.* are concerned with protecting the authenticity of an object via a watermark, without compromising its "quality". To that end, they define security as the inability, given an oracle Protect, to produce an object O and its protected version $\overline{O}$, without querying $\mathsf{Protect}(O)$. They assume the existence of procedures Join and Separate such that for an object O, $\mathsf{Separate}(O)$ returns a pair (A_O, B_O) such that B_O can be compressed, and $\mathsf{Join}(A_O, B_O) = O$. No further security assumptions on Protect nor functionality constraints on $(\mathsf{Join}, \mathsf{Separate})$ are given.

The scheme involves a signature scheme Sig, a symmetric encryption scheme with secret key K, and a cryptographic hash function H. $\mathsf{Protect}(O)$ computes $(A_O, B_O) \leftarrow \mathsf{Separate}(O)$, sets $C_O = \mathsf{Compress}(O)$, and sets $X \leftarrow E_K(C_O \| H(O))$, $s \leftarrow \mathsf{Sig}(A_O \| X)$, and $\overline{O} = \mathsf{Join}(A_O, X \| s)$; verification of an object $\overline{O}$ runs $\mathsf{Separate}(\overline{O})$ to obtain A_O and $X \| s$, and checks that $s \in \mathsf{Sig}(A_O \| X)$.

We show that this scheme requires some further assumption on Join and Separate, at a minimum, in order to be secure. Specifically, an adversary can query an object O to the Protect oracle to obtain an object $O' = \mathsf{Join}(K_W, A_O, X \| s)$. The adversary then runs $\mathsf{Separate}(K_W, O')$ to obtain A_0 and $X \| \mathsf{Sig}(A_0 \| X)$. Let $X = X_1 \| X_2$, where, e.g., $|X_1|$ is one byte. From this, the adversary forms A_P as $A_O \| X_1$ $W' = X_2 \| s$, and $P \leftarrow \mathsf{Join}(A_P, W')$. Verification on the resulting P will succeed, but P was not the result of a query to the Protect oracle.

B.2 Adelsbach et al.'s Definition

Adelsbach, Katzenbeisser, and Veith define a watermarking scheme as a triple $\langle G, E, D \rangle$ of probabilistic polynomial time algorithms. Algorithm G is the key generator: on input 1^{n_k}, where k is the security parameter, G outputs a watermarking key $K \in \{0,1\}^k$ of length k.

The algorithm E is the watermark embedding process. On input of a digital object O, a watermark message $W \in \{0,1\}^n$, and a key K, it outputs a watermarked object O'. The object O' is required to be "perceptually similar" to the original object O.

Finally, the algorithm D is the watermark detector. Given a possibly marked object O', a candidate original object O, a candidate watermark W, a key K, and an auxiliary input Aux that does not depend on the object O, algorithm D either outputs true or false . The output true indicates the presence of the watermark W in the object O'. Adelsbach et al. require with overwhelming probability that $D(E(O, W, K), O, W, K, Aux) = \mathsf{true}$ for all objects O, watermarks W, and keys K. The authors then formally define security against copy attacks and ambiguity attacks as follows.

Definition 1. *Let W be a watermark, K be a watermarking key, O_1 be an arbitrary object, and O_1' its watermarked version, i.e. $D(O_1', O_1, W, K, Aux) = \mathsf{true}$ for some auxiliary input Aux. A* copy attack *on the watermark is a probabilistic algorithm* $\textsc{Copy}(O_1', O_2, Aux)$ *that either succeeds and outputs O_2' such that $D(O_2', O_2, W, K, Aux) = \mathsf{true}$ or fails and outputs a special failure symbol. We say that a watermarking scheme* *is (t, ϵ)-secure against copy attacks if all copy attacks running in time at most t have success probability at most ϵ.*

An ambiguity attack *on the watermark is a probabilistic algorithm* $\textsc{Ambig}(O', Aux)$ *that either succeeds and outputs (W, K, O) such that $D(O', O, W, K, Aux) = \mathsf{true}$. or fails and outputs a special failure symbol. We say that a watermarking scheme is (t, ϵ)-secure against ambiguity attacks if all ambiguity attacks running in time at most t have success probability at most ϵ.*

We note several shortcomings in these definitions. First, there is no requirement that a watermark be hard to remove in the underlying definition of a watermarking scheme. Another limitation of these definitions is that they do not consider attacks where the adversary learns many (O_i, O'_i) pairs. For instance, in a chosen-object attack, the attacker chooses an object O_i, convinces a legitimate participant to watermark it, and learns O'_i. The adversary might be able to repeat this process many times, adaptively. Their definitions do not consider this attack model. Our notion of strong watermarking, in contrast, allows an adversary adaptive access to a marking oracle and provides an easy way to quantify this access by measuring the number of oracle queries allowed.

B.3 Adelsbach et al.'s Scheme

Adelsbach et al. also propose several constructions that use digital signature schemes to improve the security of watermarks against copy and ambiguity attacks. In spirit, these schemes are similar to the work we present in Section 4.1. We now describe Scheme C of Adelsbach et al. The scheme requires a trusted third party (TTP) with signature public key P and secret key S. To mark an object O, we first pick an arbitrary identity string ID and a watermarking key K. We then set the watermark W to $ID||\mathsf{Sig}_S(O \otimes (ID||K), S)$ and embed W to obtain O'.

Here, the $||$ denotes concatenation. The operator $s_1 \otimes s_2$ is a special XOR operation; if $|s_1| = |s_2|$, then $\otimes$ denotes the ordinary XOR. If $|s_1| < |s_2|$ or $|s_1| > |s_2|$ the smaller string is repeated in a cyclic manner and cut off at the appropriate position, before computing the XOR operation. The authors assume that the length of a digital signature is constant and known in advance. The detection process for Scheme C is expressed by the following pseudocode:

$D_{P,C}(O', O, W, K, P)$:
1.Parse W as $W_1||W_2$
2.if $D(O', O, W, K) = \mathsf{false}$ then
3. return false
4.if $\mathsf{Ver}(O \otimes (W_1||K), W_2, P) = \mathsf{true}$ then
5. return true else
6. return false

First, we notice that a signature on the embedding key K is provided to the adversary as part of the watermark W. Depending on the signature scheme, this may reveal K (e.g. if the signature scheme has the message recovery property). Therefore, it is not clear what role is played by K, since it cannot be considered secret.

We further observe that Scheme C has the following property. Suppose that knowledge of the key K allows insertion of arbitrary watermarks. Then, given a marked object O_1 with watermark W_1 an adversary can create an object O' with a mark $W' = W'_1||W'_2$ that depends on the mark W. This follows because the adversary may find an O' such that $O' \otimes (W'_1||K) = O_1 \otimes (W_1||K)$.

This property does not count as a "copy attack" under Adelsbach et al's definition. The reason is that their definition requires that the embedded watermark string be exactly the same between the two marked objects. We suggest that this property is undesirable for a watermarking scheme, and therefore the fact that it does not fall under the definition of a copy attack is a shortcoming of the definition of Adelsbach et al. While one could extend the definition of copy attack to preclude these types of similar watermarks, there would remain the question of whether this extension went far enough. In contrast, our notion of strong watermarking prevents this attack because the adversary cannot find an object O' such that O' appears marked, yet O' was not submitted to the mark oracle.

B.4 Li and Chang's Definition

Li and Chang propose a definition of security against ambiguity attacks. In their definition, an ambiguity attack algorithm B is given an object O' which may or may not be marked with a key K. If O' is not marked, then the algorithm B succeeds if it outputs a special symbol $\perp$. Otherwise, if O' is marked, the algorithm B succeeds if it outputs a pair (W, K') such that O' contains the watermark W under key K'. They note that their definition differs from previous definitions of security against ambiguity attacks in that their adversary's success condition changes depending on whether the object has been marked or not. Unfortunately, this definition is too restrictive – it does not allow for the adversary to output a (W, K') that succeeds with an object O'' that is "close" to the target O'. Their definition also makes no requirement that the watermark be hard to remove.

C On the need for the Challenge Oracle

Oracle $\mathsf{Mark}^*(O)$:	**Oracle** $\mathsf{Detect}^*(O)$:	**Experiment** $\mathbf{Exp}_W^{wm1}(A)$:
1. $O' \leftarrow \mathsf{Mark}(K, O)$	1. $b \leftarrow \mathsf{Detect}(K, O)$	1. $K \leftarrow \mathsf{WMGen}(1^k)$
2. $\mathsf{Marked} \leftarrow \mathsf{Marked} \cup \{O'\}$	2. **if** $\exists O' \in \mathsf{Marked} \ : \ O \sim O'$	2. $\mathsf{bad} \leftarrow \mathsf{false}$
3. **return**(O')	3. **then** $b' = \mathsf{true}$	3. $\mathsf{Marked} \leftarrow \emptyset$
	4. **else** $b' = \mathsf{false}$	4. $A^{\mathsf{Mark}^*, \mathsf{Detect}^*}()$
	5. **if** $b \neq b'$	5. **return** (bad)
	6. **then** $\mathsf{bad} \leftarrow \mathsf{true}$	
	7. **return**(b)	

Fig. 6. Watermarking security definition based on direct comparison with the ideal scheme

In Section 3 we state that the intent of our security definition is to compare a watermarking scheme (which is typically stateless) to the ideal watermarking scheme. The most obvious way of formalizing this comparison, from a cryptographic perspective, is the experiment shown in Figure 6. Here, an adversary is given access to Mark and Detect queries for some uniformly chosen key K, and succeeds in attacking the scheme if he can find an object O so that $\mathsf{Detect}(K, O)$ and the ideal watermarking scheme disagree: either $\mathsf{Detect}(K, O) = \mathsf{true}$ and O is not similar to any object queried to $\mathsf{Mark}(K, \cdot)$ or $\mathsf{Detect}(K, O) = \mathsf{false}$ yet O is similar to some object resulting from a query to $\mathsf{Mark}(K, \cdot)$. We define the advantage of the adversary A against the scheme W to be $\mathbf{Adv}_W^{wm1}(A) = \Pr[\mathbf{Exp}_W^{wm1}(A) = \mathsf{true}]$ and say W is (t, q_M, q_D, ϵ) secure if every adversary that runs in time at most t and makes at most q_M Mark^* queries and q_D Detect^* queries has advantage at most ϵ.

Unfortunately, the following result shows that no scheme that is δ preserving (where δ is the "similarity" threshold for metric space $\mathcal{M}$) with high probability can be secure in this sense. Formally, suppose that $\mathcal{D}$ is a distribution on $\mathcal{M}$ such that $\Pr[K \leftarrow \mathsf{WMGen}(1^k); O \leftarrow \mathcal{D} : \mathsf{Mark}(K, O) \sim O] = p$; if the watermark is ρ-preserving for any $\rho \leq \delta$ then p is negligibly close to 1. The following adversary has advantage at least p:

$\mathbf{A_1}$:
1. $O \leftarrow \mathcal{D}$
2 $\mathsf{Detect}^*(O)$
3. $O' = \mathsf{Mark}^*(O)$
4. $\mathsf{Detect}^*(O)$.

To see that $\mathbf{Adv}_W^{wm1}(\mathbf{A_1}) \geq p$, let us denote by A2 the event that $\mathsf{bad} \leftarrow \mathsf{true}$ in line 2 and A4 the event that $\mathsf{bad} \leftarrow \mathsf{true}$ in line 4, but not line 2. Then we have $\mathbf{Adv}_W^{wm1}(\mathbf{A_1}) = \Pr[\mathsf{A2}] + \Pr[\mathsf{A4}]$. Now suppose that once we draw $O \leftarrow \mathcal{D}$ and $K \leftarrow \mathsf{WMGen}(1^k)$, we do not have $\mathsf{Detect}^*(K, O) = \mathsf{true}$; then in line 4, $\mathbf{A_1}$ queries an object O that is similar to the result of a Mark^* query (with probability p) but unmarked;

Thus $\Pr[\mathsf{A4}] = \Pr[O' \sim O \wedge \neg \mathsf{A2}]$. And since it is also true that $\Pr[\mathsf{A2}] \geq \Pr[\mathsf{A2} \wedge O' \sim O]$, we then have $\mathbf{Adv}_W^{wm1}(\mathbf{A_1}) \geq \Pr[O' \sim O \wedge \neg \mathsf{A2}] + \Pr[O' \sim O \wedge \mathsf{A2}] = \Pr[O' \sim O] = p$.

Oracle $\mathsf{Mark}^*(O)$:	**Oracle** $\mathsf{Detect}^*(O)$:	**Experiment** $\mathbf{Exp}_{\mathcal{D},W}^{WM2}(A)$:
1. $\mathsf{MarkQueries} \leftarrow \mathsf{MarkQueries} \cup \{O\}$	1. $b \leftarrow \mathsf{Detect}(K, O)$	1. $K \leftarrow \mathsf{WMGen}(1^k)$
2. $O' \leftarrow \mathsf{Mark}(K, O)$	2. **if** $\exists O' \in \mathsf{Marked} \; : \; O \sim O'$	2. $\mathsf{bad} \leftarrow \mathsf{false}$
3. $\mathsf{Marked} \leftarrow \mathsf{Marked} \cup \{O'\}$	3. **then** $b' = \mathsf{true}$	3. $\mathsf{Marked} \leftarrow \emptyset$
4. **return**(O')	4. **else** $b' = \mathsf{false}$	4. $A^{\mathsf{Mark}^*, \mathsf{Detect}^*}()$
	5. **if** $b \neq b'$ **and** $O \notin \mathsf{MarkQueries}$	5. **return** (bad)
	6. **then** $\mathsf{bad} \leftarrow \mathsf{true}$	
	7. **return**(b)	

Fig. 7. Experiment that rules out success based on submitting unmarked originals

At first glance, this problem seems similar to the problem of having chosen-ciphertext security for an encryption scheme without restricting the adversary to disallow querying the challenge ciphertext to his decryption oracle. This suggests a second possible security experiment which prevents the adversary from defeating a watermarking scheme by querying Detect^* on "unmarked originals", shown in Figure 7. Unfortunately, this definition rules out strong watermarks that are $\delta/2$ preserving with high probability. Suppose, for some sampleable distribution $\mathcal{D}$ on $\mathcal{M}$, we have that $\Pr[O \leftarrow \mathcal{D}; K \leftarrow \mathsf{WMGen}(1^k) : d_{\mathcal{M}}(O, \mathsf{Mark}(K, O)) \leq \delta] \geq \psi$. We give an adversary $\mathbf{A_2}$ with advantage at least $(\psi/2)^3$ in $\mathbf{Exp}_W^{wm2}$:

$\mathbf{A_2}$:
1. $O \leftarrow \mathcal{D}$
2.. $K' \leftarrow \mathsf{WMGen}(1^k)$
3. $O' \leftarrow \mathsf{Mark}(K', O)$
4 $\mathsf{Detect}^*(O')$
5. $O'' = \mathsf{Mark}^*(O)$
6. $\mathsf{Detect}^*(O')$.

In analyzing this attack, we note that the adversary succeeds with probability at least $\Pr[O' \sim O'']$, since whenever this happens, either O' is marked before O is queried to Mark^*, or O' is not marked according to $\mathsf{Detect}(K, \cdot)$, but should be marked according to the ideal scheme, in line 6. A probabilistic lemma used in the proof of the so-called "forking lemma" [48] implies that $\Pr[O' \sim O''] \geq (\psi/2)^3$. We also note that if the metric space $\mathcal{M}$ is such that it is easy to find an object O' such that $d_{\mathcal{M}}(O, O') < \mu$, then a similar attack is possible with probability $\Pr[K \leftarrow \mathsf{WMGen}(1^k); O \leftarrow \mathcal{D} : d_{\mathcal{M}}(O, \mathsf{Mark}(K, O)) < \delta - \mu]$.

This leads us to the conclusion that the adversary should not be allowed to know the "unmarked originals" for the objects he tries to unmark. Thus, we must introduce some sort of "challenge" oracle that draws objects from a probability distribution and marks them. An additional question then becomes: what should be the probability distribution of the $\mathsf{Challenge}^*$ oracle? The most secure definition would allow the adversary some control over the distribution of "originals" marked by the oracle. However, it is easy to construct examples, under reasonable assumptions on the metric space $\mathcal{M}$ where this still fails. For example, if the adversary chooses a distribution with low entropy he may still guess what the originals picked by $\mathsf{Challenge}^*$ will be and carry out the unmarked original attack anyway. Or, he may choose a distribution $\mathcal{D}$ with high entropy, but such that each point in the support of $\mathcal{D}$ is the only in a radius much wider than the expected distortion of the marking procedure. Rather than attempt to address all the necessary conditions on adversaries that make this definition interesting, we choose to be agnostic about the existence of such distributions and include the distribution $\mathcal{D}$ as a security parameter.